future internet

MDPI

Review

Methods of Annotating and Identifying Metaphors in the Field of Natural Language Processing

Martina Ptiček * and Jasminka Dobša

Faculty of Organization and Informatics, University of Zagreb, 42000 Varaždin, Croatia; jasminka.dobsa@foi.hr
* Correspondence: marpticek@student.foi.hr

Abstract: Metaphors are an integral and important part of human communication and greatly impact the way our thinking is formed and how we understand the world. The theory of the conceptual metaphor has shifted the focus of research from words to thinking, and also influenced research of the linguistic metaphor, which deals with the issue of how metaphors are expressed in language or speech. With the development of natural language processing over the past few decades, new methods and approaches to metaphor identification have been developed. The aim of the paper is to map the methods of annotating and identifying metaphors in the field of natural language processing and to give a systematic overview of how relevant linguistic theories and natural language processing intersect. The paper provides an outline of cognitive linguistic metaphor theory and an overview of relevant methods of annotating linguistic and conceptual metaphors as well as publicly available datasets. Identification methods are presented chronologically, from early approaches and hand-coded knowledge to statistical methods of machine learning and contemporary methods of using neural networks and contextual word embeddings.

Keywords: metaphor; metaphor annotation; metaphor identification; neural networks; word embeddings; large language models

Citation: Ptiček, M.; Dobša, J. Methods of Annotating and Identifying Metaphors in the Field of Natural Language Processing. *Future Internet* **2023**, *15*, 201. https://doi.org/10.3390/fi15060201

Academic Editor: Paulo Quaresma

Received: 25 March 2023
Revised: 25 May 2023
Accepted: 29 May 2023
Published: 31 May 2023

1. Introduction

Metaphors are an integral part of speech and enrich communication, but they also, as shown by Lakoff and Johnson [1], form our opinion about the world and the phenomena that surround us. For instance, we often use the metaphorical concept TIME IS MONEY (Lakoff and Johnson also introduced writing the concepts in capital letters, which is generally accepted in the literature dealing with (conceptual) metaphors). (e.g., "You're *wasting* my time.") or LOVE IS A JOURNEY (e.g., "This relationship is a *dead-end street*.") (Examples taken from [1]). These concepts allow us to map knowledge about one domain, which is physical and more basic (money, travel), to a domain that is more complex and abstract (time, love).

Metaphors are used in speech and expressed with words, which makes them an area of interest in linguistic research, but they also represent a great challenge in the field of natural language processing. Precisely because they are an integral part of speech and of how we express ourselves, if we want to achieve further progress in natural language processing and artificial intelligence, metaphors are a phenomenon that needs to be studied in an interdisciplinary manner, so that we can understand their cognitive linguistic theory and recognize the reach of computational linguistics as well as machine and deep learning.

In this context, this paper aims to map the methods of annotating and identifying metaphors in the field of natural language processing and to give a systematic overview of how relevant linguistic theories and natural language processing intersect. This is done through firstly providing an outline of cognitive linguistic metaphor theory, which is given in the section following this introduction. The theoretical framework is then followed by presenting the approaches and procedures for metaphor annotation in text. In addition

to giving an overview of the approaches and procedures, this section also shows how linguistic theories of metaphor are applied in the creation of data sets, which are then used in identifying metaphors in machine learning methods. Finally, the last section of the paper provides an overview of research on computer methods of metaphor identification, from approaches with hand-coded knowledge and lexical resources, to the latest state-of-the-art models that use neural networks and contextual word embeddings (i.e., large language models).

The approach of the paper consists of systematic and chronological presentation of annotation methods and developments in identifying metaphors in the field of natural language processing. Acknowledging the fact that the field of natural language processing is vast and that new approaches are constantly (and simultaneously) being generated, this paper brings the results of the most relevant current research relating to metaphor identification in one place. Therefore, its ambition is not to decide on the success of the methods/models presented but rather to map the field and present the current state of play. The ever-growing tendency which characterizes this field makes it difficult for researchers to grasp the developments without investing significant time. With providing a systematic and comprehensive overview, this paper serves as an introduction to the field and presents a source of information for a broad spectrum of researchers who (are starting to) work in the field of natural language processing.

2. Theories of the Metaphor

Since Lakoff and Johnson [1] published their Conceptual Metaphor Theory (CMT), in the book "Metaphors we live by" in 1980, this has been the dominant theory in metaphor research. The theory of the conceptual metaphor is considered to have been most influential in shifting the focus from words and expressions to the cognitive process, prompting the development of the cognitive theory of the metaphor, a subfield of cognitive linguistics that further developed new theories, criticism, and approaches [2].

While the theory of the conceptual metaphor deals with how the metaphor impacts our understanding of the world, linguistic metaphor research deals with the metaphor at the level of language, and how metaphors are expressed in language.

Below is an overview of the conceptual metaphor theory, followed by an overview of the linguistic metaphor.

2.1. Conceptual Metaphor

In "Metaphors we live by", Lakoff and Johnson present their theory stating that metaphors influence our thinking and actions, and that our conceptual system is metaphorical in nature. They support the theory with concepts such as ARGUMENT IS WAR (e.g., "Your claims are *indefensible.*", "He *attacked every weak* point in my argument.")–stating that it is "important to see that we don't just talk about arguments in the terms of *war*", but that we actually can "win or lose arguments". Furthermore, the authors conclude that "the essence of metaphor is understanding and experiencing one kind of thing in terms of another". In the specific example of ARGUMENT IS WAR, we comprehend ARGUMENT and talk about it by means of WAR. The authors support their theory by citing a series of concepts and examples from everyday life and communication (for example, TIME IS MONEY, LOVE IS A JOURNEY, MIND IS A MACHINE, THEORIES ARE BUILDINGS).

At the core of the conceptual metaphor is the mapping of one domain (source) to another (target), as shown in the example ARGUMENT IS WAR, where WAR is the source and ARGUMENT the target domain. This mapping takes place not only at the linguistic level but also at the cognitive level, which affects our understanding of the world and the phenomena that surround us. Lakoff and Johnson focus precisely on the discovery of such concepts (in the mind) and primarily deal with conventional metaphors (cf. [3]), which are so common in everyday speech (and life) that we do not even notice them.

It is important to distinguish among common, conventional metaphors and novel metaphors. (It could be argued whether the conventional—novel metaphor distinction

is a binary one or a scale, but this discussion is out of the scope of this paper.) What is interesting about novel metaphors is how some of them become conventional. Those that do are often accepted in many languages [3,4]—which means that these languages, apart from sharing the same metaphors on the linguistic level, share the same concepts, i.e., understanding of the world.

Some of the concepts cited by Sulivan [4] that are applicable in several languages are, for example, describing a person's intelligence through brightness, which we find, for example, in English "*bright* student" and Spanish "una persona *brilliante*". Describing anger as a hot liquid (e.g., "let off *steam*", "*steaming* at the ears") is also a concept that we find in a number of languages—English, Japanese, Chinese and Hungarian. The existence of the same or similar concepts in different languages is further evidence for the conceptual metaphor theory.

Another characteristic of the conceptual metaphor is its asymmetry, i.e., its unidirectionality—if we use the same example of ARGUMENT IS WAR, we can see that we can express argument using the experience of war, but we will not explain war through the experience of the argument. The same applies, for instance, to LOVE IS A JOURNEY—we do not talk about a journey using the experience of love. Lakoff and Johnson [1] explain this phenomenon by noting that concepts that are less clear and less concrete are explained through concepts that are more clear, concrete and grounded in our experience.

Relying on the theory of the conceptual metaphor, Grady [5] in his doctoral thesis develops the theory of primary metaphor, which shows us how metaphors can be divided into primary and complex metaphors, which we create from primary metaphors. Primary metaphors represent simple patterns such as MORE IS UP (e.g., "My income *rose* last year") (Example for MORE IS UP concept taken from [1]), and these are the atoms that make up molecules, i.e., complex metaphors [5,6]. Grady gives the example of the concept THEORIES ARE BUILDINGS and decomposes this concept into simpler and more primary concepts (atoms) making up a more complex concept (molecule)—ORGANIZATION IS PHYSICAL STRUCTURE and VIABILITY IS ERECTNESS.

Lakoff and Johnson [7] accepted the primary metaphor theory and developed it further. The authors state that the theory of the integrated primary metaphor has four parts, i.e., four individual theories that they develop further:

1. Theory of conflation [8] which proposes that children combine sensory and non-sensory experiences from a very early age, resulting in later understandings of metaphors such as "close friend", "warm smile" and "big problem".
2. The theory of the primary metaphor [5,6], which states that all complex metaphors are "molecules" composed of "atomic" metaphorical parts, which we call primary metaphors.
3. Neural theory of metaphor [9] that claims that the associations created during the conflation period are realized through permanent neural connections in the neural network that defines the conceptual domains.
4. Theory of conceptual blending [10] which states that distant domains can be connected and create new deductions.

These four parts taken together form the integrated theory of the primary metaphor, which claims that people acquire a large system of primary metaphors automatically and unconsciously, in everyday and ordinary situations, from an early age, which leads to us thinking by using hundreds of primary metaphors [7]. The authors further present a short representative list containing 23 primary metaphors (AFFECTION IS WARMTH, IMPORTANT IS BIG, HAPPY IS UP, INTIMACY IS CLOSENESS, BAD IS STINKY, DIFFICULTIES ARE BURDENS, MORE IS UP, CATEGORIES ARE CONTAINERS, SIMILARITY IS CLOSENESS, LINEAR SCALES ARE PATHS, ORGANIZATION IS PHYSICAL STRUCTURE, HELP IS SUPPORT, TIME IS MOTION, STATES ARE LOCATIONS, CHANGE IS MOTION, ACTIONS ARE SELF-PROPELLED MOTIONS, CAUSES ARE PHYSICAL FORCES, RELATIONSHIPS ARE ENCLOSURES, CONTROL IS UP, KNOWING IS SEEING, UNDERSTANDING IS GRASPING, SEEING IS TOUCHING, PURPOSES ARE DESTINATIONS, PURPOSES ARE DESIRED OBJECTS).

In an appendix to his doctoral dissertation [5], Grady also provides a list of primary metaphors, which he divides into 5 main categories: Atemporal relations; Quantity and degree; Time, action and event structure; Affect, evaluation and social relations; Thought and consciousness, which contain a total of about 100 primary metaphors, i.e., concepts. It should be emphasized that none of these lists is definitive.

Grady [6] states that "humans everywhere share the basic patterns of experience that are reflected in primary metaphor" so it is possible to find these patterns in various languages, and primary metaphors are the same in different languages. For example, warmth is associated with affection which is considered as something good, and this concept has its roots in the warmth children feel when their parents hold them in their arms (e.g., "*Warm* welcome", "Thank you for your *warm* words"). Lakoff and Johnson [7] state that primary metaphors are part of unconscious cognition, and that we acquire them unconsciously as part of the normal learning process that we have no influence over. This process is the same for all people, which points to the fact that the acquisition of primary metaphors is universal. The authors emphasize that they are not innate, but learned, and are manifested through language, i.e., words.

2.2. Linguistic Metaphor

Lakoff and Johnson's theory of the conceptual metaphor is undoubtedly extremely important and has greatly influenced metaphor research, but the authors do not deal with the linguistic metaphor, that is, with how the conceptual metaphor is expressed in language. Knowledge of the theory of the conceptual metaphor is certainly necessary in order to recognize or identify a metaphor in a sentence (or text), but a metaphor is expressed by language and in language (There are, for example, visual metaphors, but this is beyond the scope of this paper), and it is language that is the subject of research when trying to identify a metaphor.

When discussing the linguistic metaphor, in the context of creating a system for identifying metaphors, Shutova [11] believes that attention needs to be given to the degree of conventionality of the metaphor, the syntactic construction of metaphorical expressions and the level at which we observe or annotate the metaphor—at the level of a word or a sentence or at the level of the source and target domains within grammatical relations. [12]

By exploring metaphors in educational discourse, Cameron [12] provides an overview of the linguistic metaphor. On a data set of 26,613 words, Cameron finds 711 instances of the linguistic metaphor and concludes that the frequency of the metaphor is 27 metaphors per 1000 words. She also states that, by type of word, metaphors are most often expressed by verbs (47%, which includes verbs, phrasal verbs, and verb phrases), followed by prepositions (34%, which also includes prepositional phrases).

