University of Naples Federico II

Department of Humanities

PhD course in
Human Mind and Gender Studies

Curriculum in
Languages, Linguistics and
English for Special Purposes
XXIX cohort

Doctoral Dissertation

Unsupervised Recognition of Motion Verbs Metaphoricity in Atypical Political Dialogues

Candidate
Fabrizio Esposito

Supervisor
Prof. Francesco Cutugno

Head of the PhD Board
Prof. Maura Striano

Naples 2017

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

The time is October 10, 2013. The venue the James S. Brady Press Briefing Room, a small theatre in the West Wing of the White House where U.S. Press Briefings usually take place. The topic at hand is internal politics and more precisely the debt-ceiling crisis affecting the United States of America (from here on U.S.).

The White House Press Secretary at the time of Barack H. Obama's first term presidency is Jay Carney. Discussing with press corps of the *extortion* strategies[1] adopted by congressional Republicans threatening "a government shutdown unless their demands were met"[2], a very significant exchange emerges between Mr. Carney and the journalists, as shown below:

> **Press Corps**: You're saying he[3] won't negotiate. You're saying he'll sign a clean extension of the debt ceiling, but he is not going to negotiate on the other stuff until the shutdown is over?

[1] Republican member party Mitchell McConnel first declared the debt-ceiling fight as "a hostage that's worth ransoming. And it focuses the Congress on something that must be done" (from the Washington Monthly webpage at `http://washingtonmonthly.com/2011/08/03/mitch-mcconnell-hostage-taker/`)

[2] `http://www.msnbc.com/rachel-maddow-show/ransom-any-other-name/`

[3] In this case the pronoun 'he' makes reference to Obama.

Press Corps: Yes? Yes?
Mr. Carney: Yes, the President will not pay a ransom for [interrupted]
Press Corps: Yes, that's all. Yes.
Press Corps: It's the "pay a ransom" [interrupted]
Press Corps: You see it as a ransom, but it's a metaphor that doesn't serve our purposes of figuring out what's actually going on.
Press Corps: We just don't want to write it wrong.
Mr. Carney: But you guys are just too literal then, right?
Press Corps: We just want to accurately report what's happening.
Mr. Carney: The closing of the American mind and failure to appreciate metaphor and simile.

The excerpt from the press briefing shows how deep the relationship between politics and metaphors is. Actually, drawing upon the words of the press corps, political rhetoric is so imbued with metaphors that it can become an obstacle for the full comprehension of one's thought. It no coincidence that Thompson (1996) titled his work stating that "[p]olitics without metaphors is like a fish without water".

In the realm of political discourse, metaphors cannot be considered just as another simple trope but they play an important role in the communication strategies employed by the speaker. The question-answer exchange between Mr. Carney and the journalists is a prime example since bringing to light the strategic use of language made by the Press Secretary. In fact, aware of the narrative advanced by congressional Republicans on the issue of the debt-ceiling, he finds the way to use it as leverage. Indeed, employing the strategic metaphor of "pay a ransom", the President and his whole administration immediately join the *good guys* team, acting at the same time as victims but also remarking that they are ones in command. In this way, the demands asked to be met by congressional Republicans become an act of extortion, the *ransom* that Obama's administration is not willing to pay. It can only be considered as fascinating (and at the same time troubling) how metaphors can become powerful tools of manipulation in the hands (or perhaps better to say words) of political figures.

Thus, being able to process and fully comprehend metaphors when it comes to political discourse represents an essential skill not only for us humans but also for machines. Indeed, as the amount of political data keeps on growing day by day due to the endless digital evolution, the development of systems being able to handle and detect metaphors in these large collection of texts becomes of foremost importance.

However, as also highlighted by the press corps, this task is not as trivial as one might be led to think, even for humans. Metaphors can be indeed obscure to one's interpretation, metaphors can be misleading, metaphors sometimes cannot be even perceived at all as such. How can a machine understand that in Mr. Carney's utterance the ransom that the President is not going to pay is not actually a *real and true* ransom?

Let us just be the machine for a few seconds. If we were just looking at the utterance itself, it would be impossible for us to detect the presence of this trope. Indeed, the syntactic structure is incomplete as we do not even know what the President is not paying the ransom for (even if there were a pronoun, we should first resolve its reference). Furthermore, metaphors do not all show the same linguistic pattern. In this case, our reading is either entirely metaphorical or not metaphorical. We begin to understand what the real meaning of Carney's words is only looking at the larger context, namely the preceding questions coming from the press corps. Even so, it is not easy at all to understand the reference of the *absent* prepositional object in the metaphorical expression, not to mention the correct understanding of terms such as *shutdown* in this context. Last but not least, if we machine wanted to interpret the metaphorical expression, we should first know that humans talk of reaching an agreement in terms of criminal negotiation.

Thus, in order to successfully (or at least in most of the cases) detect (and interpret) a metaphor, a machine should be primarily provided with two things: a grammatical and cultural knowledge. In the field of the computational modelling of metaphors, several studies have focused on this aspect, endowing the machine with knowledge coming from hand-crafted resources. However precious and accurate they are, not only do these repositories require effort and time for building but they are often limited to a handful of languages. Thus, in order to overcome this limitation, the development of metaphor processing systems that do not rely on hand-coded knowledge is arising as a very important issue.

1.1 About this thesis

This thesis deals with the problem of the automatic recognition of the novel metaphorical or literal use of lexical items in dialogical naturally-occurring continuous texts without the recourse to hand-coded knowledge. The focus is on the political genre since the building of an English for Special Purposes (henceforth ESP) corpus collecting the White House Press Briefings[4] represents one of the first objectives of the present research. Furthermore, it is important to highlight that metaphors play an important role in political discourse at both a linguistic and cognitive-pragmatic level. The lexical items under investigation are represented by the verbs of motion identified by Levin (1993), due to their role in the communication strategies deployed in public and political discourse. The problem of metaphor recognition is addressed employing unsupervised techniques which theoretical foundations primarily lie in the *Distributional Hypothesis* theory developed by Harris (1954), namely word embeddings and topic models. Systems are tested on continuous data from real-world situations, although specifically constrained by the genre of the corpus used. The metaphor processing models developed in this thesis adopt two different approaches, defined as *local* and *global*. The first one leverages syntactic knowledge for the recognition of metaphoricity. The second one drifts away from the use of the syntactic structure as feature of the system hence only using the information inferred from the discourse context. This method leads to the following research

[4] As the excerpt shown in the previous section.

questions:

RQ1 Can we move towards a *syntax-agnostic* approach when dealing with the automatic recognition of metaphors employing broadly-based distributional semantics techniques such as word embeddings and topic models? More precisely:

- How does the use of these two unsupervised techniques perform on lexical items when no syntactic information is taken into consideration?
- Does the information coming from topic modelling improve the performance of the system when syntactic information is taken into account?

RQ2 On the basis of Dunn (2013b)' claim that the linguistic structure of an utterance influences its metaphorical reading leading to a binary distinction between *saturated*[5] and *unsaturated*[6] utterances:

- Can it be observed any influence of the particular linguistic structure in the final performance of the metaphor recognition system?
- Does the *global* approach help improve the performance of the algorithm in the detection of metaphors in saturated utterances?

RQ3 Avoiding the recourse to any task-specific hand-coded knowledge and labelled data, does the joint use of word embeddings and topic modelling compare favourably with metaphor processing systems based on unsupervised approaches present in literature?

The thesis is structured as follows.

[5]Either an entirely metaphorical or non-metaphorical reading is possible.
[6]Only a metaphorical reading is possible.

Part I provides the background: Chapter 2 discusses the theoretical background on research on metaphors that will be supporting the framework of the computational models developed in this dissertation. The lexical items that will be object of analysis in the present dissertation are also introduced and the choice motivated. Chapter 3 describes the necessary background on the employed unsupervised methodologies and on the research field of the computational modelling of metaphors. Existing approaches employing unsupervised techniques are reviewed.

