Chapter 3 What the MIPVU protocol doesn't tell you (even though it mostly does)

Susan Nacey (Inland Norway University of Applied Sciences, Norway), **Tina Krennmayr** (VU University Amsterdam, The Netherlands), **Aletta G. Dorst** (Leiden University, The Netherlands) and **W. Gudrun Reijnierse** (Centre for Language Studies, Radboud University Nijmegen, The Netherlands)

The citation of the published chapter:

Nacey, S., Krennmayr, T., Dorst, A. G., & Reijnierse, W. G. (2019). What the MIPVU protocol doesn't tell you (even though it mostly does). In S. Nacey, A. G. Dorst, T. Krennmayr, & W. G. Reijnierse (Eds.), *Metaphor identification in multiple languages: MIPVU around the world*. Amsterdam: John Benjamins.

See also here: https://doi.org/10.1075/celcr.22.03nac

3.1 Introduction

Do I really have to do this for each and every word? But that will take forever!

Whether they be BA, MA or PhD students, postdocs or fellow researchers – anyone who is taught the basics of MIPVU eventually utters these words, or something very similar. We have held a number of metaphor identification courses at the Metaphor Lab's Summer and Winter Schools[1] and have taught the procedure to many cohorts of bachelor and master students from different universities. We have seen recurring aspects of metaphor identification using MIPVU that researchers who are new to the procedure typically find challenging. The clear step-by-step MIPVU protocol may create the impression that identifying metaphor is straightforward, easy and fast. However, while the procedure provides explicit instructions, we have found that applying them correctly and consistently requires explicit training, practice and experience.

The first part of this chapter discusses various 'nitty-gritty' practical aspects about the original MIPVU intended for the English language. Our focus in these first three sections is on common pitfalls for novice MIPVU users that we have encountered when teaching the

[1] The Metaphor Lab is an expertise center for metaphor studies in Amsterdam; see http://metaphorlab.org/.

procedure. First, we discuss how to determine what comprises a lexical unit (section 3.2). We then move on to how to determine a more basic meaning of a lexical unit (section 3.3), and subsequently discuss how to compare and contrast contextual and basic senses (section 3.4). We illustrate our points with actual examples taken from some of our teaching sessions, as well as with our own study into inter-rater reliability, conducted for the purposes of this new volume about MIPVU in multiple languages. Section 3.5 shifts to another topic that new MIPVU users ask about – namely, which practical tools they can use to annotate their data in an efficient way. Here we discuss some tools that we find useful, illustrating how we utilized them in our inter-rater reliability study. We close this part with section 3.6, a brief discussion about reliability testing.

The second part of this chapter adopts more of a bird's-eye view. Here we leave behind the more technical questions of how to operationalize MIPVU and its steps, and instead respond more directly to the question posed above: Do we really have to identify every metaphor in every bit of our data? We discuss possible approaches for research projects involving metaphor identification, by exploring a number of important questions that all researchers need to ask themselves (preferably before they embark on a major piece of research). Section 3.7 weighs some of the differences between quantitative and qualitative approaches in metaphor research projects, while section 3.8 talks about considerations when it comes to choosing which texts to investigate, as well as possible research areas where metaphor identification can play a useful role. We close this chapter in section 3.9 with a recap of our 'take-away' points – that is, a summary of the highlights from our entire discussion.

Our rationale for including this chapter in a volume devoted primarily to the application of MIPVU to languages other than English is that it is clearly necessary to have a shared understanding of what MIPVU is and how it is intended to work on English discourse before the procedure can be adapted to other languages. Moreover, all of the issues discussed here – including pitfalls, suggested tools, reliability tests and considerations when applying MIPVU – are applicable to all languages, even though the examples are all taken from English. Even the most English-centered section, section 3.2 on lexical units, is relevant to researchers interested in MIPVU for other languages, because the main point concerning the necessity of sharing a clear understanding of the unit of analysis is a crucial consideration for all MIPVU practitioners. Note that much of this chapter is written in an 'approachable' style, less formal than the other chapters in this volume. Our hope is that, in addition to contributing to a wider common understanding of the MIPVU protocol, this chapter will also prove useful as a resource for

novice researchers trying to master the procedure, as well as to advanced users who would like to teach the procedure to others.

Part 1: Practicalities of MIPVU

Part 1 of this chapter explores some of the details of MIPVU that novice users frequently overlook. Our discussion centers around three important steps of the procedure: demarcation of lexical units (section 3.2), determination of a more basic meaning (section 3.3), and comparison of contextual and basic meanings (section 3.4). Sections 3.5 and 3.6 then discuss practical matters concerning tools that may help in annotation and analysis, together with reliability testing.

3.2 How do I determine what comprises a lexical unit?

MIPVU uses the *lexical unit* as its unit of analysis. In most cases, the lexical unit is identical to an orthographic word – that is, a written sequence of letters with spaces at the end and none in the middle. Because of this frequent one-to-one equivalence, the term *lexical unit* is often used interchangeably with the term *word*. There are, however, a handful of exceptions to the general rule: certain types of lexical units consist of two or more orthographic words that form one semantic unity. Such multiword lexical units are considered by MIPVU as single lexical units, and the protocol provides detailed instructions for how to identify them: *polywords, proper nouns, compounds,* and *phrasal verbs* (see Chapter 2, this volume).

Demarcation of lexical units in a consistent way is important because it affects the total word count of the text at hand. This in turn impacts subsequent quantitative analyses, such as the determination of metaphor density, i.e. the number of metaphors per total number of lexical units in the sample. Such figures have been documented in past research, but prove impossible to compare in any meaningful way. Cameron (2003: 56-58), for example, surveyed nine studies from 1977 to 1999 and finds that the reported metaphor densities vary from 0.87 to 57 units per 1000 words. Inter-study comparison is hampered because researchers have not only employed different (sometimes unexplained) means of metaphor identification, but because they have also utilized varying units of analysis. For this reason, it is important to be explicit about the way in which lexical units were identified.

In the course of teaching MIPVU, we have observed a number of pitfalls with respect to the demarcation of lexical units. We thus decided to conduct a simple experiment whereby three experienced metaphor researchers – all of whom had previously published research involving the application of either MIP or MIPVU – independently applied MIPVU to two English-language newspaper articles, roughly 1,500 words.[2] In so doing, they each demarcated all lexical units into one of five categories: *word* (i.e. a single-word lexical unit), *polyword, phrasal verb, compound,* or *proper noun*. Then we compared their results: these are presented in Table 1 showing both the three-way and pairwise inter-rater agreement between the analysts.[3] We can see here that, for all intents and purposes, agreement between the analysts was nearly none.

Table 1 Three-way and pairwise inter-rater reliability for demarcation of lexical units

Analysts	**κ**	**95% confidence interval**	**Interpretation**
1-2-3	0.291	0.26-0.32	minimal
1-2	0.051	0.017-0.089	none
1-3	0.067	0.029-0.11	none
2-3	0.606	0.56-0.66	moderate

What went wrong? Analyst discrepancies concerning lexical units were mainly the result of procedural misunderstandings about how to identify characteristics of multiword lexical units, even though the MIPVU guidelines explain how to do so. By way of example, consider Table 2, which presents the three analysts' coding of the lexical units in sentence (1).

(1) Boris Johnson says Brexit will not be triggered straight away.

[2] These researchers are Linda Greve (Aarhus University, Denmark), Marlene Johansson Falck (Umeå University, Sweden), and Susan Nacey (Inland Norway University of Applied Sciences, Norway).