Furthermore, Shutova and Teufel [13] give a statistical overview of corpora that contains texts from six domains (literature, politics, news, sociology, speeches and scientific articles on literature). These authors state that metaphor expressed in verbs present 68% of all metaphors, which clearly indicates that metaphors are much more often expressed by verbs than by other types of words.

3. Annotating Metaphors

In order to be able to analyze and research the metaphor, the necessary first step is data—texts in which the metaphor is annotated—regardless of whether we are doing a machine computer analysis or whether we are manually researching the metaphor. For getting an insight into currently available data sets, at the end of the section *Annotating Linguistic Metaphors* a table is presented with publicly available data sets, which are commonly used in metaphor identification models (Table 1).

In this section we provide an overview of relevant methods and procedures in metaphor annotation, those used in the humanities, and those that approach annotation from a computer and information science perspective.

Annotating a metaphor is approached from two main directions—(1) annotating a linguistic metaphor and (2) creating a list and database of metaphorical concepts (cf. [14]). Annotating a linguistic metaphor implies annotating the very words or expressions that are used metaphorically, while the creation of lists and databases focuses on listing metaphorical concepts, that is the source and target domains, accompanied by examples. These two approaches are sometimes combined, whereby annotating the linguistic metaphor also finds its theoretical framework for determining whether something is or is not a metaphor in the theory of the conceptual metaphor, i.e., in mapping the source onto the target domain (cross-domain mapping).

3.1. Annotating Linguistic Metaphors

The best-known and most frequently used procedures of annotating linguistic metaphors are Metaphor Identification Procedure (MIP) [15] and Metaphor Identification Procedure VU University Amsterdam (MIPVU) [16]. These deal with the recognition of linguistic metaphors in discourse at the level of words, i.e., lexical units. It should be noted that both procedures were primarily developed with the aim of using the texts (corpus) annotated by them for the purpose of further research in the field of social sciences and the humanities, which is why the needs and specifics of annotating metaphors in natural language processing were not taken into account. Figure 1 shows the flow of both procedures, as well as how they overlap, while below a brief overview and comparison of both procedures is given.

The MIP method is the predecessor of the MIPVU, which is its upgraded version. The MIP was developed by the Pragglejaz group (The name of the group is a series of initial letters of the names of the researchers who participated in the research and development of the procedure) and as the authors state, the aim of the procedure is to determine for each lexical unit in the discourse whether its use in a certain context can be described as metaphorical [15]. Furthermore, it is stated that the procedure has a maximalist rather than a minimalist approach, whereby a large number of words could be considered a metaphor, taking into account their use in a certain context.

The MIP method consists of three (seemingly) simple steps [15]:

1. Read the whole text or transcript to understand what it is about.
2. Decide about the boundaries of words.
3. Establish the contextual meaning of the examined word.
    a. Determine the basic meaning of the word (most concrete, human-oriented and specific).
    b. Decide whether the basic meaning of the word is sufficiently distinct from the contextual meaning.
    c. Decide whether the contextual meaning of the word can be related to the more basic meaning by some form of similarity.

The MIP primarily identifies the indirect use of words (simply put, it is only decided whether a word is a metaphor or not), but the MIPVU adds steps aimed at "capturing" direct and implicit metaphors. As indicated by Krennmayr and Steen [17], the need to annotate direct and implicit metaphors arose in the course of their research, which began with the application of MIP on a corpus composed of texts from four registers from the BNC Baby corpus (The corpus is available at http://www.natcorp.ox.ac.uk/corpus/babyinfo.html; last accessed on 11 February 2023)—conversation, fiction, academic texts and newspaper articles.

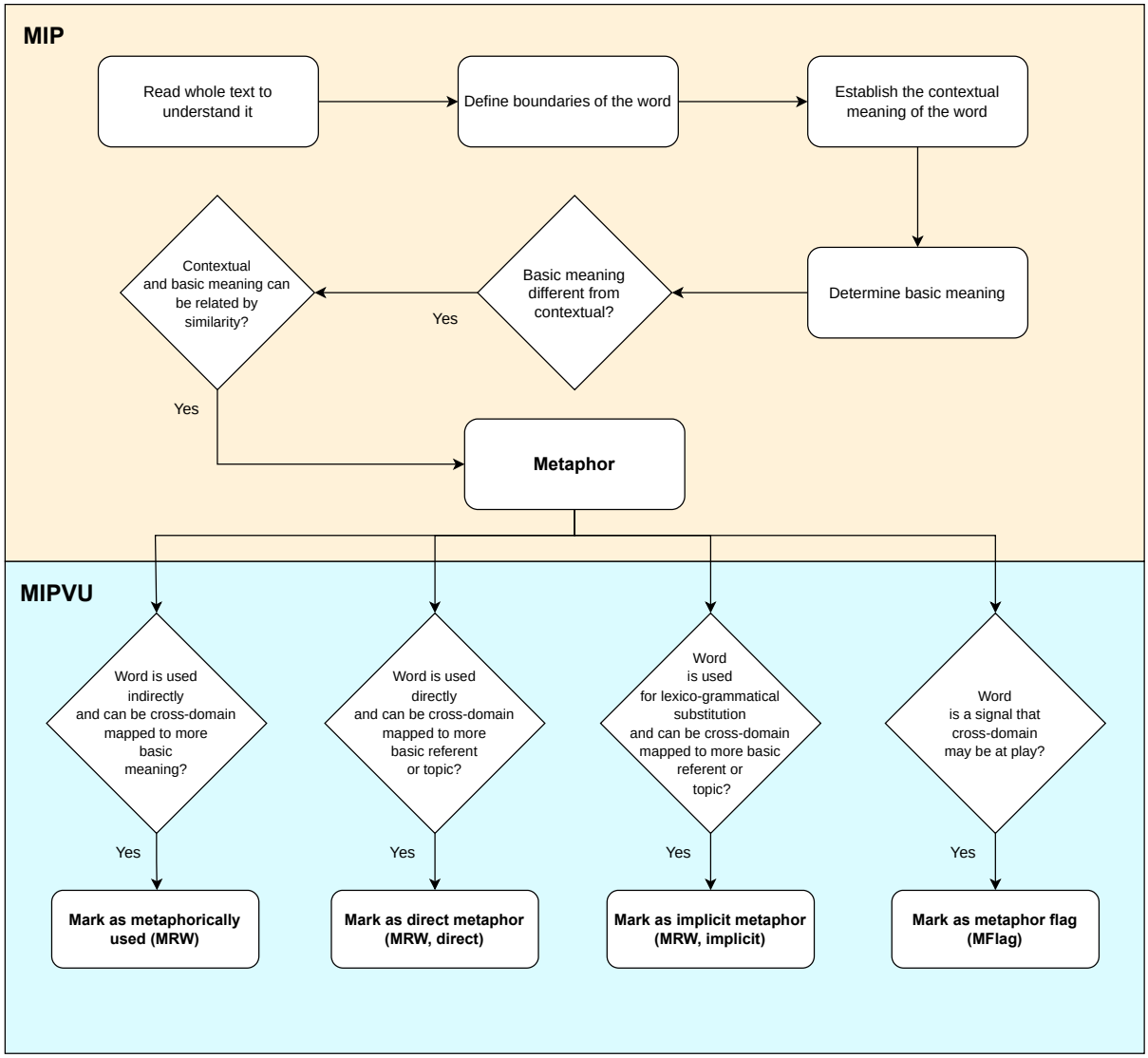

Figure 1. Flow diagram of MIP and MIPVU, showing the overlap between them.

The MIPVU procedure is as follows [16]:

1. Find metaphor-related words (MRWs) by examining the text on a word-by-word basis.
2. When a word is used indirectly and that use may potentially be explained by some form of cross-domain mapping from a more basic meaning of that word, mark the word as metaphorically used (MRW).
3. When a word is used directly and its use may potentially be explained by some form of cross-domain mapping to a more basic referent or topic in the text, mark the word as direct metaphor (MRW, direct).
4. When words are used for the purpose of lexico-grammatical substitution, such as third person personal pronouns, or when ellipsis occurs where words may be seen as missing, as in some forms of co-ordination, and when a direct or indirect meaning is conveyed by those substitutions or ellipses that may potentially be explained by some form of cross-domain mapping from a more basic meaning, referent, or topic, insert a code for implicit metaphor (MRW, implicit).
5. When a word functions as a signal that a cross-domain mapping may be at play, mark it as a metaphor flag (MFlag).
6. When a word is a new-formation coined, examine the distinct words that are its independent parts according to steps 2 through 5.

Points 1 and 2 of MIPVU procedure are, as authors state, essentially the same as MIP, and as it is shown in Figure 1, MIPVU procedure upgrades the MIP—once a metaphor is detected, MIPVU guide us on determining the metaphor type. As we can notice, the MIPVU mentions "cross-domain mapping", which indicates that the metaphor needs to be considered through the conceptual metaphor theory, while Krennmayr and Steen [17] state that in order to identify direct and implicit metaphors, metaphor identification absolutely needs to be approached from the level of the conceptual metaphor. Direct metaphors are often expressed through comparisons using the preposition *like* or *as*, which means that, if we do not apply the theory of the conceptual metaphor here, the metaphor will not be identified. The implicit metaphor, on the other hand, is expressed by pronouns and ellipsis, where the pronoun or ellipsis forms an indirect or direct meaning that can be explained by cross-domain mapping.

While the MIP states that it is based on a maximalist approach, Krennmayr and Steen find that, despite the MIPVU procedure giving the impression that everything could ultimately be a metaphor, practice shows that only 13.6% of the lexical items in the corpus are annotated as metaphors (The distribution is also interesting—18.5% academic texts, 16.4% newspaper articles, 11.9% fiction and 7.7% conversation. We would certainly expect a higher percentage of metaphors in fiction since conventional wisdom is that metaphors are reserved for fiction. Shutova [11] also reports the highest percentage of metaphors in newspaper articles, politics, and in a language and literature journal). Also, they state that direct and implicit metaphor each make up only 0.2% of that.

The authors of both procedures complicate the definition of the lexical unit—most often the lexical unit coincides with the word, but this is not always the case. For instance, in the English language, there are phrasal verbs, such as *get up*, *get out*, and *give up*, which must be viewed as one lexical unit. In addition to phrasal verbs, there are also phrases, compounds, expressions, and the like for which it also needs to be determined whether they are one lexical unit or whether each word should be viewed as a separate lexical unit. The MIPVU procedure recommends that all words, except phrasal verbs, compounds (e.g., *underpass, power plant*) and names be considered as one lexical unit.

Both procedures state that in identifying a metaphor, i.e., in determining the basic meaning of lexical units as well as in determining the lexical units themselves, it is necessary to consult monolingual dictionaries. The MIP, for instance, uses the *Macmillan English Dictionary for Advanced Learners* as a primary dictionary and the *Shorter Oxford English Dictionary* as a supplementary dictionary for information about the etymology of words. As its two main dictionaries, the MIPVU uses the *Macmillan English Dictionary for Advanced Learners* and the *Longman Dictionary of Contemporary English Online*, and additionally consults the *Oxford English Dictionary Online*.

In the development of the MIPVU procedure, data were also annotated, which led to the publicly available annotated data corpus, the VU Amsterdam Metaphor Corpus (The data set is available at http://www.vismet.org/metcor/documentation/home.html (last accessed on 11 February 2023), and is referred to as VUA in the rest of this paper, as is usual in research papers and relevant literature), which contains 186,695 lexical items, annotated as metaphors (or non-metaphors). It is this data set that is most often used when developing models for machine learning and natural language processing for the purpose of identifying metaphors.

It should be noted that both the MIP and MIPVU procedures are primarily designed to identify metaphors in the English language, and the application of these methods in other languages requires adaptation to the specifics of each language. Nacey et al. [18] provides an overview of the challenges faced by researchers when applying the MIPVU procedure in Dutch, German, Lithuanian, Polish, Serbian, Uzbek, Chinese and Sesotho. For instance, Bogetić et al. [19] adapted the MIPVU to the Serbian language and found that reflexive verbs need to be viewed as one lexical unit, and a preposition followed by a negative pronoun should be divided into two lexical units (e.g., *ni s kim (with no one)*) should be divided into the first unit, "*ni s*" ("*with no*"), and the second unit, "*kim*" ("*one*").

