VU Amsterdam Metaphor Corpus

Krennmayr, T.; Steen, G.J.

published in
Handbook of Linguistic Annotation
2017

DOI (link to publisher)
10.1007/978-94-024-0881-2_39

document version
Publisher's PDF, also known as Version of record

VU Amsterdam Metaphor Corpus

Tina Krennmayr and Gerard Steen

Abstract

The VU Amsterdam Metaphor Corpus consists of manual annotations of metaphors in four different registers—news texts, fiction, academic texts, and conversations. The goal of building this corpus was to investigate which metaphors are used in which forms, in which discourse contexts, in which registers, and for which purposes. This chapter reports on the development of the annotation scheme and its physical representation, describes the annotation process, and reports on inter-annotator agreement and quality control as well as current usage of the corpus. It also includes some quantitative results on the interaction between metaphor, register, and word class.

Keywords

Linguistic metaphor · Manual annotation · Register analysis

1 Background and Rationale

The VU Amsterdam Metaphor Corpus was built within the five-year research program "Metaphor in discourse: linguistic forms, conceptual structures, and cognitive representations" (Netherlands Organization for Scientific Research, NWO, VICI-

T. Krennmayr (✉)
Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
e-mail: t.krennmayr@vu.nl

G. Steen
University of Amsterdam, Amsterdam, The Netherlands

N. Ide and J. Pustejovsky (eds.), *Handbook of Linguistic Annotation*,
DOI 10.1007/978-94-024-0881-2_39

program, 277-30-001). The whole research program comprised two main phases: (1) developing a procedure for identifying metaphor in discourse and building the corpus and (2) describing the linguistic forms, conceptual structures and cognitive representations of metaphor in four different registers (academic texts, conversations, fiction, newspaper texts). The overall goal of the research was to determine which metaphors are used in which forms, in which discourse contexts, in which registers and for which purposes. An almost 200,000 word corpus was compiled from the BNC-Baby, a four-million word subcorpus of the British National Corpus. The corpus was annotated for metaphor by a team of researchers and was made available to the public for free.

Cognitive linguistics puts forward the idea that metaphor is ubiquitous in everyday language [20] because we actually *think* metaphorically. In other words, metaphor in language reflects conventional thought structures in our minds. Consider the following examples [20, pp. 7–8]:

Is that the *foundation* for your theory?
The theory needs more *support.*
We need *to construct* a *strong* argument for that.
We need *to buttress* the theory with *solid* arguments.
The theory will *stand or fall* on the *strength* of that argument.
So far we have *put together* only the *framework* of the theory.

These sentences describe the abstract topic of developing a theory through the more concrete concept of building something concrete (*foundation*, *support*, *construct* etc.). These expressions in italics are 'linguistic metaphors'; they express a cross-domain mapping from a usually more concrete source domain (e.g. building a building) to a more abstract target domain (e.g. developing a theory). The thought patterns underlying these linguistic expressions are called 'conceptual metaphors'. The metaphorical expressions in the examples above reflect the conceptual metaphor THEORIES ARE BUILDINGS.

However, the so-called conceptual metaphor theory was developed using artificial examples and was largely based on impressions rather than numbers. Although researchers later started to look at metaphor in language as it is actually used by people, many studies remain small-scale or restricted in their focus, or lack a rigorous, explicit method of identifying metaphor in the linguistic data. For example, many researchers are interested in the ways a particular metaphor may shape our thought and may consequently influence our actions (e.g. [18,22,26]). They therefore do not look at all but only a particular set of metaphors in a corpus. Other research concentrates on metaphor in a subset of a broader register (e.g. [9,18]), such as business news or sports reporting (e.g. [9]). Apart from a small number of exceptions (e.g. [9,27–29]), research on metaphor variation across different kinds of registers is scarce. Existing work relies on predefined search strings or it focuses on only those expressions that have been identified in small hand-annotated sample corpora (e.g. [29]) or is limited to selected semantic fields [27]. What was lacking was a more encompassing comparison between various registers that considers *all* metaphors in language.

That there is important variation in the distribution of metaphor was shown by for instance Cameron [7] who compared the metaphor density of three different conversation samples (classroom talk, doctor-patient interviews, reconciliation talk). Goatly's [15] investigation of metaphor variation covered a broader range of registers and reported a similar finding. However, a more precise, reliable and valid description of metaphor in diverging contexts of usage requires a quantitative comparison of metaphor use identified by a transparent and reliable technique. Building the VU Amsterdam Metaphor Corpus addressed this need.

Since a major goal of the project was to study the relation between metaphor and register, the VU Amsterdam Metaphor Corpus was compiled from four registers of the BNC-Baby – conversation, fiction, academic texts and news texts. This set was chosen to parallel the registers described in [6]. Biber pioneered the description of parameters of linguistic variation across a range of texts from different registers but did not include metaphor as one of the investigated features. Setting up a small parallel sample allowed for a description of metaphor in four registers of English that have been well studied from a grammatical point of view. One of the aims of our project was to investigate how metaphor contributes to the relation between register and linguistic features described by Biber et al. [6]. It is the first study to establish the proportion of metaphors in four registers, using a rigorous methodology for annotating metaphor. The systematic annotation of the corpus forms the basis for conducting these further analyses.

Quantitative analysis revealed that, overall, 13.6% of all words in the complete corpus are related to metaphor. However, metaphor is distributed unequally across the four registers (news: 16.4%; fiction: 11.9%; conversation: 7.7%; academic texts: 18.5%) and interacts with register properties in complex ways: there is a three-way interaction between metaphor, register and word-class. The relations between the three variables can largely be accounted for by the functional variation between word classes across registers, but metaphor also has some role of its own to play. Quantitative and qualitative results have been reported in detail in four publicly available Ph.D dissertations [12,16,17,19]).