[3] Fleiss' kappa was calculated for three raters, while Cohen's kappa was calculated for two raters. As McHugh (2012) suggests, the value of kappa was interpreted as indicating the following level of agreement: 0-0.20 *none*; 0.21-0.39 *minimal*; 0.40-0.59 *weak*; 0.60-0.79 *moderate*; 0.80-0.90 *strong*, above 0.90 *almost perfect*. The kappa measures were calculated in R using the 'irr' package, while the confidence intervals were calculated using the bootstrap function 'boot' function in the 'boot' package, using the percentile method of estimating the CIs from the bootstrap; see our references for the full citations for R and both packages. Our data and R code are available as supplementary material at this volume's Open Science Framework website; see the Chapter 3 folder at https://osf.io/vw46k/.

Table 2 Sample demarcation of lexical units

Element	**PoS**	**Lexical unit demarcation**		
		Analyst 1	**Analyst 2**	**Analyst 3**
Boris	NP0	word	proper noun	word
Johnson	NP0	word	proper noun	word
Says	VVZ	word	word	word
Brexit	NN1	word	proper noun	proper noun
Will	VM0	word	word	word
Not	XX0	word	word	word
Be	VBI	word	word	word
Triggered	VVN	word	word	word
Straight	AV0	polyword1	polyword1	word
Away	AV0	polyword2	polyword2	word

When it comes to **proper nouns**, we see here that the analysts disagreed on the demarcation of *Boris, Johnson*, and *Brexit*. Analyst 1 only marked proper nouns as such if they consisted of two or more elements, and followed a certain stress pattern. Analyst 2, by contrast, marked all elements beginning with capital letters as proper nouns, while Analyst 3 was inconsistent. Which analyst is correct? According to the MIPVU guidelines, proper nouns must 1) be codified in dictionaries, 2) consist of two or more elements, and 3) have the primary stress on the first element. Otherwise, they are treated as individual components. In this way, items such as *Labour Party* should be classified as a proper noun, while neither *Boris* nor *Johnson* (nor *Boris Johnson*, for that matter), nor *Brexit* should be.

In Table 2, we also see that the analysts disagreed about the demarcation of *straight away*. Analysts 1 and 2 marked *straight away* as a **polyword**, and thus a single lexical unit. Analyst 3, by contrast, marked *straight away* as two individual lexical units. This is a disagreement that would lead to a difference in the total word count if left uncorrected. According to the guidelines, analysts should consult the *List of Multiwords and Associated Tags*

in BNC2: if the particular expression is on this list and is also annotated with a multiword tag (rather than alternative tag) by the Part of Speech tagger, it should be coded as a polyword and counted as a single lexical unit. In the case of *straight away*, we find it on the BNC list and see that it is also PoS-tagged as a polyword, i.e. with both elements tagged as a general adverb (AV0). It turned out that Analyst 3 appeared to check the multiword list either inconsistently or not at all, and thus inadvertently overlooked polywords.

Marking **compounds** also led to discrepancies. Compounds can be spelled as single orthographic words (in which case they are treated as one word; e.g. *shortlist*), as hyphenated words (e.g. *two-tier*), or as two separate elements (e.g. *per cent*). MIPVU considers hyphenated compounds as single lexical units if codified in dictionaries, and separate units if not codified: *two-tier*, for example, is not found in the dictionary and thus should be treated as two separate lexical units. Coding of such hyphenated compounds resulted in disagreement when analysts paid undue attention to the orthographic writing conventions (making it look like the term was a single unit), rather than to dictionary codification (indicating the term was actually two units).

Furthermore, MIPVU says that spaced compounds are counted as single lexical units if 1) they are codified in dictionaries *and* 2) the primary stress is on the first element. If the primary stress is on the second element, then the compound is considered two separate lexical units, despite codification. What this means is that a possible compound should not be marked as a single lexical unit just because it is codified in dictionaries. Instead, the stress pattern must also be checked. The main pitfall lies in this second step, which may easily be overlooked. Information about the stress pattern of compound nouns can be found in the printed copy of the Macmillan dictionary (the procedure's suggested 'go-to' dictionary), or by closely listening to the pronunciation provided in the online version of Macmillan.[4]

A similar challenge concerns the demarcation of **phrasal verbs**, because the MIPVU guidelines also contain two main criteria for their identification: 1) they must be codified in dictionaries, *and* 2) the particle needs to be annotated through PoS tagging as a particle rather than a preposition (in the latter case, the item in question is a prepositional verb, and thus two lexical units). The problem is that dictionaries are much more liberal in their definition of what constitutes phrasal verbs (and also compounds, for that matter). A common pitfall is that analysts mark an item as a phrasal verb because the dictionary says it is, even though it is actually some other type of multiword verbal construction. They neglect to check the part of

[4] The online version of the Macmillan dictionary is available at https://www.macmillandictionary.com/.

speech assigned to the particle. For word count and reliability purposes, it is important that analysts closely follow the MIPVU guidelines.

One benefit of MIPVU is that its application supposedly allows for comparability, either between analysts or across studies. At the heart of any such comparability is the unit of analysis, which must be the same. Based on our experiences teaching MIPVU and on our experiment comparing inter-rater reliability between experienced coders, we suspect that this is often not the case, in that studies purporting to having employed MIPVU do not apply the same standards when demarcating lexical units. As banal as it may sound, our advice is to read section 2.2 of the guidelines carefully and adhere to them when demarcating lexical units (see Chapter 2, this volume: from p. **Error! Bookmark not defined.**). Do not rely on spelling conventions when demarcating lexical items, and check the Multiword list, dictionary codification, stress patterns, and/or PoS tagging. Alternatively, read the guidelines but deviate from them; this is fine, as long as you explain what you did and why. This is also true when it comes to answering any questions where the MIPVU guidelines provide no clear answer, such as how to deal with numerals: Should they be counted as single lexical units (e.g. 51.9 = 1 unit) or as multiple lexical units (e.g. *fifty/one/point/nine* = 4 units). Decide how you will deal with such cases, explain, and be consistent. Transparency is paramount, allowing any quantitative measures you produce to be properly interpreted.

3.3 How do I determine a 'more basic meaning' of a lexical unit?

The MIPVU guidelines clearly state that "a more basic meaning of a lexical unit is defined as a more concrete, specific, and human-oriented sense in contemporary language use" (Chapter 2, this volume: **Error! Bookmark not defined.**). We have nevertheless found that identifying a more basic sense of a lexical unit is sometimes easier said than done.[5] In the next paragraphs, we discuss four common pitfalls related to this matter.

3.3.1 The 'it feels basic to me' pitfall

Dictionaries such as Macmillan are frequency-based, meaning that they list the most frequently used word senses first. They are primarily aimed at advanced learners of English, who – more

[5] Note that one challenge not discussed in this chapter involves cases where two meanings compete with each other, in the sense that each of them exhibits a different criterion for basicness and it is not clear which one should take precedence. This issue is discussed in Dorst et al. (2013) and Dorst and Reijnierse (2015).

often than not – need to know something about that most frequent sense. The challenge for MIPVU users is that the most frequent sense is often also the most salient sense for most of us, in most circumstances. It 'feels' basic because we are so familiar with it; if someone were to ask what the word means out of context, that is the one we would probably offer (see e.g. Deignan 2005: 95; van der Meer 1999). And so we find that people applying MIPVU for the first time tend to automatically select the most frequent meaning as the basic meaning. They neglect to carefully check the sense descriptions against MIPVU's criteria for a more basic meaning. Discussion 1, retrieved from an online discussion among participants at a Metaphor Lab Winter School exemplifies this type of reasoning. Here the participants were discussing the metaphorical status of the adverb *desperately* in (2), which annotator B had marked as metaphorically used.

(2) "I was *desperately* seeking answers as to WHY I was still having problems communicating about (…)."

Discussion 1

Participant A: Why is this metaphorical? Help!