Part II introduces the data to be used for the task of metaphor recognition. Chapter 4 deals with the description of the corpus developed for this thesis, focusing on its genre characteristics and on the building of the resource. Chapter 5 discusses the lexical units under investigation, namely the verbs of motion, and the criteria of selection for the final task. The guidelines provided to human judges for the annotation of the literalness/metaphoricity of motion verbs are also here presented.

Part III deals with the task of metaphor identification[7]. In Chapter 6 the design and the intuitions at the basis of the developed systems are discussed. The three algorithms proposed for the task are described in detail. Chapter 7 discusses the results of the performance of computational metaphor processing systems hence addressing the three research questions.

Part IV concludes with Chapter 8, formulating the answers to the research questions.

1.2 Publications

Some chapters of this thesis contain material from or are extended version of peer-reviewed publications:

Chapter 3 contains material from Esposito et al. (2016).

[7]This task is also known as *metaphor recognition*. The two expressions are interchangeable and will be used as such in the present dissertation.

Chapter 4 is an extended version of Esposito et al. (2015) and contains material from Cimmino et al. (2016).

PART I

Walking along the Metaphor River

CHAPTER 2

Theoretical background

2.1 Overview of the chapter

In this chapter, the theories that have informed and shaped the path of computational research on metaphor are described and discussed. In Section 2.2, a general introduction to the concept of metaphor is provided. Section 2.3 describes Lakoff and Johnson's Conceptual Metaphor Theory, since representing the line of interpretation and investigation of metaphors in the present work. Section 2.4 discusses the role played by the linguistic structure in modelling the metaphorical meaning of the utterance. In Section 2.5, the cognitive-pragmatic notion of *proximization* is presented due to its importance as a political communication strategy in the realm of public discourse. Section 2.6 concludes the chapter with the description of Levin's classification of verbs of motion since these lexical items represent the target units of analysis in the present thesis.

2.2 Talking of metaphors

Metaphors are all around us. Well, figuratively speaking of course. Metaphors have been around us for centuries. Perhaps, it is more safely to say since the dawn of human race. Some have indeed suggested that cave paintings could be actually considered as the first signs of non-verbal metaphors (Lichtman, 2013, p.356). However, as argued by Lakoff and Johnson (1980, p.180), it is in the Ancient Greek that the power of metaphor begins to be discussed and widely acknowledged. Evidence of this is found in the words of Aristotle, who used to sing its praises in his *Poetics* and *Rhetoric* stating that "[i]t is a great thing, indeed, to make proper use of the poetic forms, ...But the greatest thing by far is to be a master of metaphor (*Poetics* 1459a)" since "ordinary words convey only what we know already; it is from metaphor that we can best get hold of something fresh (*Rhetoric* 1410b)" (Lakoff and Johnson, 1980, p.180).

Metaphors permeate our lives, shape the way we perceive our reality and the way we act on it. They help us convey complex messages reinvigorating the meaning of ordinary words in a quest of clarity, while at the same time adorning and enriching our communication. Metaphors are also powerful devices of communication: they allow us to move between different domains, from a local to a more general perspective, reshaping the message to be subsequently delivered to the wider audience. Metaphors are fascinating tools of reasoning. They are indeed associations between concepts that at first sight may look distant and unrelated. The creative process behind the metaphor is not confined only to its production, in the "corresponding use of strange words [resulting] in a barbarism (Poetics 1458a)" (Barnes, 2014, p.2333). Receiving and interpreting a metaphor is a process that indeed stimulates our reasoning and generates a number of analogical associations which are only limited by our linguistic and cultural knowledge and by the way we interpret the reality we live in.

Thus, it comes as no surprise that metaphors are so entrenched in our daily use of language, in every and each aspect of it. They are found in the newscasts we watch, in the songs we listen to, in the newspaper we read, in the conversations we overhear. They indeed pervade every text genre, be them written or spoken. Their meaning is sometimes so embedded in our personal lexicon that we humans seem barely able to recognise them as such. However, taking a closer look, it is still possible to detect and observe the linguistic creativity push that originally *ignited* them.

Over the years, the *power* of metaphor and their ubiquity across texts have inevitably drawn the attention of scholars from a variety of research fields, ranging from linguistics to philosophy, from politics to cognitive neuroscience. In this regard, the 1980s represented a turning point for the academic studies since being the cradle of landmark works that would deeply impact the future lines of research. In particular, the contribution given by Lakoff and Johnson (1980) forged the path of the Natural Language Processing (henceforth NLP) research on metaphor which extends until today.

Veale et al. (2016, p.3) legitimately point out that the study of this rhetorical trope still "remains a niche area in the computational study of lanaguage". Indeed, being the metaphor a very complex phenomenon in language involving the analysis of many and different linguistic layers – e.g. syntax, semantics, co-reference, inference, just to name a few – NLP community has tended to postpone any work on its investigation, focusing instead on *problems* of "more practical and immediate [linguistic] interest" (Veale et al., 2016, p.3).

However, this *niche* area has been widely investigated during recent years, producing a thriving literature in which not only have different and varied approaches to the study of metaphors been proposed, but significant step towards its computational identification and interpretation have been observed. Furthermore, the greater availability of data, the building of *ad-hoc* resources and the growing computational and linguistic interests are among those factors that have boosted the research on this direction. Indeed, leveraging the state-of-the-art technologies developed in literature and cognitive theories and intuitions pursued by researchers, the computational modelling of metaphors can represent an important integration in many NLP-oriented applications and an advancement for a more comprehensive understanding of this fascinating and complex language phenomenon. Furthermore, innumerable are the benefits that this research could bring to the study of more qualitative aspects. One need only think of those fields focusing on the impact and on the effects triggered by the use of metaphors in real-world communication (e.g. sociology and political science just to name a few).

2.3 Conceptual Metaphor Theory

As previously said in Section 2.2, the main source of influence and inspiration in NLP research on metaphor resides in the extensively investigation of this rhetorical trope carried out by Lakoff and Johnson (1980). In fact, if browsing through the literature of the last three decades, it can be observed that much of the work on the computational modelling of metaphor proposed so far have in common the same theoretical foundation, namely Lakoff and Johnson's landmark work.

In *Metaphors We Live By*, the authors presented a view on metaphor that would definitely change the panorama of the research field for the years to come. Their perspective is known across scholars and laymen alike as Conceptual Metaphor Theory (henceforth CMT).

The keystone of Lakoff and Johnson's work lies in the claim that our conceptual system – namely the way in which we both think and act – is "fundamentally metaphorical in nature" (Lakoff and Johnson, 1980, p.4). According to the authors, metaphors are indeed not just a phenomenon exclusively belonging to the realm of language, "just a matter of words" (Lakoff and Johnson, 1980, p.4), but rather pervasive in our thoughts and in our actions. What this means it is that the metaphorical *power* that we can observe unleashing on the surface of the linguistic evidence resides primarily in our conceptual system. Quoting authors' words, "[t]he concept is metaphorically structured, the activity is metaphorically structured, and, consequently, the language is metaphorically structured" (Lakoff and Johnson, 1980, p.5). Since all starting from the concept, the linguistic realisations of metaphors as we know them "are possible precisely because there are metaphors in a person's conceptual system" (Lakoff and Johnson, 1980, p.6). Thus, language becomes a source of evidence for the definition of this system. The importance of this claim is further reinforced by the role played by these concepts in our daily lives, as our conceptual system structures the way we interact with the world surrounding us, ergo defining (the vision of) the reality we live in.

In order to explain the way the conceptual metaphor takes place and how it inevitably, and often unconsciously, shapes our actions, Lakoff and Johnson discuss the existence of the conceptual metaphor *argument is war* that seems common to many societies nowadays, especially the western ones. Let us look at the following examples drawn from different text genres that support authors' claim:

(i) Mary demolished John's argument with her newly found evidence (from Dunn (2013b, p.38))

(ii) Trump battles CNN reporter in heated exchange at press conference (from the Business Insider Nederland[1])

(iii) How to win an argument (About 69,200,000 results from Google search engine[2])

(iv) Trump can be impulsive. But his war with the press is strategic. (from Vox.com[3])

The point made by Lakoff and Johnson is that not only do we talk about arguments using a *war* terminology, but we actually live our arguments as

[1] https://www.businessinsider.nl/cnn-fake-news-donald-trump-cnn-jim-acosta-question-press-conference-2017-1/?international=true&r=US

[2] https://www.google.it

[3] https://www.vox.com/policy-and-politics/2017/2/24/14730546/trump-press-briefing-fake-news

a *war*. We use *evidence* as weapons for final victory. We see our arguing counterpart as an opponent to be *battled*. Every step is carefully planned with the necessary *strategy* to gain ground. And we seem very interested in understanding and exploring the right procedure to win an argument (or at least, a great part of the websurfers). Furthermore, the four examples can be considered as reflections of real-life situations (only (i) is carefully designed for proving a point but it is not hard to be found in a conversation) proving that this conceptual metaphor "is one that we [still] live by in this culture [...] [structuring] the actions we perform in arguing" (Lakoff and Johnson, 1980, p.4).