An additional specificity of Serbian, which does not exist in English, are cases or metaphors expressed by cases—oblique cases (genitive, dative, instrumental and accusative), which, when used without a preposition, can encode metaphor. In order to "capture" a metaphor in oblique cases, the authors propose an additional step in the MIPVU procedure, noting that their proposal is a preliminary solution that will certainly be further developed and improved over time.

When determining contextual and basic meaning and whether cross-domain mapping has occurred, the MIP and MIPVU recommend the use of dictionaries, but here again we encounter a challenge if we adapt the methods to other languages, since languages with small resources simply do not have such a developed lexicography as is the case with the English language.

At the time when Steen et al. [16] developed the MIPVU, Shutova and Teufel [13] also developed an annotation scheme based on the MIP method, but extended it to annotate conceptual metaphors, in addition to annotating linguistic metaphors themselves, i.e., metaphorical expressions. Shutova and Teufel concentrate on verbs, since, as they state, previous research has shown that in more than 50% of cases metaphors are expressed by verbs. In order to determine the source and target domains, using the Master Metaphor List [20], the authors define a list of the most common and general categories of the conceptual metaphor. The data that have been annotated are texts from the British National Corpus, from five registers—fiction (5293 words), newspaper articles (2086 words), research articles (1485 words), essays on politics, international relations, and sociology (2950 words) and transcribed radio broadcasts (1828 words)—a total of 13,642 words or 761 sentences [11]. Annotation takes place in two steps, where in the first step the MIP method is applied, and in the second the source and target domains are annotated for each identified linguistic metaphor.

In their research, Wallington et al. [21] divide annotators into two groups, Team A and Team B, where both groups have the same data set—conversations between patients and doctors about arthritis and selected texts from the British National Corpus. Each team receives different instructions on how to annotate. Team A is instructed to annotate interesting stretch expressions or words whose (1) actual meaning is not physical, (2) meaning in another context may be physical and of the same syntactic nature, (3) physical meaning is related to the abstract meaning. It is evident from the instructions that they describe a conceptual metaphor, but the metaphor is not mentioned in the instructions themselves, and the interesting stretch words do not refer to the metaphor in any way. The authors of the report state that Team A was soon disbanded because it was noticed that a lot of polysemy was annotated. Team B was instructed to annotate metaphors, including cases when "something abstract is represented as something physical, something abstract is represented as something abstract, something physical is represented as something physical, something physical is represented as something abstract." In the annotation process, a number of parameters were recorded, such as whether it is a novel or conventional metaphor, the level of certainty that it is a metaphor (on a scale from -2 to 2), and the metaphors were linked to concepts in the Master Metaphor List [20], if the concept does not exist in the list, the annotators defined a new concept.

Trope Finder (TroFi) data set that contains sentences in which 50 selected verbs appear, in their literal and figurative meanings, was created by Birke and Sarkar [22]. This data set was created as a part of their research of clustering method for distinguishing the literal and figurative meaning of verbs. Data set contains 6435 sentences, out of which 2740 are metaphorical. Despite the fact that the authors emphasize that the TroFi is not a metaphor recognition system, the data set created as part of this paper is publicly available (https://natlang.cs.sfu.ca/software/trofi.html; last accessed on 11 February 2023) and is often used in machine learning models for metaphor recognition, and we cite it here as well. The data set contains sentences in which 50 selected verbs appear, in their literal and figurative meanings.

As part of their metaphor identification research, Tsvetkov et al. [23] created an annotated dataset for English and Russian (TSV) for metaphors expressed by adjectives and nouns. The dataset is publicly available (www.cs.cmu.edu/~ytsvetko/metaphor/datasets.zip; last accessed on 11 February 2023) and is divided into a training set containing 884 metaphorical and 884 literal expressions/sentences and a testing set containing 111 metaphorical and 111 literal expressions. The training data set was annotated by 5 annotators and the testing set by only one annotator.

MOH-X [24] data set was created as part of the research aimed at answering the question of whether metaphorical expressions create stronger emotions and this data set is also publicly available (https://saifmohammad.com/WebPages/metaphor.html; last accessed on 12 February 2023). Mohammad et al. used a crowdsourcing platform and each annotated instance was annotated by ten annotators, while the final data set included those that 70% or more of the annotators annotated as a metaphor. The dataset was extracted from WordNet [25,26] based on the selected 440 verbs, where the criterion for selecting verbs was that they have more than three and less than ten meanings in WordNet, to ensure a higher probability of the verb being used metaphorically but also to avoid words with a double meaning. In addition to annotating the metaphor, the strength of the emotion expressed in each instance was also annotated for each instance, in order to test the hypothesis that metaphorical expressions are more emotional than literal ones.

To the best of our knowledge, KOMET 1.0 [27], annotated corpus of metaphors in the Slovenian language, is the only publicly available (https://www.clarin.si/repository/xmlui/handle/11356/1293; last accessed on 8 January 2023) annotated data set for any of the South Slavic languages. Antloga approaches annotation using the guidelines provided by the MIPVU method, and annotates indirect and direct metaphors, borderline cases and signals that a unit could be a metaphor. In addition to the MIPVU method, the frame is also marked in the corpus, which connects the metaphor with the concept to which it belongs. The corpus consists of newspaper texts, fiction and online texts, and contains 218,730 words.

Table 1. Overview of publicly available data sets which are commonly used in metaphor identification models.

Data Set	Procedure	Data Set Size
VUA [16,17]	MIPVU	186,695 words
TroFi [22]	Sentences containing 50 selected verbs	Total 6435 sentences, 2740 metaphorical
TSV [23]	Metaphors expressed by adjectives and nouns	995 metaphorical, 995 literal expressions
MOH-X [24]	440 verbs with >3 and <10 meanings in WordNet	Total 647 sentences, 315 metaphorical
KOMET 1.0 [25]	MIPVU	218,730 words

3.2. Lists and Databases of Metaphorical Concepts

The first list of metaphorical concepts was created by Lakoff, Espenson and Shwartz [20], and is called the Master Metaphor List (MML). The list is organized ontologically and specific concepts are presented as belonging to a more general concept—for instance, PURPOSE IS DESTINATION belongs to the more general STATES ARE LOCATIONS (cf. [14]). The MML contains source and target domains primarily in the field of the mind, feelings, and emotion.

While MML is presented as a "regular" list in the document, the MetaNet (More about the project can be found on https://metanet.arts.ubc.ca/; last accessed on 5 January 2023) project is presented as a Wiki application (https://metaphor.icsi.berkeley.edu/pub/en/; last accessed on 5 January 2023), containing a repository of conceptual metaphors. At

the time of writing this paper, MetaNet contains 685 conceptual metaphors, which are interrelated (e.g., TO ACCEPT AN IDEA IS TO EAT is related to IDEAS ARE FOOD and TO ACCEPT IS TO SWALLOW) and are accompanied by examples.

3.3. Metaphor Annotation for Natural Language Processing

The MIP and MIPVU annotation procedures originated in the need for research in linguistics and deal with the annotation of linguistic metaphors. Also, other methods, i.e., other data sets that have been used for the last dozen years in machine identification of metaphors, such as TroFi, TSV and MOH-X, annotate linguistic metaphors and do not deal with how to annotate a metaphor for the purposes of machine identification, i.e., for the purposes of processing natural language. However, Shutova and Teufel [13] approach metaphor annotation from the perspective of computer science, and in a later paper Shutova [11] presents interesting questions and thoughts on what kind of annotation is necessary for natural language processing—whether conceptual metaphor annotation is necessary at all, since natural language processing primarily interprets textual data. Shutova concludes that it is not necessary to bother with annotating the source and target domains, but rather that it is enough to annotate the linguistic metaphor.

In the same paper, Shutova also presents interesting findings regarding conventional metaphors—it was precisely these that created the biggest disagreements among annotators, whether a conventional metaphor is actually a metaphor, since it has become completely common in speech/language. Shutova also suggests that, in some future research, the conventional metaphor be annotated on a scale from, for example, "completely literal" to "completely metaphorical".

These observations are important in the context of natural language processing—if something is entirely conventional and the speakers themselves do not consider it a metaphor, the question is how large language models can "understand" that it is a metaphor, and whether conventional metaphor annotation can potentially confuse machine learning models that use large language models.

Undoubtedly, it is precisely linguistics, together with related fields of the humanities, that shows us what a metaphor is, how it occurs and what role it plays in speech and in society. If we annotate a metaphor for the purposes of natural language processing, it is necessary to keep in mind why we identify a metaphor in the first place and what our ultimate goal is—is the goal just to identify a metaphor or are we doing machine translation, because in that case we need to know how to translate the identified metaphor into some another language (In Croft and Cruse [3], we find an interesting observation that, although not all novel metaphors become conventional over time, those that do, mostly become conventional in many languages). If, on the other hand, we are doing a sentiment analysis, we have to understand the sentiment of the metaphor, whether it is positive or negative, while in conversational agents we have to fully understand the metaphor in order to be able to conduct a conversation (cf. [14]).

4. Identifying Metaphors

Depending on the approach used in metaphor identification, systems and models for machine metaphor identification can be divided into 3 main groups (The division is inspired by the division given by Shutova [14], with the addition of neural network approaches and vector representations of words. In her paper, Shutova divides the methods of identifying linguistic metaphors into (1) Approaches using hand-coded knowledge and lexical resources, (2) Statistical learning, (3) Metaphor identification as a Classification Problem):

1. Hand coded knowledge and use of lexical resources
2. Statistical and machine learning methods
3. Neural networks and word embeddings

Early approaches to metaphor identification were based on hand coded knowledge and dictionary creation, which is a challenging, time-consuming task, and it is questionable whether it is even possible to annotate and list all metaphors in a language, bearing in

mind that novel metaphors appear every day and that language is a living organism that is constantly changing. With statistical and machine learning methods, the problem of metaphor identification is approached as a problem of supervised [22,23,28,29] or unsupervised learning [22,30,31], while some authors combine both approaches [32]. These approaches use traditional statistical methods, and some of the models also use static word embedding (e.g., Ref. [23]).

Approaches with neural networks also use, in addition to neural networks, some of the word embeddings, being static [33–40] or contextual word embeddings [41–51] (Brief overview of word embeddings is presented in the section *Neural networks and word embeddings*). It is these models that have given the best results in identifying metaphors in recent years.

Below is a more detailed overview of the above systems and models. In the Appendix A, a table with an overview of metaphor identification research, organized by the year in which each of the research was published, is presented.

4.1. Hand Coded Knowledge and Use of Lexical Resources

In their paper, "The structure of a semantic theory" [52], Katz and Fodor when discussing the form of semantic theory and emphasizing the importance of semantics in the very understanding of language, introduce the theory of *selectional restrictions*. A selectional restriction is a restriction in the relations between words in a sentence, particularly the predicate and its associated arguments. In his papers from 1975 [53] and 1978 [54], Wilks, one of the pioneers of natural language processing, notes that rather than being restrictions, these are preferences, and introduces the theory of *selectional preferences*.

It is the theories of Katz and Fodor and Wilks that form the basis of the Fass's [55] system for machine metaphor recognition, met* (read as *met star*), and inspired most other early approaches to machine metaphor identification. In his paper, Fass provides an overview of previous research dealing with metaphor identification and proposes that they can be divided into four different approaches—(1) metaphor as comparison, (2) metaphor as an interaction, (3) violation of selectional restrictions/preferences and, finally, (4) the conventional metaphor. He bases his met* system on selectional restrictions/preferences, and finds, like Wilks, that the metaphor violates selectional restrictions/preferences and can, therefore, be recognized. However, Fass does not ignore other approaches to identifying a metaphor, and his met* system also takes into account the importance of other approaches, which are implemented in the system. The met* system recognizes metonymy in addition to metaphors.

Wilks himself continued to work on the theory of selectional preferences to develop a method of identifying metaphors based on that theory. In Wilks et al. [56], the authors present an algorithm for metaphor recognition using WordNet as a lexical resource. The hypothesis of the paper is that we have a metaphor on our hands if the first meaning that a word has in WordNet is not the one that would correspond to the selectional preferences that are needed in the place in the sentence where the word is found, but has some less common meaning.