Participant B: I guess it is because the basic meaning will be #1 in the dictionary which is "in a very worried or angry way" and the contextual will be #2 which is "very much".

While basic meanings *can be* the most frequent meanings (and thus listed first in the dictionaries), this is by no means always the case. In fact, we find that metaphorical senses are often more frequently used than the basic meaning(s) they are derived from. Especially learners and non-native speakers may not even have encountered the word employed with its basic meaning, even though they are familiar with the metaphorical sense. Nacey (2013: 99), for example, illustrates this point by the verb *undermine*, which Macmillan defines with two distinct sense entries:

1. to make something or someone become gradually less effective;
2. to dig under something, especially so that it becomes weaker.

The more basic meaning is the second entry, because it concerns physical excavation beneath something concrete. But the first entry is the verb's most frequent meaning, making it the most salient for most people (except, maybe, for builders?). It is a metaphorical extension of the more concrete meaning, and should therefore not be mistaken for the basic sense by virtue of appearing first (see also Chapter 14 on English as a Lingua Franca, this volume).

3.3.2 The 'but there **has** to a single basic meaning' pitfall

Another source of confusion is that researchers new to the procedure may think they need to find *the* most basic meaning and get stuck when trying to determine which of a word's various sense entries is more basic than all the others. A good example of this is the verb *to serve*, which is defined with nine sense entries in the Macmillan dictionary, ranging from entry 1 *provide food/drink* to entry 9 *hit ball to start play*. Both of these senses are concrete, as are (at least) two others: 7 *help customers in shop* and 8 *officially give document*. While the novice MIPVU user might struggle to determine which of these sense is *the* most concrete (and hence, the most basic), all four of these senses share the same underlying meaning of giving something (concrete) to someone; in other words, these four senses mentioned here are *equally basic*.

Similarly, in the case of the adverb *desperately*, both entries pointed out by Participant B in Discussion 1 denote intensity of feeling and/or action, and neither can be said to be more basic than the other. Our conclusion is therefore that *desperately* is not metaphorically used, because there is no sufficient distinction between the contextual sense (sense 1 *in a very worried and angry way*) and the more basic sense (which encompasses sense 1).

3.3.3 The 'no contextual meaning' pitfall

Another stumbling block for novice coders is when they find a more basic meaning of the lexical unit in question but the contextual meaning is not listed in the dictionary. We illustrate this by our students' online discussion when deciding whether *mining* in (3) is metaphorically used.

(3) (…) journalists are *mining* this rich new vein of material…

Discussion 2

Participant A: Here I had the problem of only one meaning.

Participant B: Considered this one as metaphorical as well, with the basic meaning of *to mine* being to work at a mine, which is different from the contextual meaning.

Participant C: Yes, I agree but the dictionary as a matter of fact does not report this second one.

The contextual sense of *mining* is not listed in the dictionary, yet clearly differs from its more basic (physical) sense, something Participant B specifically notes. This type of situation is not unusual. Dictionaries are finite, and cannot possibly include a sense entry and illustrative sentence for every lexeme in every context in which it may appear. Our conclusion is thus that *mining* in (3) is metaphorically used. Its contextual meaning (about journalists taking information from a source) is sufficiently distinct from its basic meaning ("to dig a large hole or tunnel in the ground in order to get coal, gold etc., or to take coal, gold etc. from such a hole or tunnel"), and the two meanings are related by comparison: we understand investigating information in terms of mining for a valuable concrete entity.

When we come across cases where the contextual senses are not codified, we could decide to mark them as novel metaphors. At a later point, we may want to further investigate all such 'novel' lexical units, to differentiate those uses that are frequent in contemporary language but not yet captured in dictionaries from those uses that are indeed rare in use. Deignan (2005), for example suggests that corpus frequencies may be taken as an indication of novelty: "any sense of a word that is found less than once in every thousand citations of the word can be considered either innovative or rare..." (Deignan 2005: 40). Another possibility is that the term in question is not novel, but rather represents technical, subject-specific jargon, which is for this reason not included in general language dictionaries like Macmillan, intended for language learners. A possible solution here would be to consult a specialist dictionary for such terms, in addition to more general dictionaries.

3.3.4 The 'grammatical category / word class' pitfall

A final aspect that novice users of MIPVU often miss is that "a more basic sense has to be present for the *relevant grammatical category of the word form* as it is used in context" (Chapter

2, this volume: **Error! Bookmark not defined.**; italics in the original). This means that grammatical patterns such as transitivity of verbs and countability of nouns are taken into account when determining metaphoricity. MIPVU focuses on the referent in the exact context in which it appears. A typical example causing problems is the use of the verb *go* in (4).

(4) Louise had *gone* completely blind before she did (Macmillan, sense 5)

Many new users of MIPVU may correctly identify the verb's contextual sense as the fifth entry in Macmillan, "to change to another condition, usually a worse one". They then zero in on Macmillan's first entry for *go* as its basic sense: "to move or travel to a place that is away from where you are now", and subsequently mark the lexeme as metaphorical in use. The problem with such an analysis is that the basic meaning is intransitive, whereas as the lexical unit in 0 is a linking verb. Because they follow different transitivity patterns, the two senses should not be compared when following MIPVU, and 'go' should not be identified as metaphorically used in this example.

In a similar vein, MIPVU does not cross word class. Consider the noun *whole* in (5), an excerpt from a newspaper article in the VU Amsterdam Metaphor Corpus.

(5) Thus the Palestinian national movement both inside and outside the occupied territories is an organic *whole*.

According to the MIPVU procedure, the basic meaning first needs to be established. But when typing *whole* into the search bar of the online version of Macmillan, the entry that first comes up is the adjective rather than the noun. Based on the adjective's more basic meaning "not divided or broken", inexperienced MIPVU users may wrongly conclude may the use of *whole* in (5) is metaphorical. However, MIPVU requires checking meanings within the word class of the lexical unit as used in context. In this case, the meanings of the noun need to be checked. The noun has only one meaning, namely "a complete thing made of several parts", which applies to both abstract and concrete things. *Whole* in (5) is thus not metaphorically used.

The question of whether the distinction between word class boundaries or other grammatical categories should be taken into consideration when determining metaphoricity was a matter of contention during the development of MIPVU. As discussed in Chapter 1 of this volume (p. **Error! Bookmark not defined.**), the original MIP developed by the Pragglejaz Group (2007) did not have any such restriction of staying within the same grammatical

category: they illustrate their rationale with the noun/verb pair of *squirrel* and *to squirrel*, contending that treating these as two distinct lexemes would entail a loss of the clear metaphorical link between them (see also Deignan 2006).

Moreover, reasons relating to genre, style, rhetorical purpose or communicative function have sometimes led researchers to formulate additional guidelines on when to 'break' this rule. Say a novel consistently describes characters in terms of animal behavior and characteristics through verbs such as *rabbit on* or adjectives such as *bitchy*. In this case, researchers may want to disregard the word-class boundary rule when identifying metaphor on the grounds that such uses reflect part of one systematic and coherent stylistic pattern of (deliberate) animal metaphors. This is clearly a very different situation than when a single instance of *bitchy* occurs in a casual conversation between friends in which no other animal metaphors appear. As always, the golden rule when deviating from the MIPVU guidelines is to be explicit and systematic when formulating additional guidelines, and to explicitly report reasons for breaking specific rules (or including/excluding specific cases) in any publications.

3.4 How do I go about contrasting and comparing meanings?

Given that a particular word under investigation has separate entries for its basic and contextual senses, some novice MIPVU users automatically jump to the conclusion that this means they have found a metaphorically used word, without any further consideration. The problem with this is that the existence of two differently numbered senses in the dictionary does not guarantee the presence of metaphor. It is necessary to complete all the MIPVU steps, to make sure that the meanings are 1) related by similarity and 2) sufficiently distinct. Separate sense entries may be (and often are) related to each other through metaphor, but they may alternatively be related through some other relationship such as metonymy, specification, or generalization.