What can be observed from these examples is that the metaphor does not lie only in its surface realisation, but rather "in our very concept of argument" (Lakoff and Johnson, 1980, p.5). According to the authors, "the essence of metaphor is understanding and experiencing one kind of thing in terms of another". It is the systematic association between the concept and the way we talk about it that allows us to comprehend one aspect of it in terms of another. Indeed, the metaphorical structuring of concepts is always only partial. Conversely, we would confuse one thing with the other, in this way withdrawing from the use of the metaphor (to say that someone is a *chicken* makes only reference to some aspects that are believed to be proper of the animal known as chicken). Indeed, "[t]he metaphor highlights certain features while suppressing others" (Lakoff and Johnson, 1980, p.141).

Thus, at the light of what said so far, when we can define metaphor as cognitive mechanism that arises when two distant and seemingly unrelated concepts are partially understood in terms of the other. In fact, "[t]he primary function of metaphor is to provide a partial understanding of one kind of experience in terms of another kind of experience" (Lakoff and Johnson, 1980, p.154). This systematic correspondence is explained by Lakoff and Johnson as a mapping between two domains of experience, the *source* and the *target*. More precisely, we infer patterns from the source domain to conceptualise the target domain.

The root of the conceptual metaphor structure described by Lakoff and Johnson can be traced back the *interaction view*, and particularly indebted to the work of Black (1962). The main contribution to the CMT can be probably summed up using the metaphor *man is a wolf* discussed in Black (1954) where the author states "[t]he effect, then, of (metaphorically) calling a man a"wolf" is to evoke the wolf-system of related commonplaces. [...] Any human traits that can without undue strain be talked about in "wolf-language" will be rendered prominent, and any that cannot will be pushed into the background. The wolf-metaphor suppresses some details, emphasises others – in short, organizes our view of man". (Black, 1954, p.288). However, it has been the work of Lakoff and Johnson to have emphasised the cognitive aspect of the production and realisation of the metaphor, influencing and pushing forward the computational work in the field of NLP so far. Let us look at the following examples for a more comprehensive understanding of the phenomenon:

(1) American Political Integrity Is in a State of Collapse (from National Review[4])

(2) Europeans can't think of building a future without the Americans (from Politico[5])

(3) "We build bridges not walls" (from Hillary Clinton's rally on BBC.com News[6])

---

[4] `http://www.nationalreview.com/article/446033/character-politics-us-political-culture-collapsing`

[5] `http://www.politico.eu/article/trump-macron-europeans-cant-think-of-building-a-future-without-the-americans/`

[6] `http://www.bbc.com/news/av/election-us-2016-37905856/hillary-clinton-we-build-bridges-not-walls`

(4) "Terrorist attacks can shake the foundations of our biggest buildings, but they cannot touch the foundation of America. These acts shatter steel, but they cannot dent the steel of American resolve." (from G.W. Bush's speech to the nation from PBS Newshour[7])

The examples above show how we systematically experience the target domain of *politics*: we infer some of the most relevant features of the source domain of *buildings* and we project them onto the target domain. In fact, we draw from the *buildings* lexicon words and expressions "to talk about corresponding concepts in metaphorically defined domain" (Lakoff and Johnson, 1980, p.52), in this case the political one. We feel the fragile political and cultural situation as a building on the verge of collapse. We perceive our political future like a path that needs to be laid down. We worry about the external attacks as menaces threatening to disrupt the foundations of the country-building we belong to.

These systematic associations confirm the existence of the conceptual metaphor *politics is building*, which is reflected in our linguistic expressions. Indeed, quoting Shutova (2015, p.580), "[the] conceptual metaphor manifests itself in language in the form of *linguistic metaphor*, or metaphorical expression". Following Shutova's categorisation, the metaphorical expression can be realised as:

- *lexical* metaphor, namely a single word extension (as in (1), (2));
- *multi-word* metaphorical expression (as in (4) and, more in general, in the case of idioms);
- *extended* metaphor, where the rhetorical trope spans over the discourse under analysis (as in (3) and (4)).

In the present dissertation, when speaking of metaphors, I will refer from now on to their *surface* linguistic manifestations in natural language – i.e. the metaphorical expressions we can bump into text or speech.

[7] `http://www.pbs.org/newshour/updates/terrorism-july-dec01-bush_speech/`

More precisely, the focus of analysis for the investigation and computational identification will be represented by their single word extensions, namely the *lexical* metaphors. The reasons motivating this choice are to be explained in next sections.

2.4 The linguistic structure in metaphors

We humans seem to be able to recognise a metaphor without too much effort when bumping into one. Quoting Dunn (2010, p.54), "[t]here is clearly an intuition possessed by native speakers that some utterances are metaphoric and others non-metaphoric. The vast literature on metaphor is enough to show that this intuition exists". Studies have shown a fair rate of agreement between annotators (Birke and Sarkar, 2006; Shutova et al., 2013; Beigman Klebanov et al., 2014) so far, as it often happens to disagree on the metaphoric *charge* of some linguistic expressions.

As pointed out by Dunn, metaphors are not all equal. Some utterances may in fact be regarded as more metaphoric than others as their degree of metaphoricity varies. "These more metaphoric expressions are more marked, in the sense that they stand out more clearly from "literal" expressions" (Dunn, 2010, p.54). The major claim of (Dunn, 2013b) is that the linguistic structure of an utterance influences and helps in modelling the metaphoric meaning of the utterance itself. More precisely, the linguistic properties of an utterance actually influence its direct and consistent metaphorical interpretation or make the metaphorical reading dependent on the it context of appearance. The author is indeed interested in understanding if metaphors do have a meaning – as exemplified by Black and by the approaches of CMT – or if metaphors behave as exemplified by Davidson (Davidson, 1978), namely they do only have a literal meaning, which can be considered as absurd or false. Nonetheless, however important the debate about the meaning of the metaphor is not only for theoretical issues but also for practical purposes, I will not engage with it discussing this aspect of Dunn's research since beyond the scope of the present work.

The interest of this dissertation lies indeed in the criteria the author presents for categorising the variation of metaphoricity in utterances. What he means for degree of metaphoricity can be explained with the example (5) repeated from the utterance (1) in Section 2.3 and drawn from Dunn (2013b).

(5) Mary demolished John's argument with her newly found evidence.

In (5) we can observe that the utterance can only be identified as metaphoric. Quoting Dunn (2013b, pp.39-40), supposing "natural language utterances have a semantic structure that consists, in part, of case role organization connecting an event with its arguments (Fillmore, 1967), and if we assume that the arguments of an event must meet certain selectional restrictions (Katz and Fodor 1963)", the utterance in (5) "contain mismatched arguments". Indeed, not physical objects such as *arguments* cannot be demolished, moreover using a not physical instrument such as an *evidence*. Ergo, the meaning of the utterance cannot be anything but metaphoric. However, what Dunn wants to point out is that some utterances presents a semantic structure that makes them result more metaphorical than others. Let us consider the following examples:

(6) Mary demolished John's stronghold with her newly found evidence.

(7) Mary demolished John's stronghold with her newly found weapon.