2 Annotation Scheme

2.1 Underlying Assumptions

The annotation scheme attempts to capture those linguistic expressions that are seen as expressions of cross-domain mappings in cognitive linguistics [20]. From that perspective, metaphor introduces a different conceptual domain into the (sometimes just locally) dominant conceptual domain of the discourse, presumably causing a lack of coherence, which can then be resolved through a mapping from that different conceptual domain (the 'source domain') to the dominant domain of the discourse ('target domain') (see also [10], pp. 21, 35). A typical example is that of *underground* in "underground leadership" (A9J-fragment01). In this context, *underground* means

'secret and usually illegal' but the word also has a more concrete, basic meaning, namely 'below the surface of the ground'. According to current theory, the word is used indirectly because it evokes a referent ('secret') that is different from the more basic (spatial) meaning of *underground*. The metaphorical (indirect) meaning is held to arise through a mapping between the two conceptual domains related to the contextual and the basic meaning. In the corpus we annotated linguistic metaphors but did not proceed to identify their related conceptual metaphors. In the example above, this means that *underground* needs to be marked as a metaphorically used word but a potential underlying mapping associating a low position in space with illegal or secret activity is not annotated.

In order to build a corpus annotated for metaphor, the underlying assumption of cross-domain mappings was turned into a set of criteria for linguistic metaphor identification. The Pragglejaz Group [24] pointed out that researchers relying on their intuitions about what constitutes a metaphor, often disagree. As a response, they formulated a set of instructions with the goal of moving away from intuition and to achieve reliable metaphor identification across analysts. Their protocol, "MIP" (Metaphor Identification Procedure), constrains metaphor identification by checking meanings of each analyzed item, preferably in a dictionary. These are the steps of MIP:

1. Read the entire text/discourse to establish a general understanding of the meaning.
2. Determine the lexical units in the text/discourse

3a. For each lexical unit in the text, establish its meaning in context, i.e. how it applies to an entity, relation or attribute in the situation evoked by the text (contextual meaning). Take into account what comes before and after the lexical unit.

3b. For each lexical unit, determine if it has a more basic contemporary meaning in other contexts than the one in the given context. For our purposes, basic meanings tend to be:
   - more concrete; what they evoke is easier to imagine, see, hear, feel, smell, and taste.
   - related to bodily action.
   - more precise (as opposed to vague)
   - historically older.

   Basic meanings are not necessarily the most frequent meanings of the lexical unit.

3c. If the lexical unit has a more basic current/contemporary meaning in other contexts than the given context, decide whether the contextual meaning contrasts with the basic meaning but can be understood in comparison with it.

4. If yes, mark the lexical unit as metaphorical.

This step-by-step approach is theoretically compatible with the notion of metaphor as a cross-domain mapping. The basic meaning of a word evokes a source domain, whereas the contextual meaning can be ascribed to a target domain (see also Shutova,

this volume). The MIP approach also works "bottom-up" in that it does not make assumptions about related conceptual metaphors that guide linguistic metaphor identification. This was optimal for building the VU Amsterdam Metaphor Corpus, since we aimed at identifying all metaphorical language in the corpus and not just a specific set. A bottom-up approach does not start out from predefined sets of conceptual metaphors as deductive approaches do (e.g., [18]). Instead, the analyst looks at linguistic evidence without any preconceived ideas of what they may find. A clear advantage, therefore, is that the coder refrains from presuming conceptual metaphors, which reduces the danger of finding precisely those linguistic expressions that match a preconceived mapping. MIP identifies the metaphorically used words by the above criteria, but no mappings. In fact, the corpus was meant as a resource for doing subsequent research on the question which metaphorically used words might be related to which underlying conceptual cross-domain mappings.

The Pragglejaz Group [24] looked for 'indirectness' at the level of language. However, while annotating bulk data, we found that this operationalization is too restricted and does not cater to other forms of metaphor. Cross-domain mappings can surface in discourse in a number of different ways, not all of which can be captured by MIP. In the following example, the source domain is not expressed indirectly but directly: "Young Riders has a cast of five pouting male actors in an attempt to make a western with good demographics. The effect is rather *like an extended advertisement for Marlboro Lights*" (A2D-fragment05). For *advertisement* and *Marlboro Lights*, there is no comparison of a contextual and a basic sense. They are used in their basic sense. Yet, it is also true that there is a comparison between 'the effect' and a Marlboro Lights ad. This is a direct metaphorical comparison, which requires a different set of instructions for identification and annotation. Direct metaphor is often, but not always, introduced by a lexical marker ([15], p. 183ff), such as *like* or *as*. We also coded such markers as metaphor signals.

A word can also be connected to a source domain implicitly: "For three reasons such a move should be welcomed. First, *it* would bring Britain into line with the best European practice (…)" (A1F-fragment09). In discourse analysis, the discourse would have to show the previous concept (move) and not the cohesive element (it). *It* is an implicit metaphor because, in the surface text, the language does not indicate the need for a nonliteral comparison. *It* substitutes the metaphorically used *move* (underlined) in the previous sentence. *It* is not itself used indirectly (i.e. there is no more basic sense that could be contrasted to the contextual one). Implicit metaphor is thus due to a cohesive link in the discourse, pointing to recoverable metaphorical meanings.

In order to capture the phenomena of direct and implicit metaphor, the notion of metaphor was therefore pitched at the level of conceptual structure. These (and other) expansions of the MIP procedure resulted in a more detailed and elaborate protocol – MIPVU (Metaphor Identification Procedure Vrije Universiteit). The complete procedure has been published in [30].

As a maximally inclusive procedure, MIPVU may create the impression that almost anything counts as a metaphor. However, our research shows that this is not true. Only 13.6% of all lexical units in the corpus are metaphorically used.