The crucial distinction between metaphor and metonymy concerns the contrast between similarity and contiguity. Contiguity is at the core of metonymy, a form of co-occurrence whereby we view X *via* Y. By contrast, viewing X *as* Y lies at the heart of metaphor (Steen 2007: 58-61). For example, consider the lexical unit *soul* in (6), also a topic of discussion among students at a Metaphor Lab Amsterdam Winter School.

(6) You have saved my *soul*.

Looking in Macmillan, we find that the basic and contextual meanings are represented in separate sense entries:

1. the part of a person that is capable of thinking and feeling (the basic sense);
2. a person, e.g. *I promise I won't tell a soul.* (the contextual sense).

Discussion 3

Participant A: I think this is indirect metaphor for 'person' (macmillan). Like, she saved herself.

Participant B: I agree.

Participant C: I think "soul" is metonymy here, not metaphor.

Although not explicitly mentioned in this student discussion, all three participants probably (correctly) identified the first sense in Macmillan as more basic than the contextual 'person' meaning: although both are human-oriented, the former is more specific than the latter. Participants A and B are also correct in realizing that the first and the second sense are related. However, the relationship between the two is one of contiguity rather than similarity. We view the person *via* the soul (metonymy) whereby the soul 'stands for' the person, rather than viewing the person *as* the soul (metaphor).

The occurrence of separately numbered sense descriptions frequently indicates that those senses are sufficiently distinct for serving as a basis for a comparison. However, sometimes discrete senses are more finely distinguished from each other (see Lew 2013). Take, for example, the noun *melee* in (7). One Metaphor Lab Winter School participant marked it as metaphorically used, which prompted disagreement from two of his fellow students.

(7) Ever since the field emerged from the postwar cybernetic melee (…)

This noun is defined by two sense entries in Macmillan:

1. a noisy confused fight involving a lot of people, and
2. a large confused group of people or things.

Discussion 4

Participant A: Why melee is coded as metaphor? the basic and contextual meanings are considered as distinct enough?

Participant B: same comment here

Participant A: Macmillan has provided two meanings for melee, that to me there are very similar

We agree with the reasoning of Participant A, who was asking the right questions. These two sense entries are not sufficiently distinct; the first meaning is simply more specific in that it refers not just to a confused group, but a confused fight. The noun *melee* in the context of 0 should have not been marked as metaphorically used.

Thinking 'I am done and have found a metaphor' once two separate sense descriptions have been found is a common pitfall for novices. As illustrated in the above examples, however, it is crucial to complete the final steps of MIPVU and check whether the meanings are related by similarity and whether they are sufficiently distinct, as entries may simply be connected through non-metaphorical relationships.

3.5 Which tools should I use to annotate my dataset?

It is no secret that MIPVU is a time-consuming affair, because the analyst must consider each lexical unit for metaphoricity. Although we cannot remove the need for manual decisions, it is possible to speed up analysis through systematizing it in a logical way. The tools used for annotating your dataset require careful consideration before you start on a project involving metaphor identification using MIPVU.

One important consideration is the long-term plans you have for your dataset: for example, whether there are plans to make it public and in what form. For publishing annotated data, the XML format has emerged as a standard, and free XML editors are available.[6] For deliberations on physical representation formats and annotation environments, see Krennmayr and Steen (2017).

[6]A list of XML editors is available at https://en.wikipedia.org/wiki/Comparison_of_XML_editors.

While the thought of using XML may be off-putting to some, there are other, quite simple ways to code data. For instance, many metaphor researchers use spreadsheet programs like Microsoft Excel. The text is entered vertically into a column, with each line representing a lexical element. All sorts of columns can be added, depending on the particular aims of your research project, such as information about Part-of-Speech tags and decisions about metaphor. An advantage of Excel is that is allows for easy manipulation of your data, both before and after analysis. In the pre-analysis stage, you ideally need a tool that allows you to automatically enter your unannotated (maybe PoS-tagged) text from other programs such as Word or Wmatrix (a corpus analysis tool including PoS and semantic tagging systems).[7] In the post-analysis stage, you will want to be able to transfer your data into other programs to allow for further (statistical) analysis or visualization of your findings. Compatibility between programs is therefore a second important consideration when it comes to choosing tools for annotating your data.

In what follows, we present one illustration of tools for annotation and subsequent statistical analysis and visualization that we have used when applying MIPVU to texts, where we align three tools: Excel, the Filemaker Pro Advanced database application, and the R software environment for statistical computing and graphics.[8] Another possibility worth looking into is provided by a University of Lancaster team involved in the "Metaphor in end-of-life care" project (Semino et al. 2018). They identify metaphor with a modified version of MIPVU, using a combination of Wmatrix, the eMargin collaboration online annotation tool, and Excel.[9]

3.5.1 Initial independent MIPVU analysis

Besides checking inter-rater reliability on the demarcation of lexical units as reported in section 3.2, we also investigated the agreement about metaphoricity of the three analysts taking part in our small experiment. The entire study involved two rounds of analysis, with a troubleshooting session in-between – the overall goal being to determine if inter-rater reliability was 1) satisfactory, and 2) could be improved. Figure 1 presents a screenshot of our Excel spreadsheet containing the 'Boris Johnson' sentence in (1) and represents the coding of one of the three

[7] Wmatrix is available at http://ucrel.lancs.ac.uk/wmatrix/. See also Chapter 1, this volume.
[8] Filemaker Pro is available at http://www.filemaker.com/. R is available at https://www.r-project.org/.
[9] MELC presentation slides from a workshop on using corpus methods to analyze metaphor are available at http://ucrel.lancs.ac.uk/melc/workshop_jan2014.php. The eMargin annotation tool is available at https://emargin.bcu.ac.uk/.

analysts, Analyst 1.[10]

A	B	C	D	E	F	G	H	I	J	K	L	M	N
ID	PoS	element	lexunit	not met	indirect met	direct met	implicit met	Mflag	WIDLII	DFMA	extra element	punct.	count
556	NP0	Boris	w	x									1
557	NP0	Johnson	w	x									1
558	VVZ	says	w	x									1
559	NN1	Brexit	w	x									1
560	VM0	will	w	x									1
561	XX0	not	w	x									1
562	VBI	be	w	x									1
563	VVN	triggered	w		x								1
564	AV0	straight	p	x									1
565	AV0	away	p								x		1
566	PUNC	.	i									x	1
567	NP0	Boris	w	x	x								2
568	NP0	Johnson	w										0
569	VVD	said	w										0
570	NP0	Britain	w										0
571	VM0	should	w										0
572	XX0	not	w										0
573	AV0	immediately	w										0
574	VVI	trigger	w										0
575	NN1	article	w										0
576	CRD	50	w										0

Figure 1 Excel spreadsheet screenshot

We briefly explain what each column represents:

- Column A assigns each element a unique identification number.
- Column B shows the PoS tag for each element; following the suggestions in the MIPVU guidelines, we had first annotated our text with the BNC C5 tagset using the CLAWS part-of-speech tagger for English.[11]
- Column C contains the lexical elements imported from a Word document with the text under investigation. Most of these elements are the same as individual lexical units, but some (like *straight* and *away*) form a single lexical unit because they are two elements in one polyword (see next point).
- Column D contains our codings for lexical unit. The default code (by virtue of being most frequent) is *w* for 'word', a single lexical unit. Other possible options are *p* (polyword), *v* (phrasal verb), *c* (compound), *n* (proper noun), and *i* (ignore, for punctuation).
- Columns E-K consist of one column for each of the codings possible within MIPVU (not metaphorically used, indirect metaphor, direct metaphor, implicit metaphor, metaphor flag,

[10] A suggested template for MIPVU data analysis is available as supplementary material at this volume's Open Science Framework website; see the Chapter 3 folder at https://osf.io/vw46k/.
[11] CLAWS is available at http://ucrel.lancs.ac.uk/claws/.