Although there is not a clear dividing line between what can be considered metaphoric and non-metaphoric, the point being made by Dunn is that metaphoricity varies continuously. In (6) the patient role is changed to *stronghold*, increasing its degree of metaphoricity. In (7), also the instrument role is changed to *newly found weapon*, with the utterance being more metaphorical than the others[8]. Having a closer look at (7), it can be observed that it is not possible to state if the utterance is metaphorical or not. From our experience, we may tend towards a metaphorical reading since it is hard to find some common people owning both a stronghold and a weapon. But, what if Mary and John make reference to two monarchs? What if John and Mary are just two kids playing in the playground? Thus, the utterance (7) can only be entirely metaphoric or entirely non-metaphoric, until we do not fully comprehend the content of its surrounding context.

To account for the above-described variation in metaphoricity, Dunn divides metaphoric utterances in two categories based on their internal linguistic structure, defining them as *unsaturated* and *saturated*. "The degree of saturation depends on how much the utterance is filled with metaphorical material" (Dunn, 2013b, p.39):

- **Unsaturated utterances**: contain elements from both target and source domains;
- **Saturated utterances**: contain elements only from the target domain.

Let us examine the utterance (8) drawn from the first sentence of (4) in 2.3:

(8) "Terrorist attacks can shake the foundations of our biggest buildings, but they cannot touch the foundation of America."

As it can be observed, (8) is an unsaturated utterance. Indeed, the sentence contains elements from both the source and target domains. A physical event such as the *terrorist attacks* cannot physically *touch* the *foundation* of a political-established entity. As the case role fillers do not meet

[8]From intuition, I agree with Dunn on certain utterances being more metaphorical than others. However, I am not discussing here the gradient distinction provided by the author.

the selectional restrictions, there is a "mismatch or divergence between the elements of the utterance" (Dunn, 2013b, p.40). Most importantly, unsaturated utterances like (8) have only one reading, the metaphoric one. Indeed, they can be identified as such just observing their linguistic structure without having to analyse the surrounding context.

On the contrary, the metaphoric or non-metaphoric reading of saturated utterances depends on the wider linguistic context. Dunn stresses that this ambiguity does not lie in the lexical items involved in the utterance but rather only on their use in a particular context. The example (9) repeated from Section 2.3 makes the case for us.

(9) "We build bridges not walls"

We recognise the metaphorical meaning of (9) because we know that the person uttering it was Hillary Clinton during one of her rally for the presidential campaign. We know indeed what the term *walls* refers to – i.e. the well-known wall wanted by Trump to mark the border between U.S. and Mexico – hence leading us to infer that both *bridges* and *walls* do not make reference to physical objects (furthermore, she is not talking of any specific object since not using any determiner). Thus, (9) is categorised as a saturated utterance, filled only with elements from the target domain (in this case, *buildings*). However, supposing we do not share this common knowledge, the entirely non-metaphoric or entirely metaphoric meaning of (9) can be ascertained only looking at its surrounding linguistic context. Until then, (9) may have also been uttered by a contractor or an engineer, as far as we know.

On the basis of his corpus study research, Dunn provides the *causes* of saturation for an utterance, as outlined in the following lines:

- **Referential ambiguity**: when it is not possible to resolve the reference of pronouns.

- **Lexical ambiguity**: the lexical item is too *general* too provide the necessary information for the detection of the metaphor.
- **Unspecified arguments**: according to Dunn, this is the most persistent cause of saturation. A missing argument is the cause of the ambiguity.
- **Background knowledge**: background or world knowledge is given for granted making it impossible to detect the metaphorical expression without it.

Furthermore, Dunn adds that both unsaturated and saturated utterances may display a different degree of metaphoricity, hence causing a not stable interpretation of the metaphor-in-use. Nonetheless, he does not provide a clear-cut definition of this gradient measure. Utterances having a high metaphoricity show in some cases "a highly metaphoric cross-domain mapping" (Dunn, 2013b, p.40). The measure of degree lies in the choice of the lexicon used as shown in (6) and (7). It is important to highlight here that both high and low metaphoricity utterances have only a metaphoric meaning[9].

On the basis of the results of his corpus study on 500 verbs from four general domains (*physical*, *mental*, *social*, *abstract*), Dunn (2013b, p.46) claims that a mutual dependence emerges between cognitive mappings and the presence of certain linguistic properties of utterances. His work is cognitive-oriented, providing evidence for an enhancement of the descriptive coverage of metaphors that could make a joint use of different approaches to meaning in language. Nonetheless, what is of interest here is the attention paid to the linguistic structure in the exploration of metaphorical meaning, providing important contributions for both a qualitative and computational perspective. In particular, starting from the claim that not all metaphors are equal, the delineation of the two unsaturated and saturated categories represents in this dissertation an important tool for the

[9] I do not delve into the interpretation aspects of the degree of metaphoricity since beyond the scope of the present work.

qualitative investigation of the metaphors in the corpus and offers significant theoretical cues to be used for the development of the computational system described in Chapter 6.

2.5 The Proximization Theory

It comes as no surprise that metaphor has been drawing the attention of many scholars coming from a variety of research fields. This fascinating phenomenon resulting from the *collaboration* between mind and language represents indeed a major device of communication. It is probably in politics that the metaphor arises not simply as a figure of speech but rather as a powerful rhetorical tool to shape the political mind (Lakoff, 2008) of the listener.

From a more practical point of view, Beer and De Landtsheer (2004, p.24) state that these tropes are used by politicians "as tools of persuasive communication, to bridge gaps and build identification between strangers; to frame issues; to create, maintain, or dissolve political coalitions; to generate votes and win elections". It is no coincidence then that Thompson (1996) provocatively states that "politics without metaphor is like a fish without water". If confirmations were needed, recent studies in linguistic and political science (Musolf, 2000; Lakoff, 2008; Lakoff and Wehling, 2012) have indeed suggested that "the use of a particular metaphor often guides the speakers' argumentation strategy throughout a piece of discourse, as well as participants' behaviour in a dialogue" (Shutova, 2015, p.585). This claim seems to be validated by the *proximization* theory (Cap, 2013), a methodological tool recently introduced in cognitive-pragmatics research for the analysis of strategic regularities in political and public discourses.

Proximization is defined by Piotr Cap as "a discursive strategy of presenting physically and temporally distant events and states of affairs (including "distant" adversarial ideologies) as increasingly and negatively consequential to the speaker and her addressee". Developed as a cognitive-linguistic, pragmatic and critical discourse analytic concept, proximization makes reference to how speakers present, and most importantly, leverage the threat of distant entities as "gradually encroaching upon the speaker-addressee territory (both physical and ideological)" (Cap, 2014, p.17), in order to *legitimise* their decisions and actions, disguising them behind the need of neutralising the upcoming negative impact.

The origin of the term *proximization* must be traced back to the work of Chilton (2004), which was the one to first introduce the verbal forms *proximize* and *proximizing* used to describe the act of "bringing [conceptually] closer" something or somebody (Cap, 2014, p.17). It was Cap (2005) to have first employed the nominal term *proximization* referring to " [an] organized, strategic deployment of cognitive-pragmatic construals of/in (originally, political) discourse" (Cap, 2013, p.5). Proximization was indeed initially investigated in the political and public domain, with a major focus on the interventionist rhetoric. It was later extended to various domains, as it started to become integrating part of theoretical frameworks in several studies (Hart, 2010; Chovanec, 2010; Kopytowska, 2010; Cienki et al., 2010; Dunmire, 2011).

The formalisation of proximization as an integrated theory has been proposed in Cap (2013), primarily motivated by the work of Paul Chilton. The root of the theory can be indeed found in two main aspects of Chilton's research: the first one is represented by the attempt to provide a cognitive-linguistic model of conceptualisation in (political) discourse (Chilton, 2004). Chilton's major claim is that in any discourse people *position* other entities in their *world* in a relation that can be expressed via three axes – i.e. space, time and modality. The second one is the attempt to use the three-dimensional space and vector geometry to investigate lexical and grammatical constructions, for the sake of a cognitive scientific attempt (Chilton, 2005). The proximization theory indeed follows the original concept delineated in Chilton (2004) and Cap (2006) accounting for "the symbolic construal of relation between entities within the Discourse Space (DS)" (Cap, 2013, p.17). The deictic center – represented by the speaker themselves realised using pronouns and names of address – is the origin of the three dimensions. Ergo, the entities of the DS are the *Self* – i.e. the speaker – and the distant entities described by them.