The CorMet system [57], proposed by Mason, also finds its foundations in selectional preferences and recognizes cross-domain mapping. CorMet analyzes a large number of documents, organized by domain, and learns the selectional preferences of verbs that are characteristic for a certain domain. By discovering the selectional preferences of verbs in one domain, the CorMet system can also recognize differences in selectional preferences between domains—for example, the CorMet system recognizes the difference in the selectional preferences of verbs used in the finance domain from those used in the laboratory domain (an example given by the author is "Funds poured into his bank account."), and can conclude that this is a cross-domain mapping. The system learns selectional preferences in two steps—the first step is Resnik's [58] algorithm for learning selectional preferences in order for CorMet to learn the selectional preferences of verbs in a certain domain, and in

the second step WordNet nodes are clustered using the nearest neighbor method in order to compare selectional preferences and determine similarities.

Krishnakumaran and Zhu [59] limit their work to metaphors containing nouns and divide metaphors into three types, depending on the connections in the sentence. Type I is the relationship between subject and object, type II is the relationship between verb and noun, and type III is the relationship between adjective and noun. To identify a type I metaphor, the authors use WordNet to identify whether there is a hyponym relationship between the subject and the object, that is the words used to express the subject and the object, and if such a connection does not exist, the sentence is considered a metaphor. When identifying type II and type III metaphors, the same method is used, whereby in the first case, the relationship between verbs and nouns is observed, and in the second case, the relationship between adjectives and nouns is observed. In this method, the authors calculate the probability of occurrence of a noun and its hyponyms and hypernyms in a pair with a verb or adjective. If the probability or frequency of occurrence is low, the verb-noun or adjective-noun pair is annotated as a metaphor. The authors test their method on the Master Metaphor List, which they expand for research purposes by completing phrases and sentences and correcting mistakes and typos. For the type I metaphor, they achieve an accuracy of 0.58, a precision of 0.70 and a response of 0.61.

4.2. Statistical and Machine Learning Methods

In approaches to statistical and machine learning methods, we find two main directions—classification, i.e., supervised learning, and clustering, i.e., unsupervised learning, while in some papers the authors combine these two approaches. A brief overview of the models that utilize statistical and machine learning methods is given below, while in the Table 2 a summary of the models is presented, showing the data sets and methodology used.

Based on the word sense disambiguation algorithm published by Karov and Edelman [60], Birke and Sarkar [22] developed a TroFi system for distinguishing between literal and non-literal sentences. The authors emphasize that TroFi is not a system for metaphor (or metonymy) processing, but despite this, we mention it here because TroFi inspired later papers dealing with metaphor identification.

TroFi is one of the first systems that approaches the problem of distinguishing between literal and non-literal text using a statistical method. To be more precise, it clusters literal and non-literal sentences by determining whether verbs are used literally or not. The paper achieves an F1 score of 0.538, but the result of the paper, apart from the algorithm itself, is also a TroFi data set, which contains sentences in which 50 selected verbs are used, and these sentences are in clusters—literal or non-literal.

Building on the work of Birke and Sakar, Turney et al. [28], using the TroFi data set to analyze metaphors expressed by verbs, create an additional data set based on adjectives, and hypothesize that the degree of abstractness of the word in context is related to the probability of how the word is used metaphorically. Turney et al. represent an algorithm for calculating the abstractness of words in relation to the nouns they are related to, based on latent semantic analysis. For the classification of literal and metaphorical words, the paper uses a logistic regression, and the authors achieve an F1 score of 0.68 on the TroFi data set, which is a better result than in the original research.

The Concrete Category Overlap (CCO) algorithm was developed by Neuman et al. [61], who took the work of Turney et al. as their starting point. In the creation of the algorithm, they relied on the assumption that a metaphor is a mapping from the concrete to the abstract domain. The CCO algorithm is further expanded by Wilks' idea of selectional preferences, i.e., by identifying whether a violation of selectional preferences has occurred. The focus of the paper is on metaphors that contain a noun, which the authors base on the work of Krishnakumaran and Zhu [59] and take over their categorization of three types of metaphors. The method presented by Neuman et al. was evaluated on a small set of data—the authors took five concepts—governance, government, God, mother, father—which they extracted

from Reuters RCVI dataset and the New York Times archive and annotated metaphors. Their system achieves a precision of 0.72 and a recall of 0.80.

In their paper, Tsvetkov et al. [23] also used the TroFi data set (to identify metaphors expressed by verbs), and for the purposes of the research created an additional data set (TSV) that contains 884 metaphorical expressions where the metaphor is expressed by an adjective and a corresponding noun. On the basis of previous work, these authors [62], present the hypothesis that metaphors are primarily conceptual and only then lexical, which they demonstrate by training the model on an English data set, but the same model with a high F1 score identifies metaphors in Russian (0.84 for metaphors expressed by verbs, 0.77 for the set of adjectives and nouns), Spanish (0.76 for metaphors expressed by verbs, 0.72 for the set of adjectives and nouns) and Farsi (0.75 for metaphors expressed by verbs, 0.74 for the set of adjectives and nouns). Tsvetkov et al. use the random forest method for the classification, defining three main categories of features: (1) abstractness and imageability, inspired by previous papers such as Turney et al. [28]; (2) supersenses categories from WordNet and (3) word embedding, which makes this one of the first papers on metaphor identification in which word embedding is used. For the TroFi data set, that is, for the metaphor expressed by verbs, Tsvetkov et al. achieved an accuracy of 0.82 and an F1 score of 0.79, while for the data set created as part of the paper and where the metaphor is expressed by an adjective, they achieved an accuracy of 0.86 and an F1 score of 0.85.

Following their previous work [63], Li and Sporleder [32] approach the distinction between literal and non-literal phrases by combining unsupervised and supervised learning, where the support vector machine is used for supervised learning and classification. In the first step, unsupervised learning on the basis of lexical cohesion is used to classify literal versus non-literal expressions, and in this way an annotated data set is created, which is submitted to the classification model in the second step. The dataset consists of 17 phrases for which a five-paragraph context has been extracted from the Gigaword corpus, to help determine whether the phrase is being used literally or not. In the paper, the authors state that their model achieves an accuracy of 0.9 in distinguishing literal and non-literal phrases.

An approach to metaphor identification using the method of unsupervised learning, i.e., clustering can also be found at Shutova et al. [30]. They start with a small seed set of data and, by clustering nouns, create a set of data that belongs to the same domain, based on the hypothesis of "clustering by association"—abstract concepts will be clustered into the same clusters if they are connected to the same domain, while concrete concepts will be clustered according to similarity in meaning. They use the subset of the British National Corpus, in which the metaphor is annotated, as a seed set while evaluating the method itself on the entire British National Corpus. They compared the results with the annotations made by 5 annotators and concluded that their method achieved a precision of 0.79.

Table 2. Overview of metaphor identification models that use statistical and machine learning methods.

Paper	Data Set	Methodology	Results
Birke and Sarkar [22]	TroFi	Clustering	F1 0.538
Turney et al. [28]	TroFi	Calculating abstractness, classification	F1 0.68
Neuman et al. [61]	Five concepts extracted from the Reuters and New Yor Times archive	Concrete Category Overlap (CCO) algorithm, Selectional preferences	Precision 0.72
Tsvetkov et al. [23]	TroFi, TSV	Classification, word embedding	F1 for TroFi 0.79, F1 for TSV 0.85

Table 2. *Cont.*

Paper	Data Set	Methodology	Results
Li and Sporleder [32]	17 phrases extracted from Gigaword	Clustering, classification	Accuracy 0.9
Shutova [30]	Subset of the British National Corpus	Clustering (by association)	Precision 0.79
Heintz et al. [31]	60 concepts from newspapers and blogs in English and Spanish	Topic modeling with Latent Dirichlet Allocation	Evaluation 1: F1 score 0.59, Evaluation 2: Average value of 0.41 for English, 0.33 for Spanish
Mohler et al. [29]	Domain of governance extracted from Wikipedia	TF-IDF, classification	Precision 0.561, F1 score 0.70

Another unsupervised learning approach is that by Heintz et al. [31]. In their paper, the authors use the Latent Dirichlet Allocation (LDA) method for topic modeling and identify metaphors in English and Spanish language texts from the governance domain. The texts were taken from newspaper portals and blogs. The hypothesis of the paper is that metaphors are contained in those sentences that contain both the original and the target concept. In order to confirm the hypothesis, the authors create a list of 60 original concepts, namely primary concepts. Their algorithm contains four steps, (1) the system architecture assigns a topic to the sentences and documents in the input corpus, (2) the words in the sentences are connected to the concepts, (3) the sentences that contain too few words and sentences whose topic is also a highly ranked topic of the entire document are thrown out, (4) based on the results of the algorithms in the previous steps, the final result is calculated. In the first evaluation method, for each concept, 5 examples are selected from the English set, and two annotators evaluate whether the examples contain a metaphor related to governance. They achieve an average F1 score of 0.59, with a Kappa agreement of 0.48. The second evaluation approach, which the authors call stricter, selects the 250 highest-ranked linguistic metaphors for each language, which are evaluated on Amazon Mechanical Turk by people who claim that English or Spanish is their mother tongue. The task was to answer the question whether a word is a metaphor with a simple yes or no. The result was an average value of 0.41 for English (with a standard deviation of 0.33) and 0.33 for Spanish (standard deviation of 0.23).

Classification preceded by a step aimed at determining the Target Domain Signature was the approach proposed by Mohler et al. [29]. The authors chose for their research the domain of governance and defined words related to that domain (e.g., law, government). In WordNet, they found the meanings associated with those words or domains in order to use them to search for articles on Wikipedia related to that domain. The authors calculate the semantic signature of the target domain, using the clustering method and the TF-IDF weighting scheme. The support vector machine classifier gives the highest precision (0.127) and F1 score (0.184) when applied to a test set containing 241 positive and 3800 negative examples, but when applied to a set containing an equal number of positive and negative examples, the results are a precision of 0.75 and an F1 score of 0.464 using the support vector machine, while the best result is obtained by the decision tree: a precision of 0.561 and an F1 score of 0.70.

4.3. Neural Networks and Word Embeddings

The word embedding has its roots in the method of vector semantics, which has been used in natural language processing since the 1950s to represent the meaning of words. The idea of vector semantics is that each word can be represented as a point in

a multidimensional semantic space that is derived from the distribution of neighboring words [64].

Word embeddings can be divided into static (e.g., word2vec [65,66], GloVe [67]) and contextual (e.g., ELMo [68], BERT [69], RoBERTa [70]), whereby the latter are also called Large Language Models (LLM) and are based on the Transformer neural network [71]. While static word embeddings show only one, static, representation for each individual word, contextual word embeddings enable the representation of words in all contexts in which that word appears in the vocabulary, i.e., the data used when training the model.

We will divide the overview of metaphor identification papers, which approach the creation of models using neural networks, into two groups, depending on whether static or contextual word embeddings are used. Prior to the overview of the papers, we will give a brief overview of relevant (We consider those word embeddings used in previous papers on metaphor identification to be relevant. The overview does not cover all the major word embeddings published up to the time of writing this paper, nor does it aim to provide an overview of all models) static and contextual word embeddings, i.e., large language models.

4.3.1. Overview of Static Word Embeddings

One of the first word embeddings is the word2vec package, introduced by Mikolov et al. [65]. In the paper, the authors propose simple models that do not use neural networks, thus minimizing computational complexity. The two proposed models are CBOW (Continuous Bag-of-Words) and the skip-gram model. The CBOW architecture predicts the current word based on the context, while the skip-gram model predicts, based on the current word, which words can occur before or after it. The skip-gram model gave excellent results, and in Mikolov et al. [66] the authors build on it and propose the use of the negative sampling method when training models.

The GloVe (short for Global Vector) model is based on capturing global corpus statistics [67]. The two main methods that GloVe combines are the global matrix factorization and the local context window like skip-gram. This approach proved to be good, since GloVe achieves better results than word2vec, either in the CBOW or Skip-gram architecture.