This includes metaphorically used prepositions (e.g. *on* Monday, where the time-related contextual meaning contrasts and can be understood in comparison with the more basic physical meaning of *on* "touching and supported by the top surface of something") or demonstratives (e.g. *this* idea, where the contextual sense of *this*, "referring to the particular thing that you are going to talk about" can be contrasted and understood in comparison with the more basic meaning "the one that is here"). If these frequent metaphorical word classes are ignored, the percentage of metaphor-related language use would be dramatically lower, showing that the bulk of language use is not metaphorical.

2.2 Choice of Annotations and Developing the Annotation Scheme

At the beginning of the project we relied on MIP. The annotation scheme was pre-set and thus limited to the kind of metaphorical language use that could be detected by MIP, namely indirect metaphor. In other words, the initial annotation scheme simply required a decision between "metaphorical" or "non-metaphorical". When the limitations of the Pragglejaz approach for metaphor identification became clear after annotating a handful of texts, however, it led to a new operationalization of metaphor and thus to an expansion of the coding scheme. As the annotators progressed, new annotations and detailed instructions were added, producing an eighteen-page protocol, in effect an extended and refined version of MIP which we dubbed 'MIPVU'.

In addition, for determining the metaphorical status of a word, MIPVU consistently uses independent reference tools to check basic and contextual meanings of a lexical item. The main tools are two corpus-based dictionaries, namely the *Macmillan English Dictionary for Advanced Learners* [25] and the *Longman Dictionary of Contemporary English Online*. While most historically older meanings are also the most basic meanings, a word's history was generally not taken into account in order to determine its basic meaning. This is because the project dealt with contemporary texts read by contemporary language users and language users are generally not aware of historical meanings of words in contemporary language use. This means that words like *ardent* in 'ardent lover' are not considered metaphorical in the MIPVU approach, since the historically older temperature sense, which fulfills the criteria of a more basic sense, has disappeared from contemporary language use. The Macmillan English Dictionary lists emotion related senses only. Only in rare cases, when a decision on a word's metaphorical status could not be made by using the contemporary dictionaries alone, was the historical dictionary *Oxford English Dictionary Online* consulted.

The bulk of metaphors in the VU Amsterdam Metaphor Corpus is conventional, the metaphorical sense being listed in the dictionary. An example of a highly conventional metaphor is *valuable* in "to do valuable work." Most people would not recognize *valuable* as metaphorically used. However, the word, meaning "very useful and important" in this context, has another meaning that fulfills the criteria of a more basic meaning, namely "worth a lot of money." Therefore it needs to be marked as metaphorically used. When the contextual meaning of a lexical unit is not in the

relation to metaphor	metaphor type	XML representation	corpus examples
metaphor	indirect	<mrw type="met">valuable</mrw>	Professional religious education teachers like Marjorie B Clark (Points of View, today) are doing *valuable* work in many secondary schools (…). (K58-fragment01)
	direct	<mrw type="lit">ferret</mrw>	(…) he's like a *ferret*. (KBD-fragment21)
	implicit	<mrw type="impl">it</mrw>	Naturally, to embark on such a *step* is not necessarily to succeed in realizing *it*. (A9J-fragment01)
WIDLII		<mrw type="met" status="WIDLII">up</mrw>	driven *up* the bumpy Forest Drive to East Kielder Farm, (…). (AHC-fragment60)
PP		<mrw type="met" status="PP">decide</mrw>	A party can't even *decide* its name (…). (A7W-fragment22)
UNCERTAIN		<mrw type="met" status="UNCERTAIN">appealed</mrw>	The council appealed by cases stated. (A7Y-fragment03)
signal		<mFlag type="lex">as if</mFlag>	It is *as if* it is walking through a minefield. (A9J-fragment01)
		mFlag type="morph">like</mFlag>	The wave-*like* pattern of the Intifada. (A9J-fragment01)
		<mFlag type="phrase" id="a9j-fragment01-mfp1">in</mFlag></w>(…)<mrw type="met">role</mrw>(…)<mFlag type="phrase" corresp=a9j-fragment01-mfp1">of</mFlag>	(…) acts *in* the *role of* field general (A9J-fragment01)

Fig. 1 Overview of all annotations used

dictionary, which happens only very seldom (at most one per cent of all cases), this points to a novel metaphor. In the project we did not make a distinction between conventional and novel metaphorical uses of words and marked both instances simply as "metaphor-related words". Of course, such distinctions can be drawn, if desired, simply by assigning separate codes for the two phenomena. All metaphors, whether conventional or novel, are *potential* metaphors, meaning that they may or may not activate a cross-domain mapping in people's minds.

Figure 1 gives an overview of all annotations used, their representation in XML format (for more on the physical representation of the annotations see Sect. 3), and concrete examples from the corpus. Except signals for metaphor, all annotated lexical units received the general code "mrw" – metaphor related word, indicating that they are candidates for expressing some cross-domain mapping, regardless of how they surface linguistically. Indirect metaphors were given the code "met", direct metaphors "lit", and implicit metaphors "impl" (see Fig. 1). In retrospect, these codes are not very transparent. Thus, coding only indirect metaphors as "met" may suggest that direct metaphors and implicit metaphors are not really metaphors after all, since they lack the "met" code. Similarly, the code "lit" (for 'literal') may suggest that a word with this code is used literally and not metaphorically, when all it means is that the word is used in its basic sense but is still part of a cross-domain mapping. These annotations reflect the fact that they were developed not before but in the process of coding the corpus.