WIDLII (When In Doubt Leave It In), and DFMA (Discard From Metaphor Analysis). Analysts are expected to enter a lowercase *x* in the appropriate column.

- Column L is intended to mark extra elements of multiword units, exemplified here by *away* which is the second element in a polyword. Only the first element, *straight*, is coded for metaphor use of the entire polyword.
- Column M records whether a line contains a punctuation mark.
- Column N provides a safeguard against inadvertent mistakes in metaphor coding, a hazard of the trade when one analyzes large amounts of text. A formula is used here to count the number of *x*'s in Columns E-M: the number *0* shows that the lexical element in question has not yet been analyzed for metaphorical use (e.g. ID numbers 568-577); the number *1* shows that the lexical element has been given a single code for metaphorical use (e.g. ID 556-566); the number *2* (or greater) indicates that the lexical unit has been coded more than once for metaphorical use (something that has not happened in the illustrative example in Figure 1). When an analyst finishes an initial round of metaphor identification, they should then use Excel's sorting function to uncover any lexical elements that were inadvertently overlooked (*0*) or double-coded (*2* or higher), and then correct them before any subsequent analysis.

This spreadsheet contains the most basic information needed for MIPVU analysis. It can thus functions as a starting template for metaphor identification projects. Additional columns can be added, of course, depending on the project's particular objectives. As an example, a column identifying the producer of each lexical unit would be crucial for any project involving more than one informant.

3.5.2 Further comparative MIPVU analysis

While Excel is a sufficient tool for many projects, we have found that adding too many columns makes it unwieldy to use because it may not be possible to view all the columns without scrolling to the right or left on the screen – all of which takes time. If you plan on coding for a larger number of variables, you might consider transferring your Excel data to a different type of database, such as FileMaker.

Figure 2 shows a screenshot illustrating the use of FileMaker Pro Advanced to create a custom-made database that suits our needs. The screenshot displays our analysis for *Boris*, ID 556 in the Excel spreadsheet presented in Figure 1. Here we find the analysis for *Boris* shown

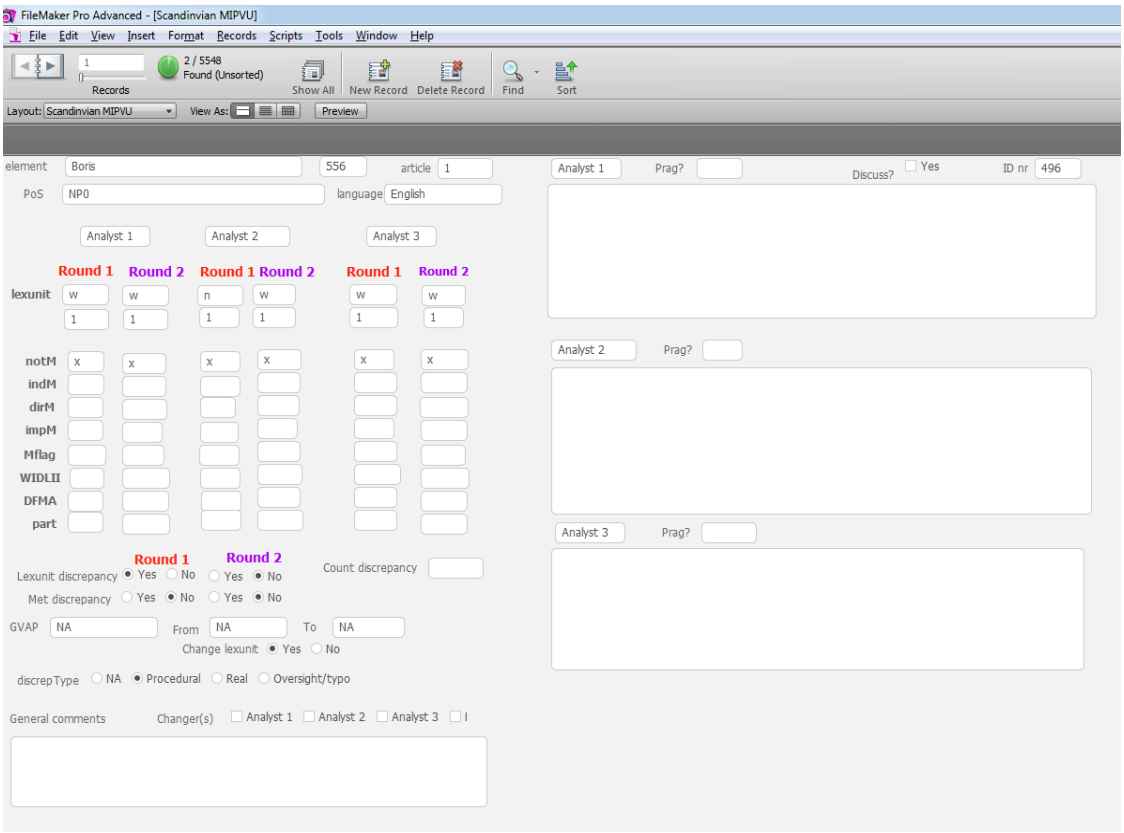


Figure 2 FileMaker Pro Advanced database screenshot

in line 556 in Figure 1 that has been imported to this new database; we can recognize it under the *Round 1* column for *Analyst 1*, where we see the *lexunit* (lexical unit) has been coded *w*, and there is an *x* indicating that this analyst marked the words as *notM* (not metaphor). Similarly, we have imported all information about this word from all three researchers and for both rounds of analysis, taken from the Excel spreadsheets they used to independently conduct their analyses.

This database was actively used in the course of our inter-rater reliability study (see section 3.6). The data from the three researchers was initially imported after their first rounds of analysis, and used to easily identify any discrepancies among analysts in their decisions about either the demarcation of lexical units or determination of metaphoricity. Such discrepancies were then discussed in a troubleshooting session before a second round of analysis. We see towards the bottom of Figure 2, for example, that a *lexunit discrepancy* is indicated under *Round 1*, and that the analysts concluded in their troubleshooting session that this resulted from a procedural misunderstanding, rather than any 'real' disagreement (the *discrepType*). In *Round 2*, however, the analysts were in agreement.

3.5.3 Statistical analysis of MIPVU results

The final step in our study into inter-rater reliability was to generate findings based on the analyses recorded in our cumulative database. To do so, we transferred the results from our Filemaker Pro database into the R software environment, where we were able to calculate inter-rater agreement for the demarcation of lexical units and for the identification of metaphor for both rounds of analysis. R requires users to write their own code, and may therefore seem rather daunting for people unfamiliar with it. There are, of course, alternative programs for statistical analysis, but any tool requires some initial investment to learn how to properly use it. We find that our combined approach using Excel, Filemaker Pro, and R provides an effective and flexible means generating findings from MIPVU analysis.

3.6 Reliability testing

A metaphor identification procedure such as MIPVU may be considered a reliable tool only if its application leads to substantial agreement between different coders. How can we tell if that is the case? Suppose we have a metaphor researcher who applies MIPVU to a text that is 5,000 words long. Following the protocol, she assigns one of the following categories to each word: 1) not metaphor, 2) indirect metaphor, 3) direct metaphor, 4) implicit metaphor, 5) metaphor flag, 6) WIDLII, and 7) DFMA. Ideally, we would like to know how reliable her metaphor identification process has been, because reliability tells us something about whether we can trust the results. To find out, we could ask a second metaphor researcher to replicate her work. We would ask this second person to independently apply MIPVU to the same text, and then compare results (similar to what we have done in our reliability experiment reported in this chapter).