4.3.2. Overview of Contextual Word Embeddings

The first contextual word embedding, ELMo (Embeddings from Language Models), was presented in Peters et al. [68]. In its architecture, ELMo takes over and modifies biLM (Bidirectional Language Model), presented in Jozefowicz et al. [72]. biLM consists of LSTM (Long Short Term Memory) [73] layers and a CNN (Convolutional Neural Network) [74] layer in which character convolution is performed, and ELMo enables training in both directions while it receives the entire sentence as input and learns the word representation depending on the context. The authors tested ELMo on a series of natural language processing tasks and achieved significantly better results than the previous best results in 6 natural language processing tasks.

The BERT (Bidirectional Encoder Representation from Transformers) language model [69] is the first word embedding trained on the encoder component of the Transformer neural network. Besides the fact that the BERT architecture is bidirectional, an important component is also the Masked Language Model (MLM), which masks or hides, by random selection, some of the tokens in the input data, with the aim of predicting the masked token based on the context.

BERT was trained in two sizes, the basic BERTBASE and the large BERTLARGE, which differ in the number of layers with Transformer blocks (L), the number of hidden layers (H) and the number of attention heads (A), which ultimately results in a difference in the number of parameters (BERTBASE—L = 12, H = 768, A = 12, Parameters = 110 M, BERTLARGE—L = 24, H = 1024, A = 16, Parameters = 340 M). With the BERT language model, better results were achieved than in previous papers in eleven natural language processing tasks, with BERT representing the best achievement at the time of publication.

BERT was also trained in a multilingual variant (https://github.com/google-research/bert/blob/master/multilingual.md; last accessed on 2 February 2023), mBERT, which includes 104 languages and is the *base* model in terms of size.

The RoBERTa [70] language model, as the authors state, is a proposal to improve BERT by correcting four areas that the authors believe are wrong in the original BERT architecture, i.e., training process—(1) longer training of the model and with larger batches, (2) removing the goal of predicting the next sentence, (3) training on longer sequences, and (4) dynamically changing the token masking pattern in the training data. An additional difference is the dataset used for training, where the authors use an additional 5 datasets in English, a total of 160GB of uncompressed text data. The RoBERTa model achieves better results than the BERT model in three tasks (GLUE, RACE and SQuAD).

While the RoBERTa model itself is not multilingual, XLM-RoBERTa [75], abbreviated XML-R, is its multilingual variant, trained on texts in 100 languages for which a corpus called CC-100 is created. The authors report that the XLM-R model gives significantly better results than the mBERT model, up to 23% improved accuracy for low-resource languages. In its architecture, the XLM-R model follows the XLM (cross-language model) presented in Lample and Conneau [76] and is available in two sizes (following the two sizes of the BERT model), XLM-RBASE and XLM-R.

4.4. Metaphor Identification with the Use of Word Embeddings

While word embeddings are the backbone of contemporary research in metaphor identification, we find it insightful also to differentiate the research by the approach taken.

Approaching the metaphor identification as the sequence labeling task is most common in early research with the use of static word embeddings [33,36–38], while there are also other approaches, e.g., enriching word embeddings with visual ones [34]. Recent research that utilizes contextual word embeddings are also experimenting with different approaches, e.g., reading comprehension task [41], learning from another type of figurative language [43], looking at the broader discourse [44], contrastive learning [45], or as relation extraction problem [51]. Lately, we notice that researchers are also trying to make the most of linguistic theories of metaphor (namely, MIP, Sectional Preferences and Conceptual Metaphor Theory) by implementing them in the architecture of their proposed models [40,46–50] and it is these models that present the current state of the art in the field of metaphor identification.

Below a brief overview of each of the research and methods in identifying metaphor with the use of word embeddings is given, while at the end of this section, in Table 3 a summary of all these approaches is presented.

4.4.1. Metaphor Identification Using Static Word Embeddings

One of the first papers that uses neural networks and word embedding to identify metaphors is by Do Dinh and Gurevych [33]. The problem of identifying a metaphor is approached as a problem of tagging i.e., sequence labeling. The authors use a multilayer perceptron in conjunction with static word embedding, word2vec. The data set for training and testing is VUA, but the authors use only those metaphors annotated as MRW while other metaphor type tags are annotated as literal words/expressions. In the research, the authors conduct three experiments—identification only at the token level, identification at the token level with information on the type of word (Part of Speech, POS), identification at the token level with information on the type of word and word concreteness [77], and conclude that adding information about the word type does not significantly affect the result (increase in F1 score for texts from newspaper articles from 0.6352 to 0.6385) nor does adding information about word concreteness (increase in F1 score for academic texts from 0.5579 to 0.5618 when POS and word concreteness are used).

Enriching linguistic word embeddings with visual ones can provide better results in identifying metaphor, as it was shown by Shutova et al. [34]. Using the entire Wikipedia available at the time, the authors created word embeddings using the skip-gram method.

Words were previously lemmatized (Lemmatization is a technique which reduces inflected words to their root word, which is called lemma), and words that appear less than 100 times were ignored. The creation of word embeddings is done in two phases, where in the first, an embedding of individual words is created, and in the second, an embedding of phrases. To create visual embeddings, a convolutional neural network is used, trained on the ImageNet classification task, and up to 10 images are downloaded from Google for each word or phrase. The final visual representation is made by averaging the extracted vector image representation for each word or phrase. The metaphorical quality of words is determined by calculating the similarity of the words and by defining whether they belong to the same domain—if the words do not belong to the same domain and the similarity between them is small, it is assumed that it is a metaphor. The cosine similarity method is used for the calculation. In order to determine whether a phrase is metaphorical, the vector representation of the phrase is compared with the individual words of that phrase, and if the vector representation of the phrase is similar to that of the individual word, it is assumed that it is a literal expression or phrase. The method was evaluated on the MOH-X and TSV data sets, and the authors achieved the best results with the multimodal model, combining linguistic and visual embeddings (F1 score for MOH-X 0.75 and for TSV 0.79), concluding that multimodal embeddings can further improve metaphor identification results.

Relying on the cosine similarity method used by Shutova et al. [34], Rei et al. [35] expanded the method by showing words in dependence on other words, thus providing context. When creating word embeddings using the skip-gram method, they use a data set with an annotated metaphor. Based on the assumption that not all dimensions of the vector representation are equally important when identifying a metaphor, the cosine similarity is expanded by adding a weighting factor, and ultimately the prediction is based on the vectors obtained in this manner. They evaluate the method on the MOH-X and TSV. In the evaluation, they use two methods of word representation—the skip-gram method and vectors based on attributes [78], while the best results are achieved by combining both methods of word representation with the supervised similarity method proposed in the paper—F1 0.811 for TSV and 0.699 for MOH-X. Since there is a difference in the size of the data sets, and the results are better on a larger data set (TSV), in order to determine whether the size of the data set affects the result, they did an additional study using the third data set, GUT, which was created by Gutiérrez et al. [79] and expand the TSV data set with a new set. They found that up to 2000 examples there is a large increase in model performance, whereas after that the increase in performance is much lower.

Research on the performance of the Recurrent Neural Networks (RNN) in the architecture of the Gated Recurrent Unit (GRU) and the feedback network in the LSTM architecture was presented by Mykowiecka et al. [36]. The authors approach metaphor identification as a sequence labeling task. Both neural networks were implemented in a two-way architecture since, as the authors state, useful information about a word can be encoded both on its left and on its right side, that is its context. For the word embeddings, the authors use GloVe and supplement the input information with POS and information from the General Inquirer dictionary [80] on whether the word belongs to one of the categories defined in the dictionary. The results, similar to those of Do Dinh and Gurevych [33], demonstrated that adding POS information does not significantly affect the result, and the same was shown for the information from the General Inquirer—the change is 0.01 for the F1 score on the training data set, while on the test data set, the addition of this information had a negative effect on the result. The research showed that LSTM provides better results than GRU architecture and that the number of layers of the network as well as the number of epochs have a significant influence on the result. The best F1 score on the VUA dataset of 0.583 for all types of words, i.e., F1 of 0.619 for a metaphor expressed by a verb, was achieved with LSTM.

LSTM was also used by Swarnkar and Singh [37] but the architecture of their system, which they call DiLSTM Contrast, has GloVe word embeddings in the first step. This is followed by a context encoder in two layers of the LSTM where the first layer observes at

the data from the front and the second from the back, which is inspired by the bidirectional LSTM architecture. After the context encoder, the system selects features and performs the classification in the final step. This paper also uses the VUA data set and approaches the metaphor identification as a sequence labeling task. On the test set the overall F1 result was 0.605 for verbs, and 0.570 for all types of words while the best result was achieved in the classification of academic texts and newspaper articles. As additional information, the authors provide POS, word meanings from WordNet and concreteness of words, which together improves the F1 score by 0.8%, but we do not learn from the paper how each of these additional linguistic information aspects affects the result separately.

Approaching the problem of identifying metaphors as sequence labeling, Wu et al. [38] developed a model that uses both LSTM and CNN. In the first step, the architecture of their model lemmatizes the input text, after which, using the word2vec word embeddings, the previously lemmatized words are placed into the vector space. This is followed by a convolutional layer that extracts local contextual information, after which the data passes into a bidirectional LSTM layer. The last step is the inference layer where the authors experiment with two approaches to compare the results—one is the Conditional Random Field (CRF), which they use to predict a metaphor for each individual word, and the second is a softmax activation function, which they use to predict a metaphor on sequences. In the paper, the authors conclude that combining CNN and LSTM provides better results than each of these networks separately, and that CRF in combination with the indicated neural networks gives better precision, but softmax gives a better response and F1 score. Lemmatization as well as adding POS information and word clusters improves the result of identifying the metaphor. The research uses the VUA data set and the authors report the best results with the combination of CNN, LSTM, softmax and the ensemble approach: an F1 score of 0.671 for verbs and 0.651 for all types of words.

The papers by Mykowiecka et al. [36], Swarnkar and Singh [37] and Wu et al. [38] were presented as part of the first workshop on figurative language processing, held in 2018 in New Orleans, as part of the NAACL conference. In addition to the above three papers, more papers were presented as part of the workshop, but these three achieved the best results. For an overview of other papers, and the approach to identifying metaphors in those papers, we suggest looking at the workshop report, which is publicly available (The report is available at: https://aclanthology.org/W18-0907.pdf; last accessed on 18 February) [81]. The second workshop on the processing of figurative language was held in 2020 (online) as part of the ACL conference, and the best three results were achieved by the papers of Su et al. [41], Chen et al. [43] and Gong et al. [42]. We give an overview of these papers in the next chapter, which provides an overview of papers that use contextual word embeddings. At the same workshop, Dankers et al. [44] presented their paper, which did not achieve the best results, but we include it in the overview due to its interesting approach of taking into consideration the wider discourse. The report from the second workshop is also publicly available, and an overview of all papers from the workshop can be found in it [82] (The report is available at: https://aclanthology.org/2020.figlang-1.3v2.pdf; last accessed on 18 February 2023).

4.4.2. Metaphor Identification Using Contextual Word Embeddings

To the best of our knowledge, first research on metaphor identification using a contextual word embedding is that of Gao et al. [39] in which authors use ELMo word embedding. The authors approach the identification in two ways—by sequence labeling where each word is annotated either as a metaphor or as a literal use, while the second approach is a classification annotating only the verb in the sentence as a metaphor or a literal use. In both cases, the authors use a bidirectional LSTM network and add an attention mechanism layer to the classification model. They evaluate the models on three existing datasets—TroFi, MOH-X and VUA. The use of the ELMo contextual word embeddings in ablation study, as the authors state, significantly affects the results of both models, which are better when ELMo is used. It has a particularly positive impact on the sequence labeling model, where

in the evaluation of the model on the VUA data set, the F1 score is 0.617 without using ELMo and 0.704 when using ELMo, while the accuracy is 0.782 without using ELMo and 0.835 with ELMo in identifying metaphors with verbs. The authors compare the results obtained by their model with the results of Wu et al. [38], whose model had at the time been the most successful, but the model of Gao et al. surpassed that of Wu et al. It should also be noted that Gao et al. feed their model data, i.e., text, that had previously been lemmatized, and they also feed it information about the type of words, which Wu et al. had demonstrated has a positive influence on the results.