In other words, initially we did not employ annotations for direct and implicit metaphors, but once we developed these based on our new operationalization of metaphor on the conceptual level, the codes were simply added as two further

phenomena besides indirect metaphors, but we continued using 'met' for indirect metaphor. This is not a problem and can be changed in later editions of the corpus. Once we included direct comparison, we also coded words signaling such comparisons (e.g. *like*, *as* etc.) with metaphor flags ('mFlag'). Type="lex", indicates the signal is one word (e.g. *like*), type="morph" indicates that the signal is part of a word (e.g. wave-*like*), and type="phrase" indicates that the signal spans over more than one word (e.g. *in the role of*). In order to create one lexical unit for the multiword flag, an id= corresp= code was added.

In order to be maximally inclusive, ambiguous cases received the additional 'status' code "WIDLII" (When In Doubt, Leave It In). Consider the following corpus example: 'By the time I had turned off the road (…) and driven *up* the bumpy Forest Drive to East Kielder Farm (…)' (AHC-fragment60). The context does not specify whether the farm is at a higher location or further down a road. Both a metaphorical and a non-metaphorical interpretation are therefore possible. In such cases, WIDLII was added in the status field. The code was also used for unclear cases for which analysts could not reach agreement during group discussion (for more on group discussion see Sect. 4). This makes it possible to quantify difficult-to-categorize cases. The refined annotation system thus makes a distinction between clear metaphors, non-metaphors and borderline (WIDLII) cases.

Another issue that came up repeatedly during the annotation process was the interaction of metaphor and metonymy when personification was involved. This is illustrated in the following example: "A party cannot even *decide* its name (…)" (A7W-fragment22). *Decide* can be interpreted as metaphorically used since *deciding* is a human activity ('to make a choice about what you are going to do') whereas in this context it is connected to an abstract entity (party). However, if the individuals making up the party are in focus, then *party* is interpreted metonymically and *decide* is not used metaphorically. Cases like these received the additional 'status' code "PP" (possible personification). Since our project focused on the annotation of metaphor and not metonymy, the noun "party" was not coded as metonymy. Overall, this phenomenon is not particularly frequent. In the complete corpus, 84,4% of the lexical units were not metaphor-related, 13% were metaphor-related and only 0.6% were metaphor-related due to possible personification.

The 'status' code "UNCERTAIN" was simply used by analysts to indicate that they were not sure how to code a certain word. This annotation alerted fellow researchers cross-checking the annotations (for details about cross-checking see Sect. 4.3) that they needed to pay particular attention to this lexical item. The "UNCERTAIN" annotations were removed when preparing the final version of the annotated text.

A small number of words from the conversation register had to be excluded from metaphor analysis. This is because it was impossible to determine the contextual meaning – often because of aborted utterances and lack of context. Analysts would indicate this by adding the comment <!–DFMA–> (Discard From Metaphor Analysis) next to the lexical unit in question.

These annotations were not used in the original MIP procedure. The main differences between MIP and MIPVU are listed in Table 1.

Table 1 Main differences between MIP and MIPVU

	MIP	MIPVU
Definition of basic meaning	More concrete, related to bodily action, more precise (as opposed to vague), historically older	More concrete, related to bodily action, more precise (as opposed to vague)
Lexical units	Crosses word class	Does not cross word class
Dictionaries	Macmillan English Dictionary for Advanced Learners	Macmillan English Dictionary for Advanced Learners; Longman Dictionary of Contemporary English Online; Oxford English Dictionary
Types of metaphors coded	Metaphor and non-metaphor	Metaphor-related words (indirect metaphor, direct metaphor, implicit metaphor), metaphor signals, ambiguous metaphor, possible personification

3 Physical Representation

The VU Amsterdam Metaphor corpus consists of annotated files taken from the BNC-Baby. The BNC-Baby is marked up in XML format. This format has the advantage of not requiring any particular software and has emerged as a standard way of publishing annotated data. Our goal was to make the annotated metaphor corpus available as part of the BNC-Baby, to be published at the Oxford Text Archive (OTA). The choice of annotating in XML format was therefore evident. We used the XML editor <oXygen/> to annotate the data. The choice for using this software was a practical one – we had a programmer who knew how to tweak it for our purposes. The annotations are represented as tags that are delimited by < and >. They contain the name of the tag, which is preceded by /. For example, all metaphors were coded with the tag </mrw>, which stands for "metaphor-related word".

Annotating in XML format had the clear advantage of allowing quick processing of the data for their publication at OTA. A disadvantage is, perhaps, that researchers who are inexperienced in programming may find that they lack the skills needed for transforming data into other formats, such as into SPSS, for further data processing. However, with the help of a programmer the advantages outweigh the disadvantages. Markup conventions can be learned relatively quickly - even by researchers unfamiliar with XML. If expert help is available, for instance for creating new annotation tags, coding in <oXygen/> is also feasible without knowledge of XML.

4 Annotation Process

The complete corpus of 186,695 words was annotated manually using the MIPVU metaphor identification protocol. Text fragments were randomly selected from the four registers fiction, newspaper texts, academic texts, and conversations in the BNC-Baby. Four annotators went through all texts from the corpus on a word-by-word basis. For each word, they determined whether or not it should be coded as related to metaphor, based on the MIPVU procedure. This involved checking the contextual and basic meanings of each lexical unit in a dictionary and deciding, for each case, if those meanings contrast and can be understood in comparison to each other. If this was the case, the lexical item was marked as metaphorically used. For direct and implicit metaphor, the instructions were slightly different. This is a laborious and time-consuming process and puts a practical limit on the amount of data that can be annotated given the resources at hand. The upshot of manual annotation is, however, that the quality is superior to automatic analysis (see e.g., [2,21]). It has produced a protocol for metaphor identification that is transparent, systematic, and, after some initial training, relatively easy to use—our lab has run several post grad courses for Ph.D students and post doc researchers to substantiate this.