The strategic narrative depicts the distant and peripheral entities referred to as ODCs (outside-deictic-center) as posing a threat since conceptualised as to be crossing the DS in order to invade the IDCs' space (inside-deictic-center). The IDCs are in turn represented not only by the speaker themselves but also by all the people the threat is posed to. The threat itself operates on three aspects according to the Spatial-Temporal-Axiological (STA) analytical model. In fact, proximization acknowledges and leverages the "primacy of spatial cognition" (Cap, 2014, p.18) in language and in discourse construction. The *spatial* aspect is *a forced construal* in which the ODCs are *physically* invading IDCs' territory. The *temporal* aspect is *a forced construal* where there is a "symbolic "compression" of the time axis" (Cap, 2013, p.85), since conflict and threat are described as upcoming and historic. The axiological proximization depicts an ideological clash between the inner values of the IDCs and the ODCs' antagonistic values. All together, these three aspects of proximization "contribute to the continual narrowing of the symbolic distance between the entities/values in the Discourse Space and their negative impact on the speaker and her addressee" (Cap, 2014, p.17). Let us consider the following examples drawn from Cap (2013) for the sake of clarity:

(10) The danger is clear: using chemical, biological or, one day, nuclear weapons, obtained with the help of Iraq, the terrorists could fulfill their stated ambitions and kill thousands or hundreds of thousands of innocent people in our country, or any other. [...] The United States and other nations did nothing to deserve or invite this threat. But we will do everything to defeat it. Instead of drifting along toward tragedy, we will set a course toward safety. Before the day of horror can come, before it is too late to act, this danger will be removed. (G. W. Bush, March 17, 2003)

(11) Now shadowy networks of individuals can bring great chaos and suffering to our shores for less than it costs to purchase a single tank. (G. W. Bush, April 24, 2003)

(12) The stakes in that region could not be higher. If the Middle East remains a place where freedom and democracy do not flourish, it will remain a place of stagnation and anger and violence for export. And as we saw in the ruins of the towers, no distance on the map will protect our lives and way of life. (G. W. Bush, November 19, 2003)

The three examples are all extracted from G.W. Bush's war-rhetoric speeches. In (10), Cap describes a mechanism of spatial proximization in act. Using lexical items and phrases such as *innocent people*, *our country*, *other nations*, *United States*, President Bush seems to establish a new shared geopolitical identity creating in this way an IDC territory that is not only menaced but also likely to be potentially invaded by the ODCs, namely the terrorists. In (11), the key lexical item is *now*, making the utterance an example of temporal proximization. *Now* is suggested by Cap as "the moment to start or decide on starting the pre-emptive action [...] informed by events from the past frame" (Cap, 2013, p.91). (12) shows how ideology can be used as a tool for the legimitisation of actions. The clash is here between the *freedom and democracy* values of the IDCs threatened by the ODCs bleak values and postures, further empowered by the analogy with the past events of 9/11. As it can be observed, the ODCs shifting towards the IDCs territory – be it spatial, temporal or ideological – is often enacted by resorting to metaphors and figurative language (e.g. *drifting along toward tragedy*, *bring great chaos*, *ruins of the towers*), which helps strengthen the message conveyed by the speaker.

The interest raised by this theory in the present dissertation does not lie only in the confirmation that politics is highly *indebted* to metaphors, as previously described. Firstly, this framework has been proved to be applicable to wide areas of the public discourse (cf. Chapter 4 for the nature of the data to be used in this work). Secondly, the contribution provided by this analytical model does not rest on a conceptual level, but also shows the role played by lexico-grammatical items. Indeed, in a study performed on a corpus of US presidential speeches and remarks on the US anti-terrorist policies and actions in Cap (2013), the author categorises the key lexico-grammatical items of the spatial proximization framework based on a keyness frequency threshold. What of interest here is the role played by the verbs (and verb phrases) of motion and directionality construed as markers of movement of ODCs towards the deictic center, and also being used as metaphors in political discourse.

The unit of analysis of metaphorical investigation in the present thesis will be indeed represented by the verbs of motion. In fact, acquiring a deictic status intrinsic in the motion they encode, these verbs may represent the *vehicles* leveraged by the speaker to portray the threat posed by the *foreign* DS entities – i.e. their addressees – on the IDCs territory, be them political adversaries or foreign enemies, be them symbolic or not.

2.6 The lexical units under investigation: the verbs of motion

In 1993 a pioneering work on the classification of over 3,000 English verbs was published by the University of Chicago. To this day, the rich reference study of Levin (1993) is indeed a reference point for scholars and researchers from NLP and linguistics communities alike.

Levin starts from the claim that "there is more to lexical knowledge than knowledge of idiosyncratic word-specific properties [...] easily illustrated with respect to verbs" (Levin, 1993, p.1). Speaker's awareness of their properties seems indeed not to be limited to the *classic* lexical knowledge represented by subcategorisation frames, namely the number and syntactic types of arguments with which the verb (in this case) co-occurs with. Instead, "knowing the meaning of a verb can be a key to knowing its behaviour" (Levin, 1993, p.5). What Levin means is that the syntactic behaviour of the verb can be actually predicted from its meaning, hence expecting that verbs belonging to a particular class also share similar aspects of meaning. In a nutshell, a verb is classified in a particular class based on the correlation that exists between its specific types of syntactic behaviour – i.e. the verb alternation, namely the realisations of its argument structure – and the semantics of the verb itself.

As pointed out by Lenci (2008), verbs classes are indeed identified using a distributional analysis where the context features are the different alternations in which verbs are found. Indeed, according to Levin, the investigation of diathesis alternations is the revealing factor for distinction and grouping of semantic coherent class of verbs. "Thus diathesis alternations can be used to provide a probe into the elements entering into the lexical representation of word meaning [...] bringing out unexpected similarities and differences between verbs." (Levin, 1993, p.14-15). In fact, each class is subsequently investigated to *draw out* the common semantic properties that characterise the verbs belonging to the class itself. As observed by Lenci (2008, p.15), "similarity in distribution is taken to be an overt consequence of some deep semantic property that explains it". By way of example, the author indicates how verbs of motion – a class often cited as large and important – are not homogeneous as it was usually thought, but they could actually be split in more than just one main class. Levin and Hovav (1992) prove indeed that this class could be divided at least into two subclasses, the one of *inherently directed motion* and the one of *manner of motion*. Thus, what Levin wants to stress is that "the important theoretical construct is the notion of meaning component, not the notion of verb class" (Levin, 1993, p.17).

As it will be described and motivated in Chapter 5, in the present thesis the whole category of verbs of motion is initially taken into account for its metaphorical investigation. Thus, based on the work carried out by Levin, the selection of the lexical items to be analysed is realised here by *extracting* the terms belonging to the cluster n°51, indeed describing the verbs falling into the *motion verbs* section. The class is further divided into different subclasses, each one describing a semantic coherent little cluster isolating particular meaning components. The clustering division is represented in Table 2.1:

Class Number	**Verb Class**
51.1	Inherently Directed Motion
51.2	Leave Verbs
51.3.1	Manner of Motion: Roll Verbs
51.3.2	Manner of Motion: Run Verbs
51.4.1	Manner of Motion using a Vehicle: Vehicle Name Verbs
51.4.2	Manner of Motion using a Vehicle: Verbs not associated with Vehicle Name Verbs
51.5	Waltz Verbs
51.6	Chase Verbs
51.7	Accompany Verbs

Table 2.1 | *Levin's classification of verbs of motion semantic (sub-)classes.*

The detailed discussion on the computational selection of verbs of motion and their distribution in the corpus is postponed to Chapter 5.