Furthermore, Mao et al. [40] presented the hypothesis that with the use of linguistic metaphor theories, when designing the architecture of deep neural network, one can improve the results of metaphor identification. The two linguistic methods on which they base their models are MIP and selectional preferences (SPV). The MIP-based model (RNN_HG) approaches metaphor classification based on the difference between the literal meaning of the word and the contextual one. For the representation of the contextual meaning, the model uses a bidirectional LSTM neural network (BiLSTM), while for the representation of the literal meaning, GloVe and ELMo word embeddings are used. The SPV-based model (RNN_MHCA) compares the target word with the context, using BiLSTM to represent the target word, while the attention mechanism is used for the context. The models were evaluated on three datasets, VUA, MOH-X and TroFi, and the results were compared with those of Wu et al. [38] and Gao et al. [39]. As an additional model, they modified the approach of Gao et al. by using BERT instead of ELMo. The RNN_MHCA model achieved a better F1 score on the MOH-X (0.80), TroFi (0.724) and VUA dataset with all types of words (0.743), while RNN_HG achieved better results on the VUA dataset containing only verbs (0.708).

That approaching the metaphor identification as a reading comprehension problem is feasible and that this gives good results, was shown by Su et al. [41]. The authors also show that information about the global and local context, as well as POS information, contribute to a better result. The reading comprehension paradigm requires three sentences, a word that needs to be understood, and an indication of whether the word is a metaphor. The model uses BERT in the embedding layer, while the Transformer neural network in two layers is used as the basis of the model. One layer is fed the global text context, the searched word, POS and FGPOS at the input while the other layer is fed the local text context, the searched word, POS, FGPOS, and the properties are separated by a special separation token [SEP]. In the training process, the RoBERTa word embedding is used, with the aim of fine-tuning the Transformer layers. The model was tested on 4 data sets—VUA, TOEFL, MOH-X and TroFi, and for all data sets (except TOEFL, which was not used in previous research at the time), a better result was achieved than in previous papers: an F1 score of 0.804 (VUA Verbs), 0.769 (VUA All Words), 0.749 (TOEFL Verbs), 0.715 (TOEFL All Word Types), 0.918 (MOH-X) and 0.761 (TroFi).

Gong et al. [42] also approach metaphor identification as a problem of sequence labeling, and classify each word in a sentence as a metaphor or a non-metaphor. Recognizing that contextual information is crucial in metaphor recognition, they base their model on the RoBERTa contextual word embedding. At the input, the model is fed linguistic properties (POS, word distribution in topics, word concreteness, WordNet, VerbNet and properties from the corpus), since previous papers have shown how useful these are in identifying metaphors. The classification itself is performed with a feed forward network. The model was tested on the VUA and TOEFL dataset. The best F1 result of 0.771 was achieved for VUA verbs (RoBERTa in ensemble architecture) and 0.719 for TOEFL verbs (RoBERTa in ensemble with all linguistic properties).

Using data from another domain and learning from another type of figurative language form the approach taken by Chen et al. [43]. Using the BERT fine-tuning method, in the domain learning experiment, they combine the training sets from TOEFL and VUA data sets into one, while in the second type of figurative language learning, they compile sets of phrases based on the idea that phrases are often metaphors. The authors compare the

results of the experiments with the results obtained in the system based on BERT, which they develop themselves, and achieved an F1 of 0.718 for VUA all types of words, 0.756 for VUA verbs, 0.624 for TOEFL all types of words and 0.657 for TOEFL verbs. They achieved the best F1 score with the VUA data set in the experiment with phrases (0.775 VUA verbs and 0.734 VUA all types of words), while for TOEFL the best results were achieved in the learning experiment from another domain (0.702 TOEFL verbs, 0.692 TOEFL all types of words).

Dankers et al. [44] based their model on the thesis that information about the broader discourse of the sentence in which the metaphor is identified contributes to a greater success in identifying the metaphor. The authors build an architecture that incorporates the wider discourse, by using a context window of $2k + 1$ sentences before and after the sentence that is the focus of identifying the metaphor. They experimented with two different architectures—in the first, GloVe and ELMo are in the embedding layer, and BiLSTM in the encoding layer, and in the second the BERT model is in the embedding and the encoding layer. Also, in the first architecture, the final layer uses a General Attention mechanism creating a discourse by applying the attention mechanism on all tokens, while the second one uses Hierarchical Attention, combining the attention mechanism at the level of words and sentences. The experiment uses the VUA data set, and the best result, an F1 score of 0.715, was achieved on the architecture with BERT and hierarchical attention, with a context window of $k = 2$ sentences before and after the sentence in which the metaphor is identified.

The Contrastive Learning approach in the CATE model [45], proposed by Lin et al., is designed in order for the model to learn the difference between the literal and metaphorical use of a word, based on their distance in the vector space representation. In the paper, the authors also propose a method called Target-Based Generating Strategy (TGS), which addresses the issue of the lack of labeled data sets for identifying metaphors. The TGS method is based on the thesis that if a word is targeted as a potential metaphor, all other sentences containing that word are potential candidates. In order to create pseudo-tags for these sentences, they use a self-learning method. In the experimental system, VUA, MOH-X and TroFi datasets were used, while RoBERTa was used for the encoder and contextual word display. The achieved results were an F1 score of 0.790 for VUA all types of words, 0.756 for VUA verbs, 0.847 for MOH-X and 0.745 for TroFi.

Metaphor-relation BERT (MrBERT) [51] model, proposed by Song et al., is focusing on token level verb metaphor detection, by approaching the task as a relation extraction problem. The whole sentence is viewed as a global context, while local context is represented by the words that have close grammatical relation to the target word. Authors, motivated by MIP, see distant context as a basic meaning of the word. These three contexts are then extracted and represented. Final transformer layer is used to get the contextual representation while output from BERT tokenizer is used to get the context independent representation. Contextual relation, relation between the target verb and one of its contexts, is utilized to determine the metaphoricity of the verb. The authors evaluate their model on VUA verb and report F1 score of 0.759, while VUA all F1 score is 0.772.

An additional model based on the linguistic theories, MIP and SPV, is MelBERT, proposed by Choi et al. [46]. Architecture of the MelBERT adopts RoBERTa as a contextualized backbone. On input, token, position, and segment embedding are fed to transformer encoder, on top of which MIP and SPV layers are implemented. MIP layer is fed with two embedding vectors representing contextualized embedding vector and isolated embedding vector for the target word. This approach allows for identification of the semantic gap for the target word, in context and isolation. SPV layer, on other hand, is fed only with the sentence encoder, with the assumption that there is a semantic gap between contextual embedding vector of word, and word in a given context. Using these two strategies, hidden vectors hMIP and hSVP are computed and combined to get the prediction score. Finally, the cross-entropy loss function for binary classification is used. For the evaluation, VUA-18 (data set used on first workshop on figurative language processing [81]) and VUA-20 (data

set used on second workshop on figurative language processing [82]) data sets are used, so that results can be compared to the previous research [39–41]. The authors report the F1 score of 0.757 for VUA verbs, 0.785 on VUA-18 and 0.723 on VUA-20.

MIP has also inspired Maudslay and Teufel [50] in proposing the model for conventional metaphor identification with the word sense disambiguation method (WSD), called Metaphorical Polysemy Detection (MPD). They state that detecting novel and conventional metaphors are fundamentally different tasks and argue that metaphoricity is best treated as a word sense in lexicon. Furthermore, they argue that step 1 in MIP is very similar to word sense disambiguation. They build a model which learns to identify metaphorical senses in WordNet and for evaluation manually annotate 250 commonly cited metaphor examples from Master Metaphor List (MML). The same authors also propose a relative metaphoricity measure, with the goal to improve usual measurement (F1, precision and recall) which only gives absolute judgment, metaphorical-literal binary, while metaphor often has edge cases. Relative metaphoricity is calculated using ROC-AUC and allows evaluation whether a model can judge which word senses are more metaphorical than others. Architecture of MPD is a multi-layer perceptron (MLP) which is given as input pair of embeddings for word sense tuple, consisting of wordform and definition. The authors experiment with two WSD models—BERT WSD as a baseline and EWISER [83] as a state-of-the-art approach. BERT WSD model achieves absolute F1 measure of 0.60 for MML, 0.41 for out of the vocabulary words of the training data (OOV), and 0.47 for verbs. Furthermore, the relative metaphoricity measure in their research was 0.78 on MML and 0.64 on OOV with MPD (EWISER WSD), while for verbs was 0.71 with MPD (BERT WSD Baseline).

MIP and SPV are also found as the backbone of MisNet method [49], proposed by Zhang et al. These two linguistic theories are combined in MisNet architecture. MIP module is used for semantic matching of the target word and the basic usage, while the goal of SPV module is to measure semantic incongruity of the target word in context. These authors emphasize that SPV can be invalid in the case of conventional metaphor as those do not have paradoxical context. In order to combine MIP and SPV, siamese framework is adopted (similar to MelBERT), where left part encodes the given sentence and right part uses the target, POS and basic usage. MIP is implemented on both, left and right encoders, while SPV is implemented only on the left one. The proposed model is validated on VUA All, VUA Verbs and MOH-X datasets and the authors report F1 score of 0.794 on VUA all, 0.759 on VUA Verb and 0.834 on MOH-X.

FrameBERT [47] is RoBERTa based model that also utilizes MIP and SPV, whose authors, Li et al., hypothesized that using external knowledge of concepts is essential for improving not just metaphor identification but also explainability and to support this, they use FrameNet (FrameNet is a lexical database, accessible at http://framenet.icsi.berkeley.edu/; last accessed on 13 May 2023). The architecture of the FrameBERT incorporate FrameNet embeddings for concept-level metaphor detection, while by using MIP to identify the gap between contextual and literal meaning, metaphorical word is identified. With SPV, semantical difference from surrounding words points out that word is used metaphorically. Authors report F1 score of 0.788 on VUA18 dataset, 0.73 on VUA20 dataset, 0.838 for MOH-X, and 0.742 on TroFi dataset.

Conceptual Metaphor Theory (CMT) has inspired Ge et al. [48] to propose a model that takes as input depended word pairs (verb-noun, adjective-noun) and gives as output labels indication metaphoricity of a word pair. This model consists of three steps, first being a common association acquisition where they count the frequency of each co-occurring word in Wikipedia as well as dependency of a word pair. In the second step, conceptualization, a noun and a word pair are conceptualized with the use of WordNet and statistical point knee algorithm. In the final, third step, RoBERTa and multitask learning model learn simultaneously, to identify metaphor and to generate a concept. Model is evaluated on MOH-X, TSV, and GUT datasets, where achieved F1 score for MOH-X was 0.756, TSV 0.866, and GUT 0.925.

Table 3. Overview of papers that use word embeddings to identify metaphors.