4.1 Annotators

The corpus was built over a two-year period. In the first year of the project, four Ph.D students annotated the corpus for metaphor use. They initially applied MIP, as laid out by the Pragglejaz Group [24]. Continuous calibrations to the procedure resulted in the refined and detailed MIPVU protocol. The students' backgrounds were in English language and linguistics. Two of them were native speakers of Dutch, one was a native speaker of Polish and one was a native speaker of Spanish. In the second year of the project the Polish and Spanish student discontinued their work and two new Ph.D students joined the project. Their background was in English and German language and linguistics. They were both native speakers of German. Only one of them had significant previous experience with metaphor research in the form of a master's thesis.

Rather than put the project in peril, the change of half the team brought an unexpected advantage. By the time the new students started working on the project, the metaphor identification procedure was almost in its final form. Neither of the new students had prior experience with either MIP or MIPVU. This served as an excellent test case to see how quickly novice annotators could use the procedure and perform at par with experienced coders. Before the new students were given texts to code individually, they first worked together for about a week on a blank copy of a text that had already been annotated. They then compared their results with the annotations of the previous team. They also met informally with the experienced team members to receive help and ask questions. In the second week, they started to work on texts independently. After three months, a reliability test (for more on reliability testing see Sect. 4.4) was carried out to measure the performance of the new team.

The new team performed as well as the old team, showing that the procedure can be transferred to new teams members without difficulty.

4.2 Annotation Environment

In the first few months of the project, when the oXygen editor was not available in its appropriate form yet, the annotators simply coded text in a Word document by inserting tags (for example, 'M' for metaphor) behind each lexical unit that was identified as metaphorically used. For example: 'How long**M** has she been there?' (ABN9-fragment01). However, the goal was to make the corpus available as part of the BNC-Baby (which is available in XML format). Therefore, after several months, the annotators switched to using the XML editor <oXygen/>. Annotations were added in angular brackets and the sentence above then looked like this: How <mrw type="met" >long</mrw> has she been there?

4.3 Annotation Process

The analysts coded the complete corpus using the MIPVU identification protocol. The annotation process was as follows:

(1) The principal investigator selected fragments from the BNC-Baby for analysis and assigned them to the individual Ph.D students. The selection process was guided by equal distribution across BNC-Baby files, with excerpts coming from beginnings, middles and ends of files, and having a somewhat variable bandwidth of words. Each student received fragments from each of the four registers in the corpus (fiction, academic texts, conversations, newspaper texts). This ensured that each of them was exposed to differences between phenomena typical of a particular register that had to be solved consistently using the same identification procedure. The following details of each sampled text were recorded in an administrative (Microsoft Access) database: the file name and fragment number, the number of words annotated, the percentage of the complete BNC-Baby file annotated, the name of annotator of the text, and the date of the annotation.
(2) Every week, the students checked which text fragments were assigned to them. Each student coded the texts individually for metaphor using the MIPVU protocol.
(3) When a text was fully annotated, they sent it to the principal investigator who uploaded the texts on a discussion website on the university intranet. This website had been specifically created for the purposes of crosschecking all texts.
(4) Each text on that website was then checked by the other three analysts. If they disagreed or had doubts about certain annotations, they posted a comment on the site. Here is an example of the discussion site on the web – a paragraph from A7Y-fragment03:

196 For Mrs Bujok it was argued that the 1936 Act was designed to <mrw type="met" morph="n" TEIform="seg">secure</mrw><mrw type="met" morph="n" TEIform="seg">in</mrw>the interests

196.2 designed: M since basic = to make a drawing or plan of something that will be made or built. This is about an act. A
196.2.1 yes, M GVAP

Two lexical units in this excerpt (sentence 196) had been marked as metaphor-related, namely *secure* and *in.* They are surrounded by mrw ("metaphor-related-word") tags. Under each sentence, analysts cross-checking the document, could add comments or queries, which they signed with their initial. Comment 196.2 was added by annotator 'A' drawing attention to the lexical unit *designed.* The annotator believed that *designed* also needed to be marked as a metaphor-related word ('M') because the word has a more basic meaning, namely 'to make a drawing of plan of something that will be made or built.'

As for comment 196.2.1: once a text had been completely checked by all team members, a group meeting was held in which cases of disagreement were discussed among the four analysts and the group leader. Decisions were recorded on the website and – if necessary – corrections were made in the annotated file by the analyst who had been in charge of the initial annotation. In the corpus example above, the group decided to follow annotator A's reasoning. The analyst-in-charge (i.e. the one who was the annotator of that text) recorded the decision to mark *designed* a metaphor-related ("yes, M") on the discussion site and signed the decision off with GVAP ("Group Validation after Pragglejaz" – "Pragglejaz" was used as shorthand for "group discussion"). The analyst recorded the date of corrections in the administrative database after which they stored the final version of the file in a group folder on the university server.

(5) Cases that were not simply errors spotted by the other researchers but which needed prolonged group discussion were entered into an Access database for future reference for increased coding consistency. The final database turns out to have 1180 entries. The following specifics were recorded: word class, word-class subcategory, basic meaning, dictionary source (i.e. the dictionary from which the basic meaning was taken), contextual meaning including the use of the lexical unit in context taken from the corpus or the dictionary, metaphorical status (metaphorical, non-metaphorical, borderline metaphorical, possible personification), and a comment on annotation decisions (if applicable). Here is a concrete example from this lexical database illustrating the entry for *attract* as in "They were beginning to attract a penumbra of gallery-goers" (FET-fragment01):

Word class	Verb
Word-class subcategory	Transitive
Dictionary source	Macmillan
Basic meaning	To make something move near someone or something (MM3)
Contextual meaning	To make someone interested in something so that they do it or come to see or hear it (MM1): "They were beginning to attract a penumbra of gallery-goers" (FET-fragment01)
Metaphorical status	Not-M
Comment on annotation decisions	As long as this sense involves physical movement towards a concrete location, it is considered a non-metaphorical extension of the basic sense

4.4 Inter-annotator Agreement

Any reliable metaphor identification protocol needs to guide analysts in a way that leads them to making highly similar judgments. Therefore, the inter-coder agreement for coded metaphor was closely monitored through six reliability tests conducted over a period of less than two years. These tests measured the performance of four analysts when they had analyzed their texts independently of each other. Reliability was measured before group discussion.