One way of comparing results would be to calculate the percentage of agreement between our researchers. We might then find that out of the 5,000 words, the two annotators agreed more than 95% of the time about their decisions to identify a word as belonging to one of the predetermined categories. This might seem impressive at first sight, we need to consider what we have just calculated before we declare ourselves satisfied with our results. We know from previous empirical studies that there is far more non-metaphorical language than metaphorical language in discourse (see Chapter 10 in Steen et al. 2010). Even though metaphor is ubiquitous in language, most of the words in the sample will be categorized ‘not metaphor’. Our researchers are (and should be) positively biased towards judging any particular word as

not metaphorical in use. But what would happen if we replace our second metaphor researcher with a monkey who has been taught to hit the 'not metaphor' key all the time? Because most words are not metaphorical in discourse, the percentage agreement between our first researcher and our trained monkey would still be rather high.

A more informative measurement of inter-rater reliability is the kappa: Fleiss' Kappa is frequently used to measure agreement between three or more analysts, while either Fleiss' Kappa or Cohen's Kappa are typically used for agreement between two analysts (see e.g. Steen 2007: 124-125). What the kappa does is correct for chance. Given that so much discourse is not metaphor, it is not really surprising if our two researchers agree that any particular word is not metaphor. But the chances of a word being indirect metaphor are much lower, so if our researchers agree on that coding, we have more reason to be pleased. Direct metaphors, implicit metaphors, WIDLIIs and (in most text types) DFMAs are even rarer than both not metaphor and indirect metaphor. So if our researchers also agree on these codings, then we can take this as a true indication that they are applying MIPVU in similar ways. The kappa will reflect this by placing more weight on those 'trickier' cases where we actually have to make a decision that is not the default 'not metaphor'. Researchers who provide kappa values will also typically explain which cut-off points they have used to interpret the degree of agreement (i.e. minimal / weak / moderate / strong / almost perfect), often referring to past precedent (see Howell 2010: 165-166 for more information about the kappa).

We also strongly recommend providing the 95% confidence interval (CI) for the kappa value. In our example with two metaphor researchers, the one kappa value tells us about the agreement between them for their first 5,000 words. But if we were to ask them to repeat the same task by identifying metaphor in another set of 5,000 words, they probably would arrive at a different kappa value. We want the kappa to reflect the degree to which they *generally* agree about metaphoricity when employing MIPVU, rather than just their agreement about one particular sample. To find this 'true' kappa, we would actually have to ask them to conduct an infinite number of metaphor identification analyses. A more practical alternative would be to provide the CI, which gives the range within which we are 95% certain that this true kappa lies; there is no need to doom our researchers to an eternity of metaphor identification.

Part 2: Choosing your approach and your data

The sections above have discussed some of the main challenges in applying MIPVU, and issues relating to data entry and reliability testing. However, we should go back to our initial question: 'Do I need to use this on each and every word in my dataset?' The answer, of course, is 'Not necessarily'. Going manually through hundreds, thousands, even millions of words is normally far beyond the capacity of any individual researcher; even collaborating researchers often do not have the time or means to take on such an endeavour. The discussion below will help guide you through the different possibilities of approaching metaphor identification projects using MIPVU, suggesting when a particular approach or perspective makes most sense.

3.7 Decision 1: Quantitative, qualitative, or both?

Decisions regarding which research approach to adopt and which selection of texts to make should be informed by research questions and hypotheses. One of the first questions to consider is: 'Can I best answer my research question(s) by carrying out a quantitative analysis (*How many? How often?),* a qualitative analysis (*How? When? Where? Why?),* or a combination of the two?' The answer to this question may have consequences for the selection of texts – the kind and the number of texts that will (or realistically can) be included in your dataset.

Especially in the case of quantitative analyses, it is important to pay attention to the selection of texts or text excerpts if the goal is to generalize to a population. In that case, the sample that is studied needs to be representative of a particular type of language use. This language use can be very general (e.g. contemporary English; Dutch used in literary discourse), or highly specific (e.g. Emily Dickenson poems about death; the *obiter dicta* of Supreme Court tort law rulings), or anything in-between. You may want to focus on a particular language or dialect, a specific domain or genre, a text type or specific subject matter, a time period or geographical region, a specific author or politician – in principle, anything goes. The broader the language variety, the rarer the phenomenon you are interested in, and the more quantitative your approach, the more texts you will need to obtain a representative selection and to ensure that your findings will be consistent across different samples, rather than idiosyncratic. However, representativeness is a tricky concept: in the end, the texts you analyze are truly representative of nothing but themselves.

When taking a qualitative, instead of a quantitative, approach to your research question(s), the goal of analysis is concerned with finding patterns that (re)occur throughout a text, or a (small) collection of texts, and describing those patterns in a detailed way. In contrast

to quantitative analyses, the aim is not to examine how often a phenomenon occurs, but to describe *how* it occurs. This, too, may affect how a dataset is collected. Independent of whether your approach is mainly quantitative or mainly qualitative, we recommend to include brief demonstrations in your publications to explain how the MIPVU works and how it is useful for your particular research question(s) and/or hypotheses.

3.8 Decision 2: Which (elements in) texts and why?

Both quantitative and qualitative metaphor analysis may be used to answer a range of research questions and to test a variety of hypotheses. One particular perspective you may wish to take is to explore differences between languages. For English, a number of quantitative studies on different genres and registers have been carried out. An example is the "Metaphor in Discourse" project (Steen et al. 2010), where MIPVU was developed and used to manually identify all metaphor-related words (MRWs) in an almost 200,000-word corpus of contemporary British English, divided between roughly 50,000 words of everyday conversations, fiction, news and written academic discourse. In your own project, you may, for example, wish to find out if the language you are investigating has similar distributions of metaphors across one or more of these different registers. Perhaps you expect your language to use more or fewer metaphor-related words overall, or to use more or fewer MRWs in one or more of the four registers (e.g. you may expect German news texts to have fewer MRWs than English news texts), or to use more or fewer MRWs of a particular type (e.g. you may expect Japanese novels to have more direct metaphors than English novels). Rather than wanting to establish frequencies for the entire language (which seems a mission impossible par excellence), or even for all four types of discourse included in the "Metaphor in Discourse" project (which had the luxury of having a team of four full-time PhD students for five years), you could focus on just one type to investigate whether this particular register is subject to different norms for metaphor usage in different languages.

Of course, a register perspective can also be taken without the need for a comparison between languages. Instead of comparing metaphor use in the same register between languages, you may want to investigate how metaphor usage differs between two registers within the same language, or explore metaphor usage in thus far unexplored or underexplored areas. For example, you may wish to establish how (particular types of) MRWs are used in pop songs or

legal contracts, or you may be interested in the occurrence or absence of metaphor in new forms of communication such as tweets or e-consulting.