CHAPTER 3

Unsupervised Methodologies for Metaphor Recognition

3.1 Overview of the chapter

In this chapter, the background of the unsupervised methodologies employed in this dissertation for the task of the metaphor recognition of motion verbs is described. In Section 3.2, distributional semantics and their computational implementation - i.e. the distributional semantic models - are described in detail since representing the theoretical foundation of the unsupervised methodologies later discussed. Section 3.3 and 3.4 illustrate the two unsupervised techniques implemented in the present work for the task of metaphor recognition, respectively represented by word embeddings and topic modelling techniques. In Section 3.5, a panorama on the computational modelling of metaphors is provided. The main characteristics in the development of a computational model for the automatic processing of metaphors are first discussed. Section 3.5.3 deals with the related work in the field of the computational modelling of metaphors, with a particular focus on unsupervised metaphor recognition systems focusing on the investigation of lexical units. Section 3.5.4 concludes the chapter with a description of the strategies performed for the evaluation of metaphor modelling systems.

3.2 Distributional Semantics: an overview

In the last three decades a fundamental question has arisen in the computational linguistics community: how can machine model meaning? Luckily for us (and for machines), in the 20th century scholars belonging to a wide spectrum of research fields – ranging from linguistics to the cognitive psychology – have been asking themselves how we humans process and model meaning. The theoretical frameworks originated from this quest – still far from being considered accomplished – represented the pivotal contribution to the study of meaning from a computational perspective that led to what it is considered today as "the most systematic and extensive application of distributional method" (Bruni et al., 2014, p.1), i.e. the Distributional Semantic Models (henceforth, DSMs). Broadly speaking, it can be said that DSMs represent the meaning of a word by looking at the context in which it appears. This definition has often been used as a gentle introduction to help get a grasp on the theory behind this models of meaning representation. Nonetheless, a more thorough explanation is due to the interested reader.

3.2.1 The theory behind the models: the *Distributional Hypothesis*

The theoretical foundation of any distributional semantic model is to be found in the assumption that a certain degree of semantic similarity between linguistic items can be determined by looking at their linguistic distributions, i.e. the words surrounding each linguistic item. To put it differently, "there is a correlation between distributional similarity and meaning similarity, which allows us to utilize the former in order to estimate the latte" (Sahlgren, 2008, p.34). This assumption is known as *Distributional Hypothesis* (henceforth DH) and it can be more formally defined quoting Lenci (2008, p.3):

> "The degree of semantic similarity between two linguistic expressions A and B is a function of the similarity of the linguistic contexts in which A and B can appear."

Thus, the lexical meaning of a word (or at least part of it) depends on its distributional properties, ergo words that share similar contexts will tend to share also similar meanings.

The rise of the DH is usually connected to the work of Zellig Harris, and more precisely to his distributional methodology which contribution is not only reflected in today DMSs' fortune but most importantly in its immeasurable support to the development of linguistics as a scientific research field. Although the study of meaning has a long tradition that can be dated back to the philosophical work of Wittgenstein (1973) claiming that the meaning of a word actually lies in their use, it has been Harris to have laid the foundations for an analysis of meaning based on "firm methodological bases" (Lenci, 2008, p.4), along the lines of the post-bloomfieldian American structuralism tradition.

Stating that "[...] difference of meaning correlates with difference in distribution" (Harris, 1970, p.786), Harris claims that the meaning of a word resides in its purely linguistic realisation. Indeed, "the linguistic meaning is inherently differential, and not referentia" (Sahlgren, 2008, p.36). Linguistics as a science should deal with the internal structure of the language itself and not be looking at extralinguistic aspects (although accepting that "extralinguistic factors do influence linguistic events" (Sahlgren, 2006, p.23)). Thus, according to Harris' view of the distributional methodology, the *explanans* for semantic similarity is to be found in the similarity in the linguistic distribution. This means that if we observe that two linguistic items share the same linguistic environment, we may deduce that they also have a related meaning.

According to Harris, the distributional approach is the procedure to "typologize the whole of language" (Sahlgren, 2008, p.35). As observed by Lenci (2008) and Sahlgren (2008), even if Harris does not take into account meaning as an explanation for linguistic phenomena (just like Bloomfield did), he views the possible investigation of meaning in its linguistic configuration only possible via a distributional analysis.

A note of caution is in order about the notion of semantic similarity in the DH. The investigation of word meaning on the basis of the distributional analysis brings out paradigmatic similarities between linguistic items occurring in the same contexts. This notion of semantic similarity must be considered as a very broad one as it does not reveal the specific semantic relation between the linguistic items under analysis (synonymy, antonymy, hyperonymy and so forth), but it rather encompasses all of them in the final representation.

As pointed out by Lenci (2008, p.14), under this paradigm "the DH only assumes the existence of a correlation between semantic content and linguistic distributions" in order to "get at a better understanding of the semantic behavior of lexical items". The distributional methodology as a discovery procedure cannot identify the specific semantic relation because it is not what it aims at doing. As stressed by Sahlgren (2006, p.24), "[t]he distributional methodology only discovers differences (or similarities) in meaning [...] If we want to claim that we extract and represent some particular type of semantic relation in the word-space model[1], we need to modify either the distributional hypothesis or the geometric metaphor, or perhaps even both".

Lenci (2008, p.14) defines the quantitative discovery procedure as proposed by Harris as a *weak* DH, where the distributional properties are not indicative of words semantic properties at a cognitive level but rather semantics is taken "as a kind of "latent varibale" which is responsible for the linguistic distributions that we observe". The Italian scholar comes up

[1]The *word-space* model will be discussed in the following sections, although using a different terminology for its definition.

with this definition to distinguish this approach from the *strong* version as shown by Miller and Charles (1991), where the DH is interpreted under a cognitive paradigm since "a word's contextual representation is not itself a linguistic context, but is an abstract cognitive structure that accumulates from encounters with the word in various (linguistic) contexts" (Miller and Charles, 1991, p.5). Ergo, under this assumption the distributional discovery procedure does not dwell only in the realm of linguistics but it also becomes the basis for the cognitive hypothesis, an *explanans* of the formation of meaning in a cognitive perspective.

Under the generativism era, the structural approach to the investigation of meaning was temporarily dismissed by the scientific community. However, the rise of corpus-linguistics from the '80s on, backed by the well-known Firthian quote "You shall know a word from the company it keeps" (Firth, 1957, p.11), and the statistical approach to the study of language represented a proof of the success of the distributional analysis as a pivotal methodology for the investigation of lexical meaning. Indeed, as corpora became available and widespread across the linguistics community, the introduction of corpus-based statistical techniques for the study of words distribution across wide collections of texts marked the success of the empirical methodology applied to the study of word meaning.

3.2.2 Distributional Semantic Models

During the last three decades, distributional semantics has played a key role in the computational linguistics community, pushing forward the research for meaning understanding and representation. The first models were introduced in the early 1990s, when the probabilistic revolution took root and machine learning started to become an established asset across NLP.

DSMs, also known as *word-space* (Schütze, 1993; Sahlgren, 2006), *vector-space* (Turney and Pantel, 2010) or *semantic-space* (Padó and Lapata, 2007) models, represent word meaning in geometrical spaces where the closeness between two words indicates their meaning similarity. More specifically, DMSs can be defined as unsupervised computational methods that "[...] represent salient aspects of lexical meaning" (Lenci, 2010, p.57) as high-dimensional vectors built out of a corpus-based statistical analysis of words co-occurrence across text(s). Every vector stores the information about the co-occurrence of the term under analysis with its context and its dimensions equal the size of the corpus vocabulary. The degree of semantic relatedness between the two linguistic items is measured as the geometric distance between their respective vectors.

In the next paragraph sections, the main features in the building of a DMS are discussed, since representing the foundational structural framework for the unsupervised approaches to be discussed in the next sections. The detailed description of a typical DSM in Section 3.2.2.1 is indebted to the works of Sahlgren (2006) and Evert (2010a).

3.2.2.1 Building and features of a DSM

As the number of DMSs have been thriving in literature so far, the underlying main idea has been shown to reside in the scientific approach to linguistics described by the DH. Even though differing in the choice of a specific parameter (context features, association measures, geometrical distance and so forth) according to the *philosophical* approach and to the task at hand, the building of DSMs shows a common structure behind the implementation of every model. A critical role is obviously played by corpora since not only do they represent the repository of the language in use – ergo influencing the final semantic representation of words – but they are typically the unique input of a DSM (if not considering the indirect supervision in the parameters setting).