The first reliability test was conducted ten weeks after the start of the research project (with the 'old' team of Ph.D researchers). It revealed shortcomings in the developing MIPVU procedure, which was not yet detailed enough to be applied to bulk data. Testing was then resumed after three months. The texts were randomly selected from the BNC-Baby files and ranged from 713 to 1,940 words, for a total of 6,659 words in five tests. The first three tests were conducted with the 'old' team. The final two tests were conducted with the 'new' team. We measured analyst agreement on a case-by-case basis (Fleiss' Kappa) and the overall degree of difference between individual researchers (Cochran's Q), both in SPSS15. Since the incidence of fine-grained codings of borderline cases, direct metaphor, indirect metaphor and personification turned out to be extremely low in the corpus (e.g. only 1% of all cases were borderline), the reliability tests looked at whether analysts coded a unit as metaphor-related or not but did not look at more-fine-grained codings.

The results were good, at significance level $\alpha = 0.05$. For the Fleiss' Kappa test statistic, which is appropriate for assessing agreement between more than two analysts, the mean value was 0.85. On average, the four analysts reached unanimous agreement on whether or not a word was related to metaphor for 92.5% of all cases, in five distinct tests spread over time (N = 713, 1180, 1940, 905, and 1921). These results held between two differently composed teams, with two analysts remaining constant. They also held across all four registers.

Cochran's Q looks at analyst bias and checks whether one or more analysts are behaving significantly differently than the others. It was significant in the second and third reliability test. It was also significant for two of the four texts in the fourth test and for three of the texts in the sixth test. In the fifth test Cochran's Q did

not reach significance. These findings suggest that one or two analysts often scored either fewer or more items than the others, per test, implying that the analysis is not entirely reliable from that perspective and displays analyst bias. However, the regular annotation protocol, as laid out in Sect. 4.3, always contained group discussion, a step designed to filter out annotation bias and concomitant errors. Group dynamics in this process must be acknowledged. However, what is crucial is that the basis of the metaphor identification procedure lies in the reliable individual case-by-case analyses as was shown by Fleiss' Kappa. The group discussions therefore function as a further step in increasing consistency and systematicity.

A major factor in analyst disagreements was the ambiguity of some of the word meanings. An example is the preposition *at* in "Jack Kahn graduated with honours *at* the University of Leeds in 1928 (...)." The question is whether the preposition refers to an actual place (which renders its use non-metaphorical since the contextual meaning then equals the spatial basic meaning) or whether the meaning is more broadly constructed, in this case referring to what someone was doing (in which case it is metaphorically used.)

5 Quality Control

Whether researchers develop an identification protocol from scratch or refine and expand an existing protocol, they will commonly have to adjust and change decisions made earlier in the annotation process based on new insights and results of group discussions. As a consequence, annotations that had been inserted into texts before new decisions were made may not be in line with these new decisions. This naturally introduces error into the annotation process. In order to guarantee the quality of the annotations, a troubleshooting round was carried out once the complete corpus had been annotated. Features that turned out to be particularly problematic during the annotation process were selected for closer inspection. The goal was to remove systematic errors and to estimate and report error margins.

The following features were checked for errors by manually examining a sample for each. Phrasal verbs, compounds, and polywords were checked for correctness in determining the unit of analysis (they all needed to be coded as one unit). As far as metaphor annotation was concerned, the following cases were selected: borderline cases (WIDLIIs), units that were discarded from metaphor analysis because lack of context made them unintelligible, and units signaling metaphors. All sampled items for which an error was detected were corrected and the error margin was calculated. During this 'clean-up' process, annotators noticed that one code, namely that of 'implicit metaphor', had barely been used. A check of a random text sample revealed that there had been no systematic coding of this phenomenon, which was mainly a result of the procedure not being fully explicit. In a time-consuming but worthwhile effort, the whole team developed a set of clear instructions and subsequently each Ph.D student went through roughly a fourth of the corpus and fixed the errors.

Manually annotating linguistic metaphors is not infallible. Both systematic and erratic errors remain. What we can do, however, is build in as many checks as possible in order to reduce error. This ranges from cross-checks by analysts, to group discussions, to systematic checks of cases that have been identified and collected as problematic during the annotation process. This way, the error rate can be reduced and, at the same time, it is possible to estimate the quality of any interpretations that arise from an analysis of the data.

6 Main Results

Counter to intuition, fiction is not the most metaphorical register. It only comes in third (11.9%) after academic texts (18.5%), news texts (16.4%) and conversation (7.7%). However, the picture is not as simple. This is because word class correlates with linguistic characteristics of registers [4,5]. For example, highly informational texts such as news articles feature a prominent use of nouns, prepositions, and adjectives. We investigated what happens to the relationship of word class and metaphor if metaphor is added into the picture and found a three-way interaction between the variables metaphor, register and word class ($\chi^2(21) = 890.95$, $p < 0.000$). This confirms that word classes are distributed differently across different registers and that metaphors are distributed unequally across word classes and registers. This means that an analysis of metaphor use in a text must take the distribution of word classes in the register into account, as this distribution impacts metaphor frequency. The main patterns of this three-way interaction can be summarized as follows ([16] p. 138, pp. 141–2):

Considering each of the four registers separately, we see that prepositions and verbs generally tend to be used more metaphorically than average, even within the varied frequencies of metaphor between registers. Thus, in academic prose, prepositions are metaphorical in 42.5% of all cases, in news, 38.1% of all cases, in fiction, 33.4%, and in conversation, 33.8%. For verbs, these percentages are 27.7% for academic prose, 27.6% for news, 15.9% for fiction and 9.1% for conversation. Even though these percentages vary markedly, they are all higher than average within each register, suggesting that prepositions and verbs are generally more metaphorically used than other word classes, which may be a reflection of their frequently abstract meanings. By contrast, the word classes labelled as conjunctions and 'rest' in the data always displayed exceedingly low scores for metaphorical usage, averaging 1.2% and 1.1% across all four registers with some (non-significant) variation; this may be interpreted as a reflection of their frequently empty or sketchy grammatical meaning which makes it hard to build a contrast between a basic and a contextual sense that can potentially express a cross-domain mapping.