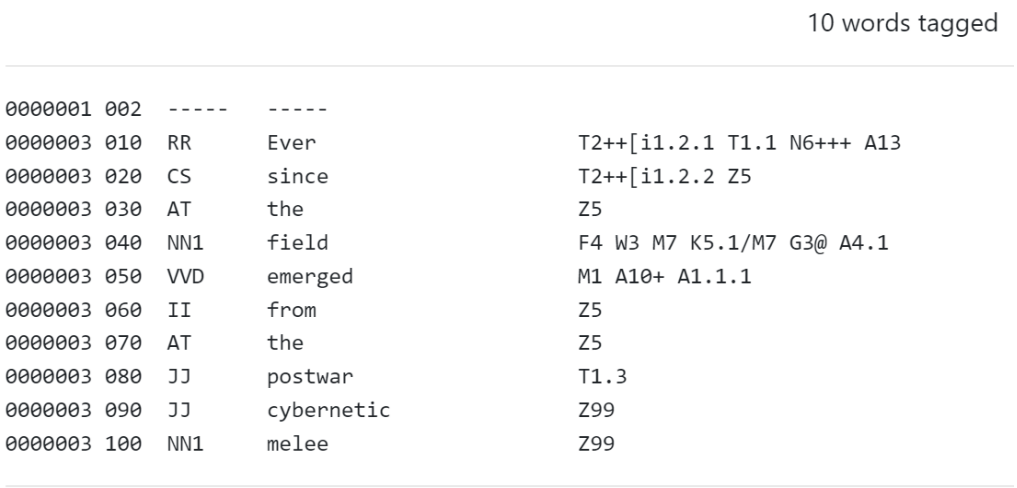
One point of critique that is sometimes lodged against the results of Steen et al. (2010) is that the reported frequencies include 'everything', including 'uninteresting' words such as deixis (e.g. *this, that*), prepositions (e.g. *in* 2016, talk *about*), delexicalized verbs (e.g. *take, make, give, get*), and empty nouns (e.g. *thing, stuff, end, point*). Whether or not to take into account these words depends on the specific questions you want to answer in your project. In some cases, it may make most sense to focus only on content words (noun, verbs, adjectives, adverbs) and leave the grammatical words for what they are. Or you may be interested in only one word class, say verbs, to establish how many of them are metaphorically used when high-frequency delexicalized verbs are not taken into account, and how this distribution differs per genre or register or author. With such a focus, you may still need to go through your texts manually, but your work will considerably speed up and be far less daunting than analysing each and every word.

Many researchers are merely interested in a particular source domain, e.g. ANIMAL metaphors, FIRE metaphors, or WAR metaphors, especially those working within the framework of the Conceptual Metaphor Theory. It is important to note that MIPVU only identifies metaphors on the *linguistic* level, not the conceptual level nor the cognitive level (production/processing). As such, the method does not make any claims about underlying conceptual metaphors. However, although MIPVU is not meant to identify conceptual metaphors or mappings, the steps of the procedure do invite the researcher to think about the potential mappings and underlying conceptual metaphors. This is because researchers must first establish the contextual meaning of a lexical unit (which relates to the target domain), then examine whether the lexical unit has a more basic meaning (which relates to the source domain), and finally determine whether the contextual and more basic senses are distinct (and thus belong to different domains) as well as related via comparison (that is, a cross-domain mapping).

As a result, MIPVU can be used as a very first step towards conceptual analysis. It should be noted, though, that the transition from linguistic to conceptual metaphor is far from straightforward (see e.g. Steen 2009). One online tool to identify potential source domain concepts is the Ucrel Semantic Analysis System (USAS).[12] Researchers can run their texts, or identified metaphors, through the tagger, which will return a list of tags for each of the words. By way of illustration, Figure 3 displays the USAS output for the sentences in discussed

[12] The USAS tagger is available online at http://ucrel-api.lancaster.ac.uk/usas/tagger.html.

previously in example (7): “Ever since the field emerged from the postwar cybernetic melee”. Here we see columns for the part of speech code, the token in question, and the semantic code(s) selected by USAS (with statistically most likely code being listed first).[13]



10 words tagged

0000001 002	-----	-----	
0000003 010	RR	Ever	T2++[i1.2.1 T1.1 N6+++ A13
0000003 020	CS	since	T2++[i1.2.2 Z5
0000003 030	AT	the	Z5
0000003 040	NN1	field	F4 W3 M7 K5.1/M7 G3@ A4.1
0000003 050	VVD	emerged	M1 A10+ A1.1.1
0000003 060	II	from	Z5
0000003 070	AT	the	Z5
0000003 080	JJ	postwar	T1.3
0000003 090	JJ	cybernetic	Z99
0000003 100	NN1	melee	Z99

Figure 3 USAS output screenshot

While this tool is potentially useful to systematically (rather than intuitively) add source domain labels to linguistic metaphors, the number of lemmas in the database remains somewhat limited (especially for languages other than English), leading to a frequent use of the ‘Z99’ tag (the USAS code for ‘unmatched’). As is shown in Figure 3, domain tags for words like ‘cybernetic’ and ‘melee’ are absent from the system. In addition, the USAS output often provides more than one domain, and as such leaves it to the analyst to determine which of those counts as ‘the’ source domain. An example of this is the output for the noun ‘field’ in Figure 3 above. This noun is tagged with more than five different (sub)tags, among which ‘Farming and horticulture’ (F4), ‘Geographical terms’ (W3), and ‘Sports’ (K5.1).

By contrast, researchers interested in examining how a particular source domain is expressed linguistically (i.e. taking a top-down perspective), can look for specific words belonging to the semantic field of that source domain, instead of analyzing each and every word in all of their texts (see e.g. Koller 2002). Previous studies or a thesaurus may be used to generate a list of possible candidates (that is, lemmas expressing the source domain under investigation) and then check their sense descriptions in the dictionary to see whether they indeed have a basic sense relating to the source domain of interest. In a subsequent step, the researcher can search his/her dataset for these items, and examine whether they are used

[13] USAS employs a tagset of 21 major discourse fields, divided into 232 different subcategories; see Archer et al. (2002).

metaphorically given the context in which they are used. Using software such as AntConc will allow researchers create concordance lines for relevant keywords and can also give information about frequencies and patterns.[14] Another related possibility is to use a 'small corpus / large corpus' approach, where you first manually analyse a small sample of your dataset to collect relevant metaphorical expressions, and then run fully or partially automated searches for these expressions in your entire dataset (see e.g. Cameron and Deignan 2003).

While some researchers are interested in particular source domains and start from there, others will look for specific target domains to find out what source domains are used to metaphorically describe the target topics they are interested in, for example, which metaphors are used to talk about migration, the euro, education or cancer (see e.g. Stefanowitsch 2006). In such cases, a dataset can be searched in similar ways to the source-domain approach described above, either manually combing through the texts to identify the metaphorical expressions (bottom-up), or starting from a list of metaphorical expressions and searching for them in the dataset (top-down), or using a small corpus/large corpus approach (see e.g. Deignan 2005). Depending on your research aims, you may even want to look for specific expressions or forms of metaphor signalling: for example, you could look for "consider education a*" or "new currencies are like*" (where the asterisk signifies a wildcard expression), thus revealing potentially deliberate metaphor (see Reijnierse et al. 2017). A very large corpus may be required to collect a considerable number of instances, depending on the search words and the specificity of the expression. On the other hand, searching for all instances of "looked like a*" may very well yield a many hits.

In the end, your research questions and hypotheses will determine whether a focus on the kinds of source domains that are used to describe a specific target might be best, or whether it might be more appropriate to investigate which target domains are described through a specific source domain. This type of consideration will help determine whether you need a small or a large dataset, as well as whether you are able to (or need to) to go through the data manually or whether you can rather utiliza search/concordancing software. Again, both quantitative and qualitative approaches (or combinations) are possible: e.g. 'Are WAR metaphors more frequent than JOURNEY metaphors to describe cancer?' (quantitative) versus 'Are WAR metaphors used positively (empowerment) or negatively (loss of agency) to describe cancer?' (qualitative). This analysis may then also be carried out comparatively between registers or domains or languages or time periods, etc. The possibilities are endless.