3.2.2.1.1 DSM as a matrix. According to Evert (2010a), a DSM can be seen as a scaled and/or transformed co-occurrence matrix where each row represents the distribution of a target term – hence the semantic vector – and each column represents the context/dimension in which the target term[2] appears. More formally, in a $r \times c$ co-occurrence matrix M, the r rows represent the target terms while the c columns are the features or dimensions, which characteristics are up to the decision of the researcher. Let us take as a way of example the following utterance – well-known in the entertainment world[3] today – to illustrate a toy co-occurrence matrix:

(13) A Lannister always pays his debts

In this toy example we define the context as the one word preceding and following the target term. Defining t as the target term, its context are $t-1$ and $t+1$: in this case, it means that the the context of 'Lannister' are respectively 'A' and 'always'. Once tabulated the information for each term, our toy co-occurrence will look like as in Figure 3.1.

Term	**Co-occurrents**					
	A	Lannister	always	pays	his	debts
A	0	1	0	0	0	0
Lannister	1	0	1	0	0	0
always	0	1	0	1	0	0
pays	0	0	1	0	1	0
his	0	0	0	1	0	1
debts	0	0	0	0	1	0

Table 3.1 | *Toy co-occurrence matrix.*

I borrow the term *co-occurrent* from Sahlgren (2006) due to its self-explanatory meaning. For each word co-occurring with the target term, a 1 is assigned to it. The sequence of numbers for each target term is known as *vector*. The vector $\vec{v}$ of the term 'Lannister' is in this case the ordered list (1,0,1,0,0,0). Thus, $\vec{v}$ is defined by its n components or coordinates that place and describe the location of the vector in a n-dimensional space.

[2]The word *term* here encompasses a wide range of linguistic items such as tokens, lemmas, phrases, morphemes and so forth.

34

Hence, a vector can be formally defined as $\vec{v} = (x_1, x_2, ..., x_n)$ where x_n is the nth coordinate of the vector. Sahlgren (2006, p.28)'s defines this vector as *context vector* and highlights the importance of this concept for moving from the distributional data to the representation in the geometrical space.

Going back to the above matrix definition, our M co-occurrence matrix can be then represented more formally as follows:

$$M = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & & \vdots \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{bmatrix}$$

3.2.2.1.2 Linguistic pre-processing. The first step when handling corpus data is to linguistically process the text. Tokenisation is typically considered the minimal requirement but more steps of the NLP pipeline can be performed (POS-tagging, lemmatisation, dependency parsing and so forth) depending on the final aim of the research.

Evert (2010a) points out that the implementation of a NLP pipeline often helps reduce data sparseness, a typical problem in distributional models. Indeed, as shown in our toy matrix, the majority of cells present zero as their entries. Quoting Sahlgren (2006, p.38) making reference to the *Zipf's law* (Zipf, 1949), "[o]nly a tiny amount of the words in language are distributionally promiscuous; the vast majority of words only occur in a very limited number of contexts".

To address this issue, both linguistic and statistical criteria may come to the aid in the development of the DSM. Sahlgren (2006, p.38) indicates that even if POS-tagging filtering removes functional terms and words with little semantic meaning (also acting as a shallow means of disambiguation), the result is "modest at best" as the majority of words belong to open grammatical classes. Evert (2010a) instead suggest that performing lemmatisation on texts often reduce data sparseness.

On the other side, dimensionality reduction via *simple* statistical measures (filtering of high- and low-frequencies terms) may help increase the number of words removed resulting in a more significant dimensionality reduction. However, the *linguistic* drawback of this procedure is that semantic meaningful terms belonging to open classes tend to be discarded during the process. More sophisticated statistical measures as feature scaling can be implemented to *handle* the sparseness issue, as it will be discussed in the next lines.

3.2.2.1.3 Co-occurrence context. As previously discussed, each cell of our toy matrix records the frequency of the target term with the predefined context. In our toy example, I define the context as the one word preceding and following the target term. Typically, the context of co-occurrence it is learnt by skimming through huge corpora and it may vary from one distributional model to another. However, in literature two are the main term-context matrices proposed: the term-context-region and the term-term matrices.

In the first one, rows represent target terms while columns are the context-regions – be them documents, paragraphs, sentences and so on. Hence, each cell entry records the frequency of the target term in each individual context-region. Let be c_n the nth context-region, the vector of the term-context-region matrix can be formally defined as $\vec{v} = (c_1, c_2, ..., c_n)$, resulting often in a very sparse matrix.

The definition of the context-region as wide portion of texts such as documents is a direct legacy of the information retrieval. As pointed out by Sahlgren (2008), one might want to carefully select the context if s/he is going to study syntagmatic or paradigmatic similarities. Indeed, in this case the similarity between vectors constructed in this matrix results in a syntagmatic relation.

In the term-term matrix, the context is regarded as the words surrounding the target term. Hence, rows are the target terms and the columns are the word types w defining the n dimensions of the model. The vector in a term-term matrix is formally defined as $\vec{v} = (w_1, w_2, ..., w_n)$. This matrix favours the emergence of a paradigmatic relation from the resulting vectors similarity.

In choosing the context of co-occurrence on which one will be operating, Evert (2008, pp.11–16) describes three main types of context used in Distributional Semantics: the surface co-occurrence (terms co-occurring in a certain distance according to the most classical Firthian tradition), textual co-occurrence (terms co-occurring in the same textual unit) and syntactic co-occurrence (terms co-occurring in specific syntagmatic relations). Sahlgren (2006) shows that matrices built according to a paradigmatic paradigm provide richer data and a more robust statistical foundations. His claim is also supported by Schütze and Pedersen (1997) stating that term-term matrices provide more significant linguistic results and a statistical basis.

3.2.2.1.4 Feature scaling. When using raw frequency in our matrix, the result is to have skewed data and context words that are not informative about our target terms. To deal with this issue, feature scaling can be performed in order to give more relevance to less frequent but more informative context features. Due to the scope of the present dissertation, I only list here the most common weighting measures employed in literature.

Tf-idf – legacy of the research field of information retrieval – has been used to measure the relevance of a term to a document in a corpus. Evert (2005) provides a wide detailed inventory of statistical association measures differing in the balancing of observed and expected co-occurrence counts: Mutual Information (Church and Hanks, 1990), Local Mutual Information Evert (2005) and Positive Pointwise Mutual Information (Dagan et al., 1993; Niwa and Nitta, 1994) stand out for having being proved to provide better results with both term-context-region and term-term matrices. The use of these measures is fundamental for having more representative data.

3.2.2.1.5 Dimensionality reduction. One of the most common issue in the building of a DSM is indeed handling very large and sparse matrices. Indeed, Evert (2010b) shows that the Web1T5-Easy term-term matrix built out of the Google Web 1T 5-gram database contains one trillion cells of which less than 0.05% presents a nonzero entry. To address this problem, the solution is to apply *dimensionality reduction* (also known as *model compression*) to represent high-dimensional data in a low-dimensional space.

The statistical technique known as Singular Value Decomposition (henceforth SVD) has been widely applied in literature. It consists in the decomposition of the original matrix into several smaller matrices that can be multiplied to reproduce the original one. This approach is typically associated to the well-known work of Landauer and Dumais (1997) where dimensionality reductions is used to uncover latent dimensions.

Lund and Burgess (1996) performs dimensionality reduction by *feature selection*, computing the variances for each row and column and keeping only the elements with high variance. Karlgren and Sahlgren (2001) present the renowned DSM implementation known as Random Indexing (RI) that solves the problem of dimensionality reduction directly from the construction of the co-occurrence matrix. More precisely, RI "removes the need for the huge co-occurrence matrix [...] [by] *incrementally accumulat[ing]* context vectors, which can then, if needed, be assembled into a co-occurrence matrix" Sahlgren (2006, p.42). As Basile et al. (2015, p.39) point out, "[t]he mathematical insight behind the RI is the projection of a high-dimensional space on a lower dimensional one using a random matrix".