More conspicuous three-way interactions can be observed in the other word classes. For instance, adjectives are close to the average of 18.5% in academic texts (with 17.6%), but exceed the register averages for news (21% versus register average of 16.4%), fiction (19.4% versus register average of 11.9%), and conversation

(13.3% versus register average of 7.7%). This comparison suggests that in general, adjectives might be more metaphorical than the register average, but that this does not hold for academic texts, where they are a little less often metaphorical than the register average. It is possible that this may be due to the relatively higher number of non-metaphorical, technical adjectives, such as *social-scientific, historical* and so on in the academic register. This picture is further complicated when comparisons are made by fixing word classes as distinct data sets and comparing the distributions of metaphor and register within each word class—for further information we refer the reader to the four publicly available Ph.D theses mentioned above [12, 16, 17, 19]).

The story is not complete unless we also briefly mention the role of metaphor form. When the distribution of metaphor across word classes and registers was split up for indirect metaphor, direct metaphor, and implicit metaphor, an interesting further interaction was obtained. It should be recalled that implicit and direct metaphor comprise only 0.2% each of the total amount of data, as opposed to indirect metaphor which accounts for 13.3% of the data. However, within this uneven distribution, direct metaphor turned out to exhibit a substantially different rank order between registers, showing that this time it was fiction that was most metaphorical, closely followed by news, whereas academic texts were almost comparable to conversation in that neither exhibited much direct metaphor. In other words, fiction and then news texts use substantially more simile and other direct and explicit comparisons than academic texts and news. This seems to be a reflection of what intuition tells us about the metaphorical nature of fiction and the rhetorical nature of a lot of news writing. In all, then, data analysis of the corpus revealed a four-way interaction between register, word class, metaphor, and metaphor type.

7 Usage

7.1 Data Availability

From the start of the VU Amsterdam Metaphor Corpus project, the goal was to make the corpus available to the public. It is currently available in two different forms: (1) as XML files (TEI P5 XML) at the Oxford Text Archive (OTA) and (2) through a simple search form hosted at Metaphor Lab Amsterdam (http://metaphorlab.org/metaphor-corpus).

(1) The XML files can be downloaded for free from OTA (http://ota.ahds.ac.uk/desc/2541). They are available for non-commercial use under the terms of the BNC License and by agreeing to the terms and conditions of use stated on the website.
(2) The corpus can also be searched using simple search forms hosted at Metaphor Lab Amsterdam (http://metaphorlab.org/metaphor-corpus). The online corpus was subjected to another round of 'clean-up' through which error was further reduced. The annotations in the online corpus therefore do not fully match up with the corpus

as published at the OTA. We are considering whether to prepare a new edition of the corpus for OTA.

Three output forms can currently be generated when searching the online corpus:

(a) KWOT (keyword-out-of-context)-listing, which is a tabular overview specifying e.g., register, document, sentence, word number, word class, relation to metaphor, and metaphor type for each hit

(b) a concordance or KWIC (keyword-in-context)-listing

(c) raw counts in table format

The online corpus is licensed under a *Creative Commons Attribution-ShareAlike 3.0 Unported License*.

7.2 Usages of the Data

The annotated corpus was made freely available for a number of reasons. First of all, there were no annotated corpora available that had been systematically coded for metaphor using a transparent, replicable procedure. As Sect. 4 illustrated, building the corpus was a major group effort and not every researcher has the time and means to embark on such an endeavor. By making the corpus available, we aim to provide the metaphor community with the opportunity to access fully annotated material and to approach the use of metaphor in academic texts, fiction, conversation and news texts with their own research questions. For example, Berber-Sardinha [3] has recently performed a multidimensional analysis [4] to examine the relation between register and the use of metaphor.

Second, we expected the corpus to be useful for anyone in the process of annotating their own data for metaphor. Researchers can check lexical items in the corpus to see how they were annotated, which may be especially helpful to those who do not have the luxury of working in a team and meeting with colleagues for group discussion. This corpus can serve as a rich learning tool, particularly for researchers attempting to apply the MIPVU procedure to their data; the online search tool is particularly suitable for this purpose. Indeed, we have made successful use of the corpus in training Ph.D students and postdocs in using MIPVU at the Metaphor Lab Summer and Winter Schools. We have also received emails from researchers who have pointed out the usefulness of this source for checking cases they have difficulties with when annotating their own corpora. As an added benefit, this furthers discussion of remaining weaknesses of the MIPVU procedure and sharpens our eye for thinking critically about subtle details in identifying metaphor in discourse.

The data have recently been used to build a Russian Metaphor corpus [1]. This corpus contains the same registers as the VU Amsterdam project and was annotated using MIPVU (adapted to the Russian language). We had not foreseen that the complete project would be used as a model to generate a corpus in another language. Evidently, projects like these can provide the first step towards a multilingual Metaphor Corpus.

While we have not used the corpus as training data for machine learning algorithms ourselves, there have been attempts by colleagues in computational linguistics to use the Metaphor Corpus for devising tools for automatic metaphor identification (e.g., [2,13,14,23]).