Paper	Data	Methodology	Approach	Word Embeddings	F1 Score
Do Dinh and Gurevych [33]	VUA	Multi layer perceptron	Sequence labeling	word2vec	VUA news 0.6385, VUA academic 0.5618
Shutova et al. [34]	MOH-X, TSV	Skip-gram, CNN, Cosine similarity	Enriching word embeddings with visual embeddings	Word and visual embeddings	0.75, 0.79
Rei et al. [35]	MOH-X, TSV	Cosine similarity, gating function	Supervised version of cosine similarity	Skip-gram	0.699, 0.811
Mykowiecka et al. [36]	VUA	BiLSTM	Sequence labeling	GloVe	VUA all 0.583, VUA verbs 0.619
Swarnkar and Singh [37]	VUA	LSTM	Sequence labeling	GloVe	VUA all 0.570, VUA verbs 0.605
Wu et al. [38]	VUA	BiLSTM, CNN	Sequence labeling	word2vec	VUA all 0.651, VUA verbs 0.671
Gao et al. [39]	VUA	BiLSTM	Sequence labeling, Classification of verbs	ELMo	VUA verbs classification 0.704
Mao et al. [40]	TroFi, MOH-X, VUA	BiLSTM	Implementing linguistic theories (MIP, SPV)	GloVe, ELMO	0.724, 0.80, VUA all 0.743, VUA verbs 0.708
Su et al. [41]	TroFi, MOH-X, VUA, TOEFL	Transformers	Reading comprehension	BERT, RoBERTa	0.761, 0.918, VUA all 0.769, VUA verbs 0.804, TOEFL all 0.715, TOEFL verbs 0.749
Gong et al. [42]	VUA, TOEFL	Feed Forward Network	Sequence labeling	RoBERTa	VUA verbs 0.771, TOEFL verbs 0.719

Table 3. *Cont.*

Paper	Data	Methodology	Approach	Word Embeddings	F1 Score
Chen et al. [43]	VUA, TOEFL	BERT	Learning from another type of figurative language	BERT	VUA verbs 0.775, VUA all 0.734, TOEFL all 0.692, TOEFL verbs 0.702,
Dankers et al. [44]	VUA	BERT, Hierarchical Attention	Broader discourse of the sentence	BERT	0.715
Lin et al. [45]	TroFi, MOH-X, VUA	CATE	Contrastive Learning, Classification of the target word	RoBERTa	0.745, 0.847, VUA all 0.79, VUA verbs 0.756
Song et al. [51]	VUA	Transformers, MrBERT	Relation extraction problem	BERT	VUA verbs 0.759, VUA all 0.772
Choi et al. [46]	VUA	Transformers, MelBERT	Implementing linguistic theories (MIP, SPV)	RoBERTa	VUA-18 0.785, VUA-20 0.723, VUA verbs 0.757
Maudslay and Teufel [50]	MML	Transformers, Word sense disambigua-tion, MPD	Implementing linguistic theories (MIP), Word sense disambigua-tion	BERT	0.60 Relative metaphori-city measure 0.78
Zhang and Liu [49]	VUA, MOH-X	Transformers, MisNet	Implementing linguistic theories (MIP, SPV)	BERT	VUA all 0.794, VUA Verb 0.759, MOH-X 0.834
Li et al. [47]	VUA, MOH-X, TroFi	Transformers, FrameBERT	Implementing linguistic theories (MIP, SPV), External knowlesge of concepts	RoBERTa	VUA-18 0.788, VUA-20 0.73, MOH-X 0.838, TroFi 0.742
Ge et al. [48]	TSV, MOH-X, GUT	Transformers, Dependent word pairs	Implementing linguistic theories (CMT)	RoBERTa	TSV 0.866, MOH-X 0.756, GUT 0.925

5. Conclusions

The cognitive linguistic theory of the metaphor provides the basis for understanding metaphor annotation methods, which are currently used for research in linguistics as well as in the field of natural language processing. This paper present these theories, gives an overview of relevant data sets used in metaphor identification over the last decade and present developments in relation to computer methods for metaphor identification.

By mapping the field and chronologically listing the research dealing with metaphor identification, the paper provides an insight into the research approaches but also into the development of the entire field of metaphor identification. While the first research in the field relied on hand-coded knowledge and lexical resources (which were also created manually), the current state-of-the-art methods are based on statistical methods, neural networks and large language models. These were, in turn, trained on extremely large amounts of texts, which could hardly be processed manually. Approaches with neural networks and large language models enable research on smaller data sets, which is a great advantage since annotating metaphors is a complex and time-consuming task.

Contemporary approaches with large language models used as their backbone, also take into account the linguistical theory of metaphor, namely, MIP, Selectional Preferences and Conceptual Metaphor Theory. With the use of this approach the results of metaphor identification are improved and it can be regarded as the current state-of-the-art.

It is certainly necessary to continue the research in this field to be able to better understand how and whether large language models understand metaphors. Namely, the existing research shows good results in the context of recognition accuracy, but the confirmation of its success should come from additional "manual" verification on which and what kind of metaphors are recognized or not recognized and providing explanations on that. Also, since most of the research was done on several same data sets (VUA, TSV, MOH-X, TroFi) further research is certainly needed on other data sets, including on data sets in languages other than English.

Author Contributions: Conceptualization, M.P. and J.D.; investigation, M.P.; writing—original draft preparation, M.P.; writing—review and editing, M.P.; visualization, M.P.; supervision, J.D.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Overview of metaphor identification research by the year published.

Year	Paper
1975	Wilks [53]
1978	Wilks [54]
1991	Fass [55]
2004	Mason [57]
2007	Krishnakumaran and Zhu [59]
2006	Birke and Sarkar [22]
2009	Li and Sporleder [32]
2010	Shutova [30]
2011	Turney et al. [28]

Table A1. *Cont.*

Year	Paper
2013	Neuman et al. [61], Heintz et al. [31], Mohler et al. [29], Wilks et al. [56]
2014	Tsvetkov et al. [23]
2016	Do Dinh and Gurevych [33]
2017	Rei et al. [35]
2018	Mykowiecka et al. [36], Swarnkar and Singh [37], Wu et al. [38], Gao et al. [39]
2019	Mao et al. [40]
2020	Su et al. [41], Gong et al. [42], Chen et al. [43], Dankers et al. [44]
2021	Lin et al. [45], Song et al. [51], Choi et al. [46]
2022	Maudslay and Teufel [50], Zhang and Liu [49], Ge et al. [48]
2023	Li et al. [47]

References

1. Lakoff, G.; Johnson, M. *Metaphors We Live By*; University of Chicago Press: Chicago, IL, USA, 1980.
2. Despot Štrkalj, K. Konceptualna metafora i dijakronija: O evoluciji metaforičkog uma u hrvatskom jeziku. In *Metafore Koje istražujemo: Suvremeni Uvidi u Konceptualnu Metaforu, by Mateusz-Milan Stanojević*; Srednja Europa: Zagreb, Croatia, 2014; pp. 63–89.
3. Croft, W.; Cruse, D.A. *Cognitive Linguistics*; University Press: Cambridge, UK, 2004.
4. Sulivan, K. Conceptual Metaphor. In *The Cambridge Handbook of Cognitive Linguistics*; Dancygier, B., Ed.; Cambridge University Press: Cambridge, UK, 2017; pp. 385–406.
5. Grady, J. *Foundations of Meaning: Primary Metaphors and Primary Scenes*; University of California: Berkeley, CA, USA, 1997.
6. Grady, J. Metaphor. In *The Oxford Handbook of Cognitive Linguistics, by Dirk Geeraerts and Hubert Cuyckens*; Oxford University Press: New York, NY, USA, 2010; pp. 188–213.
7. Lakoff, G.; Johnson, M. *Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought*; Basic Books: New York, NY, USA, 1999; ISBN 0465056741.
8. Johnson, C.R. Metaphor vs. Conflation in the Acquisition of Polysemy. In *Proceedings of the Cultural, Psychological and Typological Issues in Cognitive Linguistics: Selected Papers of the Bi-Annual ICLA Meeting in Albuquerque*; Hiraga, M.K., Sinha, C., Wilcox, S., Eds.; John Benjamins: Amsterdam, The Netherlands, 1999.
9. Feldman, J.; Narayanan, S. Embodied meaning in a neural theory of language. *Brain Lang.* **2004**, *89*, 385–392. [CrossRef]
10. Fauconnier, G.; Turner, M. *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities*; Basic Books: New York, NY, USA, 2002.
11. Shutova, E. Annotation of Linguistic and Conceptual Metaphor. In *Handbook of Linguistic Annotation*; Ide, N., Pustejovsky, J., Eds.; Springer Science + Business Media: Dordrecht, The Netherlands, 2017; pp. 1073–1100.
12. Cameron, L. *Metaphor in Educational Discourse*; Continuum: London, UK; New York, NY, USA, 2003.
13. Shutova, E.; Teufel, S. Metaphor Corpus Annotated for Source—Target Domain Mappings. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*; European Language Resources Association (ELRA): Valletta, Malta, 2010.
14. Shutova, E. Design and Evaluation of Metaphor Processing Systems. *Comput. Linguist.* **2015**, *41*, 579–623. [CrossRef]
15. Pragglejaz Group. MIP: A Method for Identifying Metaphorically Used Words in Discourse. *Metaphor. Symb.* **2007**, *22*, 1–39. [CrossRef]
16. Steen, G.J.; Dorst, A.G.; Herrmann, J.B.; Kaal, A.A.; Krennmayr, T.; Pasma, T. *A Method for Linguistic Metaphor Identification*; John Benjamins: Amsterdam, The Netherlands, 2010; ISBN 9789027239044.
17. Krennmayr, T.; Steen, G. VU Amsterdam Metaphor Corpus. In *Handbook of Linguistic Annotation*; Ide, N., Pustejovsky, J., Eds.; Springer Science + Business Media: Dordrecht, The Netherlands, 2017; pp. 1053–1071.
18. Nacey, S.; Dorst, A.G.; Krennmayr, T.; Reijnierse, W.G. *Metaphor Identification in Multiple Languages*; John Benjamins: Amsterdam, The Netherlands, 2019; ISBN 9789027204721.
19. Bogetić, K.; Bročić, A.; Rasulić, K. Linguistic Metaphor Identification in Serbian. In *Metaphor Identification in Multiple Languages*; Nacey, S., Dorst, A.G., Krennmayr, T., Reijnierse, W.G., Eds.; John Benjamins: Amsterdam, The Netherlands, 2019; pp. 203–226.
20. Lakoff, G.; Espenson, J.; Schwartz, A. *Master Metaphor List*; Cognitive Linguistics Group, University of California at Berkeley: Berkeley, CA, USA, 1991.
21. Wallington, A.M.; Barnden, J.A.; Buchlovsky, P.; Fellows, L.; Glasbey, S.R. *Metaphor Annotation: A Systematic Study*; Technical Report CSRP-03-04; University of Birmingham: Birmingham, UK, 2003.