[14] AntConc is a free corpus analysis toolkit for concordancing and text analysis; see http://www.laurenceanthony.net/software/antconc/.

3.9 Concluding thoughts

A large part of this chapter has dealt with pitfalls into which, in our experience, novice MIPVU users frequently stumble. We would nevertheless like to stress that difficult cases typically form only a small part of any dataset. Most cases are unproblematic, a fact that might be easy to forget when reading a chapter that primarily focusses on the tricky cases and the finer procedural details.

Further, we believe that there is no need for researchers to slavishly follow every step of MIPVU, exactly as set out in the procedure. There may be good reasons to deviate from the protocol, a point we have emphasized at various points in this chapter. But you should first have a thorough knowledge of what the procedure calls for, in order to truly understand that you are indeed deviating from the protocol as it stands, in which way(s), and why. Part of the rationale for the development of MIPVU was to promote greater transparency in metaphor identification, and to allow for greater comparability between findings in different studies. If researchers who do not fully understand the procedure nonetheless claim that they have used it, both of these objectives are undermined. Transparency is threatened because we then cannot know how metaphor was actually identified in any given project, which in turn means that we have reason to doubt the validity of cross-study comparisons. We therefore hope that this chapter will contribute towards accomplishing the objectives of realizing a valid and reliable means of metaphor identification.

To end, we would like to recapitulate the highlights of this chapter – the most important points that we want novice MIPVU to retain about the basics of the protocol:

About demarcation of lexical units:

- Do not rely on spelling conventions.
- Remember to check the BNC Multiword list, dictionary codification, stress patterns, and/or PoS tagging for the different multiword units, as relevant.

About identification of more basic meanings:

- A more basic meaning is not necessarily the most frequent (or salient) meaning;
- There may be more than one basic meaning; senses may be equally basic.

- The contextual meaning may not be codified; this may indicate a novel metaphor.
- When you compare contextual and more basic senses, make sure to compare lexical units from the same grammatical class and word category.

About comparing and contrasting basic and contextual senses:

- Finding a contextual and a more basic sense for a lexical unit listed as two separate sense descriptions does not automatically mean that the meanings are related by metaphor: senses may be equally basic, they may be related by metonymy, or one of the senses may just be more specific or general than the other.

About tools for applying MIPVU:

- Before you begin, decide what your long-term plans are for your data.
- When choosing tools for annotating your data, compatibility between programs is an important consideration. How well do different tools 'play' with each other?
- Investing time in developing an effective set of tools to help you in your annotation is a worthwhile investment for future projects.

And finally, we leave you with a "Metaphor Identification Project Checklist" that may help you plan research projects dealing with metaphor:

1. Will you include one language or more?
2. Will you include one domain/register/time period/etc. or more?
3. Do you need whole texts or only (representative) specific parts of the texts?
4. Will you look at all metaphors or only specific types of metaphors?
5. Will you look at all words or only particular word classes?
6. Will you look for metaphors from specific source domains or target domains or mappings?
7. Will you look for flagged metaphors?
8. How many texts will you need to get a sample that is representative of the phenomenon you are interested in?
9. Will you use a quantitative, qualitative or mixed approach?
10. Will you use a top-down or bottom-up approach?

And lastly:

11. Will you apply MIPVU?

Acknowledgements

Thanks to Bård Uri Jensen (Inland Norway University of Applied Sciences) for his input about statistical measures for inter-rater reliability and confidence intervals.

References

Archer, D., Wilson, A., & Rayson, P. (2002). *Introduction to the USAS category system*. http://ucrel.lancs.ac.uk/usas/usas_guide.pdf

Cameron, L. (2003). *Metaphor in educational discourse*. London: Continuum.

Cameron, L., & Deignan, A. (2003): Combining large and small corpora to investigate tuning devices around metaphor in spoken discourse. *Metaphor and Symbol, 18*(3), 149-160. doi:10.1207/S15327868MS1803_02

Canty, A., & Ripley, B. (2015). boot: Bootstrap R (S-Plus) Functions, R package version 1.3-20. https://cran.r-project.org/web/packages/boot/index.html

Deignan, A. (2005). *Metaphor and corpus linguistics*. Amsterdam: John Benjamins.

Deignan, A. (2006). The grammar of linguistic metaphors. In A. Stefanowitsch & S. T. Gries (Eds.), *Corpus-based approaches to metaphor and metonymy* (pp. 106-122). Berlin: Mouton de Gruyter.

Dorst, A. G., & Reijnierse, W. G. (2015). A dictionary gives definitions, not decision. Response 1 to 'On using a dictionary to identify the basic senses of words'. *Metaphor and the Social World, 5*(1), 137-144.

Dorst, A. G., Reijnierse, W. G., & Venhuizen, G. (2013). One small step for MIP towards automated metaphor identification? Formulating general rules to determine basic meanings in large-scale approaches to metaphor. *Metaphor and the Social World, 3*(1), 77-99.

Gamer, M., Lemon, J., & Singh, I. F .P. (2012). irr: Various Coefficients of Inter-rater Reliability and Agreement, R package version 0.84. http://CRAN.R-project.org/package=irr

Howell, D. C. (2010). *Statistical methods for psychology*, Seventh Ed. Belmont, CA: Cengage Wadsworth.

Koller, V. (2002). “A shotgun wedding”: Co-occurrence of war and marriage metaphors in mergers and acquisitions discourse*, Metaphor and Symbol*, *17*(3), 179-203. doi:10.1207/S15327868MS1703_2

Krennmayr, T., & Steen, G. J. (2017). VU Amsterdam Metaphor Corpus. In J. Pustejovsky & N. E. Ide (Eds.), *Handbook of linguistic annotation* (pp. 1053-1072). Berlin: Springer Verlag.

Lew, R. (2013). Identifying, ordering and defining senses. In H. Jackson (Ed.), *The Bloomsbury companion to lexicography* (pp. 284-302). London: Bloomsbury Publishing.

McHugh, M. L. (2012). Inter-rater reliability: The kappa statistic. *Biochem Med, 22*(3), 276-282.

Nacey, S. (2013). *Metaphors in learner English*. Amsterdam: John Benjamins.

Pragglejaz Group (2007). MIP: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol, 22*(1), 1-39.

R Core Team (2013). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/

Reijnierse, W. G., Burgers, C., Krennmayr, T., & Steen, G. J. (2017). DMIP: A method for identifying potentially deliberate metaphor in language use. *Corpus Pragmatics 2*(2), 129-147. doi:10.1007/s41701-017-0026-7

Semino, E., Demjén, Z., Hardie, A., Payne, S., & Rayson, P. (2018). *Metaphor, cancer and the end of life*. New York, NY: Routledge.

Steen, G. J. (2007). *Finding metaphor in grammar and usage*. Amsterdam: John Benjamins.

Steen, G. J. (2009). From linguistic form to conceptual structure in five steps: Analyzing metaphor in poetry. In G. Brône & J. Vandaele (Eds.), *Cognitive poetics: Goals, gains, gaps* (pp. 197-226). Berlin: Mouton du Gruyter.

Steen, G .J., Dorst, A. G., Herrmann, J. B., Kaal, A. A., Krennmayr, T., & Pasma, T. (2010). *A method for linguistic metaphor identification: From MIP to MIPVU*. Amsterdam: John Benjamins.

Stefanowitsch, A. (2006). Words and their metaphors: A corpus-based approach. In A. Stefanowitsch & S. T. Gries (Eds.), *Corpus-based approaches to metaphor and metonymy* (pp. 63-105). Berlin: Mouton de Gruyter.

Van der Meer, G. (1999). Metaphors and dictionaries: The morass of meaning or how to get two ideas for one. *International Journal of Lexicography, 12*(3), 195-208.