8 Conclusion

Building the VU Amsterdam Metaphor Corpus has yielded a valuable resource that is available for other researchers for free. Most importantly, it has served as a test-bed showing that metaphor can be systematically annotated in real language data. It has also shown that annotating as a team yields reliable results. Analysts systematically collected metaphorically used expressions by applying the MIPVU protocol and monitored their performance through reliability tests. They further developed and refined the original MIP procedure, making it possible to account for different kinds of linguistic manifestations of cross-domain mappings, such as indirect, direct and implicit metaphor. While manual annotation put a limit on the size of the corpus, manual coding allowed for building in control mechanisms (e.g. cross-checks, group discussion, troubleshooting systematic errors, reliability testing) in order to control quality. The resulting database is a unique effort to add validity and comparability to metaphor research.

The corpus annotation has also demonstrated that it is possible to collect metaphor data at the linguistic level alone, without making assumptions about related conceptual structures, which has also been advocated by Cameron and Low, Charteris-Black, and Deignan ([8,10,11]). The dataset serves as a basis for further analysis of the conceptual structure underlying the metaphorically used words identified in the dataset. The corpus can also serve as a source for creating experimental material to research metaphor processing. Overall, this research contributes to a better view of the role of linguistic forms of metaphor in discourse.

References

1. Badryzlova, Y., Isaeva, Y., Shekhtman, N., Kerimov, R.: Annotating a Russian corpus of conceptual metaphor: a bottom-up approach. In: Proceedings of the Workshop on Metaphor in NLP, pp. 77–86 (2013)
2. Berber Sardinha, T.: A tool for finding metaphors in corpora using lexical patterns. Paper presented at Corpus Linguistics 2009, Liverpool (2009)
3. Berber Sardinha, T.: Register variation and metaphor use: a multi-dimensional perspective. In: Herrmann, J.B., Berber Sardinha, T. (eds.) Metaphor in specialist discourse (2015)
4. Biber, D.: Variation across speech and writing. Cambridge University Press, Cambridge (1988)
5. Biber, D.: Dimensions of register variation. A cross-linguistic comparison. Cambridge University Press, Cambridge (1995)
6. Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E.: The longman grammar of spoken and written english. Longman, London (1999)

7. Cameron, L.: Metaphor and talk. In: Gibbs, R.W. (ed.) The cambridge handbook of metaphor and thought, pp. 197–211. Cambridge University Press, Cambridge (2008)
8. Cameron, L., Low, G. (eds.) Researching and applying metaphor. Cambridge University Press, Cambridge (1999)
9. Charteris-Black, J.: Metaphor and vocabulary teaching in ESP economics. engl. specif. purp. **19**, 149–165 (2000)
10. Charteris-Black, J.: Corpus approaches to critical metaphor analysis. Palgrave Macmillan, Houndmills (2004)
11. Deignan, A.: Metaphor and corpus linguistics. John Benjamins, Amsterdam (2005)
12. Dorst, A.G.: Metaphor in fiction: linguistic forms, conceptual structures, cognitive representations. BOXpress, Oisterwijk (2011)
13. Dunn, J.: What metaphor identification systems can tell us about metaphor-in-language. In: Proceedings of the Workshop on Metaphor in NLP, pp. 1–10 (2013)
14. Florou, E.: Detecting metaphor by contextual analogy. In: Proceedings of the ACL Student Research Workshop, pp. 23–30 (2013)
15. Goatly, A.: The language of metaphors. Routledge, London (1997)
16. Herrmann, J.B.: Metaphor in academic discourse: linguistic forms, conceptual structures, communicative functions and cognitive representations, vol. 333. LOT, Utrecht (2013)
17. Kaal, A.A.: Metaphor in conversation. Boxpress, Oisterwijk (2012)
18. Koller, V.: Metaphor and gender in business media discourse: A critical cognitive study. Palgrave Macmillan, Basingstoke (2004)
19. Krennmayr, T.: Metaphor in newspapers, vol. 276. LOT, Utrecht (2011)
20. Lakoff, G., Johnson, M.: metaphors we live by. University of Chicago Press, Chicago (1980)
21. Mason, Z.: CorMet: a computational, corpus-based conventional metaphor extraction system. Comput. Linguist. **30**(1), 23–44 (2004)
22. Musolff, A.: Political imagery of Europe: a house without exit doors? J. multiling. multicult. dev. **21**(3), 216–229 (2000)
23. Niculae, V., Yaneva, V.: Conceptual considerations of comparisons and similes. In: Proceedings of the ACL Student Research Workshop, pp. 89–95 (2013)
24. Pragglejaz Group.: MIP: a method for identifying metaphorically used words in discourse. metaphor symb. **22**(1), 1–39 (2007)
25. Rundell, M. (ed.): Macmillan english dictionary for advanced learners. Macmillan, Oxford (2002)
26. Santa Ana, O.: 'Like an animal I was treated': anti-immigrant metaphor in US public discourse. discourse and society **10**, 191–224 (1999)
27. Semino, E., Hardie, A., Koller, V., Rayson, P.: A computer-assisted approach to the analysis of metaphor variation across genres. Paper presented at the Corpus Linguistics Conference 2009, University of Birmingham, July 2009
28. Skorczynska, H.: Metaphor in scientific business journals and business periodicals: an example of the scientific discourse popularization. Ibérica **3**, 43–60 (2001)
29. Skorczynska, H., Deignan, A.: Readership and purpose in the choice of economics metaphors. metaphor Symb. **21**(2), 87–104 (2006)
30. Steen, G.J., Dorst, A.G., Herrmann, J.B., Kaal, A.A., Krennmayr, T., Pasma, T.: A Method for linguistic metaphor identification: From MIP to MIPVU. John Benjamins, Amsterdam (2010)