22. Birke, J.; Sarkar, A. A Clustering Approach for Nearly Unsupervised Recognition of Nonliteral Language. In Proceedings of the EACL 2006, 11st Conference of the European Chapter of the Association for Computational Linguistics, Trento, Italy, 3–7 April 2006; pp. 329–336.
23. Tsvetkov, Y.; Boytsov, L.; Gershman, A.; Nyberg, E.; Dyer, C. Metaphor Detection with Cross-Lingual Model Transfer. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, MD, USA, 22–27 June 2014; Association for Computational Linguistics: Stroudsburg, PA, USA, 2004; pp. 248–258.
24. Mohammad, S.; Shutova, E.; Turney, P. Metaphor as a Medium for Emotion: An Empirical Study. In Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics, Berlin, Germany, 11–12 August 2016; Association for Computational Linguistics: Stroudsburg, PA, USA, 2016; pp. 23–33.
25. Miller, G.A. WordNet: A lexical database for English. *Commun. ACM* **1995**, *38*, 39–41. [CrossRef]
26. Fellbaum, C. *WordNet: An Electronic Lexical Database*; MIT Press: Cambridge, MA, USA, 1998; ISBN 9780262061971.
27. Antloga, Š. Metaphor Corpus KOMET 1.0. *Clarin.si.* 2020. Available online: https://www.clarin.si/repository/xmlui/handle/11356/1293 (accessed on 8 January 2023).
28. Turney, P.; Neuman, Y.; Assaf, D.; Cohen, Y. Literal and Metaphorical Sense Identification through Concrete and Abstract Context. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, Edinburgh, Scotland, UK, 27–31 July 2011; Association for Computational Linguistics: Stroudsburg, PA, USA, 2011; pp. 680–690.
29. Mohler, M.; Bracewell, D.; Tomlinson, M.; Hinote, D. Semantic Signatures for Example-Based Linguistic Metaphor Detection. In Proceedings of the First Workshop on Metaphor in NLP, Atlanta, Georgia, 13 June 2013; Association for Computational Linguistics: Stroudsburg, PA, USA, 2013; pp. 27–35.
30. Shutova, E.; Sun, L.; Korhonen, A. Metaphor Identification Using Verb and Noun Clustering. In Proceedings of the 23rd International Conference on Computational Linguistics, Beijing, China, 23–27 August 2010; Association for Computational Linguistics: Stroudsburg, PA, USA, 2010; pp. 1002–1010.
31. Heintz, I.; Gabbard, R.; Srivastava, M.; Barner, D.; Black, D.; Friedman, M.; Weischedel, R. Automatic Extraction of Linguistic Metaphors with LDA Topic Modeling. In Proceedings of the First Workshop on Metaphor in NLP, Atlanta, Georgia, 13 June 2013; Association for Computational Linguistics: Stroudsburg, PA, USA, 2013; pp. 58–66.
32. Li, L.; Sporleder, C. Classifier Combination for Contextual Idiom Detection Without Labelled Data. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–7 August 2009; Association for Computational Linguistics: Stroudsburg, PA, USA, 2009; pp. 315–323.
33. Do Dinh, E.-L.; Gurevych, I. Token-Level Metaphor Detection Using Neural Networks. In Proceedings of the Fourth Workshop on Metaphor in NLP, San Diego, CA, USA, 17 June 2016; Association for Computational Linguistics: Stroudsburg, PA, USA, 2016; pp. 28–33.
34. Shutova, E.; Kiela, D.; Maillard, J. Black Holes and White Rabbits: Metaphor Identification with Visual Features. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, CA, USA, 12–17 June 2016; Association for Computational Linguistics: Stroudsburg, PA, USA, 2016; pp. 160–170.
35. Rei, M.; Bulat, L.; Kiela, D.; Shutova, E. Grasping the Finer Point: A Supervised Similarity Network for Metaphor Detection. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 7–11 September 2017; Association for Computational Linguistics: Stroudsburg, PA, USA, 2017; pp. 1537–1546.
36. Mykowiecka, A.; Wawer, A.; Marciniak, M. Detecting Figurative Word Occurrences Using Recurrent Neural Networks. In Proceedings of the Workshop on Figurative Language Processing, New Orleans, LA, USA, 6 June 2018; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; pp. 124–127.
37. Swarnkar, K.; Singh, A.K. Di-LSTM Contrast: A Deep Neural Network for Metaphor Detection. In Proceedings of the Workshop on Figurative Language Processing, New Orleans, LA, USA; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; pp. 115–120.
38. Wu, C.; Wu, F.; Chen, Y.; Wu, S.; Yuan, Z.; Huang, Y. Neural Metaphor Detecting with CNN-LSTM Model. In Proceedings of the Workshop on Figurative Language Processing, New Orleans, LA, USA, 6 June 2018; Association for Computational Linguistics: New Orleans, LA, USA, 2018; pp. 110–114.
39. Gao, G.; Choi, E.; Choi, Y.; Zettlemoyer, L. Neural Metaphor Detection in Context. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; pp. 607–613.
40. Mao, R.; Lin, C.; Guerin, F. End-to-End Sequential Metaphor Identification Inspired by Linguistic Theories. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July 2019; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 3888–3898.
41. Su, C.; Fukumoto, F.; Huang, X.; Li, J.; Wang, R.; Chen, Z. DeepMet: A Reading Comprehension Paradigm for Token-Level Metaphor Detection. In Proceedings of the Second Workshop on Figurative Language Processing, Online, 9 July 2020; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 30–39.
42. Gong, H.; Gupta, K.; Jain, A.; Bhat, S. IlliniMet: Illinois System for Metaphor Detection with Contextual and Linguistic Information. In Proceedings of the Second Workshop on Figurative Language Processing, Online, 9 July 2020; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 146–153.

43. Chen, X.; Leong, C.W.; Flor, M.; Klebanov, B.B. Go Figure! Multi-Task Transformer-Based Architecture for Metaphor Detection Using Idioms: ETS Team in 2020 Metaphor Shared Task. In Proceedings of the Second Workshop on Figurative Language Processing, Online, 9 July 2020; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 235–243.
44. Dankers, V.; Malhotra, K.; Kudva, G.; Medentsiy, V.; Shutova, E. Being Neighbourly: Neural Metaphor Identification in Discourse. In Proceedings of the Second Workshop on Figurative Language Processing, Online, 9 July 2020; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 227–234.
45. Lin, Z.; Ma, Q.; Yan, J.; Chen, J. CATE: A Contrastive Pre-Trained Model for Metaphor Detection with Semi-Supervised Learning. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Online and Punta Cana, Dominican Republic, 7–11 November 2021; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 3888–3898.
46. Choi, M.; Lee, S.; Choi, E.; Park, H.; Lee, J.; Lee, D.; Lee, J. MelBERT: Metaphor Detection via Contextualized Late Interaction Using Metaphorical Identification Theories. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Stroudsburg, PA, USA, 6–11 June 2021; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 1763–1773.
47. Li, Y.; Wang, S.; Lin, C.; Guerin, F.; Barrault, L. FrameBERT: Conceptual Metaphor Detection with Frame Embedding Learning. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Dubrovnik, Croatia, 9 February 2023; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 1558–1563.
48. Ge, M.; Mao, R.; Cambria, E. Explainable Metaphor Identification Inspired by Conceptual Metaphor Theory. *Proc. Conf. AAAI Artif. Intell.* **2022**, *36*, 10681–10689. [CrossRef]
49. Zhang, S.; Liu, Y. Metaphor Detection via Linguistics Enhanced Siamese Network. In Proceedings of the 29th International Conference on Computational Linguistics, Gyeongju, Republic of Korea, 12–17 October 2022; International Committee on Computational Linguistics: Stroudsburg, PA, USA, 2022; pp. 4149–4159.
50. Maudslay, R.H.; Teufel, S. Metaphorical Polysemy Detection: Conventional Metaphor Meets Word Sense Disambiguation. In Proceedings of the 29th International Conference on Computational Linguistics, Gyeongju, Republic of Korea, 12–17 October 2022; International Committee on Computational Linguistics: Stroudsburg, PA, USA, 2022; pp. 65–77.
51. Song, W.; Zhou, S.; Fu, R.; Liu, T.; Liu, L. Verb Metaphor Detection via Contextual Relation Learning. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Stroudsburg, PA, USA, 1–6 August 2021; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 4240–4251.
52. Katz, J.J.; Fodor, J.A. The Structure of a Semantic Theory. *Language* **1963**, *39*, 170. [CrossRef]
53. Wilks, Y. A preferential, pattern-seeking, Semantics for natural language inference. *Artif. Intell.* **1975**, *6*, 53–74. [CrossRef]
54. Wilks, Y. Making preferences more active. *Artif. Intell.* **1978**, *11*, 197–223. [CrossRef]
55. Fass, D. Met*: A Method for Dicriminating Metonymy and Metaphor by Computer. *Comput. Linguist.* **1991**, *17*, 49–90.
56. Wilks, Y.; Dalton, A.; Allen, J.; Galescu, L. Automatic Metaphor Detection Using Large-Scale Lexical Resources and Conventional Metaphor Extraction. In Proceedings of the First Workshop on Metaphor in NLP, Atlanta, GA, USA, 13 June 2013; Association for Computational Linguistics: Atlanta, GA, USA, 2013; pp. 36–44.
57. Mason, Z.J. CorMet: A Computational, Corpus-Based Conventional Metaphor Extraction System. *Comput. Linguist.* **2004**, *30*, 23–44. [CrossRef]
58. Resnik, P.S. *Selection and Information: A Class-Based Approach to Lexical Relationships*; Technical Report No. IRCS-93-42; University of Pennsylvania: Philadelphia, PA, USA, 1993.
59. Krishnakumaran, S.; Zhu, X. Hunting Elusive Metaphors Using Lexical Resources. In Proceedings of the Workshop on Computational Approaches to Figurative Language, Rochester, NY, USA, 26 April 2007; Association for Computational Linguistics: Stroudsburg, PA, USA, 2007; pp. 13–20.
60. Karov, Y.; Edelman, S. Similarity-Based Word Sense Disambiguation. *Comput. Linguist.* **1998**, *24*, 41–59.
61. Neuman, Y.; Assaf, D.; Cohen, Y.; Last, M.; Argamon, S.; Howard, N.; Frieder, O. Metaphor Identification in Large Texts Corpora. *PLoS ONE* **2013**, *8*, e62343. [CrossRef] [PubMed]
62. Tsvetkov, Y.; Mukomel, E.; Gershman, A. Cross-Lingual Metaphor Detection Using Common Semantic Features. In Proceedings of the First Workshop on Metaphor in NLP, Atlanta, GA, USA, 13 June 2013; Association for Computational Linguistics: Stroudsburg, PA, USA, 2013; pp. 45–51.
63. Sporleder, C.; Li, L. Unsupervised Recognition of Literal and Non-Literal Use of Idiomatic Expressions. In Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009), Athens, Greece, 30 March–3 April 2009; Association for Computational Linguistics: Stroudsburg, PA, USA, 2009; pp. 754–762.
64. Jurafsky, D.; Martin, J.H. Speech and Language Processing. 2023. Available online: https://web.stanford.edu/~jurafsky/slp3/ (accessed on 26 December 2022).
65. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. 2013. Available online: https://arxiv.org/pdf/1301.3781.pdf (accessed on 25 December 2022).
66. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; Dean, J. Distributed Representations of Words and Phrases and Their Compositionality. 2013. Available online: https://arxiv.org/pdf/1310.4546.pdf (accessed on 25 December 2022).

67. Pennington, J.; Socher, R.; Manning, C. GloVe: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; Association for Computational Linguistics: Stroudsburg, PA, USA, 2014; pp. 1532–1543.
68. Peters, M.E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. Deep Contextualized Word Representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, LA, USA, 2–4 June 2018; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; Volume 1, pp. 2227–2237.
69. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the NAACL-HLT 2019, Minneapolis, MA, USA, 2–7 June 2019; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 4171–4186.
70. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. 2019. Available online: https://arxiv.org/pdf/1907.11692.pdf (accessed on 25 December 2022).
71. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In Proceedings of the 31st Annual Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Neural Information Processing Systems: Long Beach, CA, USA, 2017; pp. 6000–6010.
72. Jozefowicz, R.; Vinyals, O.; Schuster, M.; Shazeer, N.; Wu, Y. Exploring the Limits of Language Modeling. 2016. Available online: https://arxiv.org/pdf/1602.02410.pdf (accessed on 22 December 2022).
73. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
74. LeCun, Y.; Boser, B.; Denker, J.S.; Henderson, D.; Howard, R.E.; Hubbard, W.; Jackel, L.D. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Comput.* **1989**, *1*, 541–551. [CrossRef]
75. Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; Stoyanov, V. Unsupervised Cross-Lingual Representation Learning at Scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 8440–8451.
76. Lample, G.; Conneau, A. Cross-Lingual Language Model Pretraining. 2019. Available online: https://arxiv.org/pdf/1901.07291.pdf (accessed on 5 January 2023).
77. Brysbaert, M.; Warriner, A.B.; Kuperman, V. Concreteness ratings for 40 thousand generally known English word lemmas. *Behav. Res. Methods* **2013**, *46*, 904–911. [CrossRef] [PubMed]
78. Bulat, L.; Clark, S.; Shutova, E. Modeling Metaphor with Attribute-Based Semantics. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Valencia, Spain, 3–7 April 2017; Association for Computational Linguistics: Stroudsburg, PA, USA, 2017; Volume 2, pp. 523–528.
79. Gutiérrez, E.D.; Shutova, E.; Marghetis, T.; Bergen, B. Literal and Metaphorical Senses in Compositional Distributional Semantic Models. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, 7–12 August 2016; Association for Computational Linguistics: Stroudsburg, PA, USA, 2016; pp. 183–193.
80. Stone, P.J.; Hunt, E.B. A Computer Approach to Content Analysis: Studies Using the General Inquirer System. In Proceedings of the AFIPS '63 (Spring): Proceedings of the May 21–23, 1963, Spring Joint Computer Conference, New York, NY, USA, 21–23 May 1963; Association for Computing Machinery: New York, NY, USA, 1963; pp. 241–256.
81. Leong, C.W.; Klebanov, B.B.; Shutova, E. A Report on the 2018 VUA Metaphor Detection Shared Task. In Proceedings of the Workshop on Figurative Language Processing, New Orleans, LA, USA, 6 June 2018; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; pp. 56–66.
82. Leong, C.W.; Klebanov, B.B.; Hamill, C.; Stemle, E.; Ubale, R.; Chen, X. A Report on the 2020 VUA and TOEFL Metaphor Detection Shared Task. In Proceedings of the Second Workshop on Figurative Language Processing, Online, 9 July 2020; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 18–29.
83. Bevilacqua, M.; Navigli, R. Breaking Through the 80% Glass Ceiling: Raising the State of the Art in Word Sense Disambiguation by Incorporating Knowledge Graph Information. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Stroudsburg, PA, USA, 5–10 July 2020; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 2854–2864.