3.2.2.1.6 Computing word similarity. The representation of vectors in a geometrical space allows to compute proximity between word in mathematical terms. Among the similarity measures proposed in literature so far, both Euclidean distance and City Block distance (also known as Manhattan distance) are considered as special cases of the general Minkwoski metric. However, as Widdows (2004) and Sahlgren (2006) point out, using these measures "frequent words will end up being *too far* from the other words" (Sahlgren, 2006, p.35). Due also to these reasons, the most frequently employed similarity measures across DMSs is the cosine similarity measure. More precisely, the scalar product of the two vectors is divided by their norms as in the following formula:

$$cos\_sim(\vec{x}, \vec{y}) = \frac{x \cdot y}{|x||y|} \tag{3.1}$$

Thus, if observing the geometrical space, the similarity between the two vectors is interpreted by looking at the size of the angle between them.

3.3 Word Embeddings

Although originally developed as an approach to language modelling from research on neural networks (Bengio et al., 2003), the distributed representations of words today known as *word embeddings* rest on the same foundational linguistic hypothesis that characterises each DSM: "words which are similar in meaning occur in similar contexts" (Rubenstein and Goodenough, 1965, p.627). Indeed, just like any DSM, a *neural* word embedding is the representation of a word as a vector, which can be considered meaningless per se if not applied to a pre-definite task. What is the difference then in using a DSM or a neural model in the building of such a vector?

The most significant and explanatory answer can be found in the words of Bengio et al. (2003, p.1137): "[w]e propose to fight the curse of dimensionality by learning a distributed representation for words". Indeed, unlike DSMs' vector representations which are usually high-dimensional and sparse (if no dimensionality reduction step is applied), word embeddings are real-valued, low-dimensional (typically up to hundreds of dimensions, rarely over a thousand), dense distributed representations (ergo presenting most of its values as non-zeroes), hence much more efficient under a computational perspective. Furthermore, as observed by Baroni et al. (2014, p.238), in this approach "[t]he traditional construction of context vectors is turned on its head". Indeed, while traditional DSMs build their vectors by first collecting and counting the co-occurrences of the target term with the context words in the pre-set *surrounding* window and then applying weighting functions, neural word embeddings models "replac[e] the essentially heuristic stacking of vector transforms in earlier models[5] with a single, well-defined supervised learning step" (Baroni et al., 2014, p.238) by predicting one term from its surrounding neighbouring words (this is why DSMs and word embeddings algorithms respectively are also known as *count* and *predictive models*[6]). Thus, the learning process of the distributed representations of words is typically performed by making probabilistic predictions: the target word w_t is predicted given the preceding one(s) w_p.

In the next section, attention is paid to the specific two-layer neural network employed in this dissertation for the production of these distributed representation of words.

3.3.1 Word2Vec

As shown in Section 3.3, word embeddings have been around in NLP for almost fifteen years so far. Bengio et al. (2003) were among the first ones to experiment and leverage neural networks to learn words distributed representations. The landmark works of Mnih and Hinton (2007) and Collobert and Weston (2008) undoubtedly provided an important boost to the study of predictive models for word representations, with the latter introducing the main structure of the neural network architecture that would be at the basis of future proposed in literature. Nonetheless, it was the work of Mikolov et al. (2013c,a) that marked the explosion of word embeddings as a *viral* phenomenon across the NLP community. Indeed, the release of the

[5]With reference here to the traditional DSMs.

[6]This terminology became widespread among NLP scholars after its use in the renowned paper of Baroni et al. (2014).

Word2Vec toolkit [7] by Mikolov prompted several researchers to investigate the use of word embeddings in different NLP tasks[8].

The huge success is probably to be found in the main differences with the previous approaches presented in literature. Indeed, unlike the neural network architectures which had been proposed previously, Word2Vec focuses explicitly on the generation of word embeddings – hence not only as a by-product – by reducing significantly its computational complexity. In fact, if compared to popular neural network models, the simple model architecture of Word2Vec allows to train high quality vectors even on corpora of large dimensions in very reasonable times.

Word2Vec toolkit is the efficient implementation of two different architectures for the generation of word embeddings: Contextual Bag-Of-Words (henceforth CBOW) and the Skip-gram models. Simplifying the network structure by removing expensive hidden layers and non-linear functions, both the architectures represent computationally-effective predictive models.

The first architecture is defined as CBOW since word distributed representation is continuous and, being based on a bag-of-words model, word order does not count. In the CBOW, the n context vectors around the target word are the input to the model. The sum of the vector representations of the context vectors is used to predict the target word. The symmetrical window to be set is up to the decision of the researcher. Embeddings for the context and target word are learnt separately. The probability of the target term given its context words is computed as a softmax function. The objective function is applied to each word in the corpus and embeddings are updated using gradient-based techniques. Since two different words with similar contexts will tend to have a similar meaning, the network will learn similar word vectors for these two different words. Furthermore, in order to speed up the process of maximising the probabilities for each word in the corpus, Word2Vec provides two alternatives to the standard computa-

[7] `https://code.google.com/archive/p/word2vec/`

[8] As the exponential number of works produced in literature so far proves.

tion via softmax classifier: the hierarchical softmax and the negative sampling. The hierarchical softmax improves training efficiency by decomposing the probability of observing a word into a tree-sequence of probabilities making the complexity logarithmic. Negative sampling instead maximises the probability of estimating the correct word by minimising the expected probability of random words drawn from a noise distribution.

The second architecture implemented in Word2Vec is the skip-gram model. The main difference between the two models lies in the notion of context construction. In fact, it can be said that the skip-gram model turns the CBOW one on its head. Indeed, instead of predicting a word from its context, each surrounding word is predicted from its target word. Hence, what differs between the two architectures is the target variable since the skip-gram follows the same topology as of the CBOW.

Let us take as a way of example the famous Shakespeare's quote in (14) drawn from his Hamlet to illustrate the difference in the prediction between the two models.

(14) "Madness in great ones must not unwatched go"

Using the CBOW approach, low capitalising our Claudius' utterance and simply defining the context as one word before and after the target word to be predicted, our dataset would be then look like as in the following list:

(15) [([<padding>, in], madness), ([madness, great], in), ([in, ones], great), ...]

where, the words in square brackets are the input words – i.e. the n context terms – used to predict the middle target term of the symmetrical window. On the contrary, as the skip-gram model inverts the context and target terms, it uses the current word to predict the n words surrounding it. Thus, using this architecture, our dataset would now look like as in the following list:

(16) [(madness, [<padding>, in]), (in, [madness, great]), (great, [in, ones]), ...]

where, the first word in the round brackets is now the input term used to predict the output context words represented in square brackets.

Using predictive models, Baroni et al. (2014) report an improvement in results in a series of tasks compared to the count-based models. Indeed, in their systematic evaluative comparison of count and predictive models the authors show that the neural network implementation represented by the optimisation of parameters of the Word2Vec CBOW architecture is able to *beat* an optimised distributional semantic model on the tasks of semantic relatedness, analogy and synonym detection while not showing a relevant improvement on selectional preferences and concept categorisation tasks. The results of their investigation even led the authors to push forward the research on the *new* wave of word embeddings stating that "the predict models are so good that [...] there are very good reasons to switch to the new architecture" (Baroni et al., 2014, p.245).

Nonetheless, the attention drawn by word embeddings (and by the Word2Vec model in particular) lies primarily in the syntactic and semantic properties that such word vectors have been shown to exhibit. In fact, as shown in the landmark work of Mikolov et al. (2013c), the neural model automatically encodes linguistic regularities in the vector representations and semantic relations between them emerge by performing simple algebraic operations.

Analogy is probably one of the most interesting and noticeable properties of these embeddings, of the kind *a is to be as c is to* __. As shown by Mikolov himself[9], using the embeddings of *king*, *man* and *woman* and computing simple vector operations it can be observed that the resulting vector is the one of the word *queen*. Figure 3.1 drawn from (Mikolov et al., 2013c) shows how concepts about countries and capitals are organised and

[9]This example is briefly illustrated in the section *Interesting properties of the word vectors* at `https://code.google.com/archive/p/word2vec/`

how they are semantically related using the statistical Principal Component Analysis technique to project the 1000 dimensions of the skip-gram model on a two-dimensional plane.

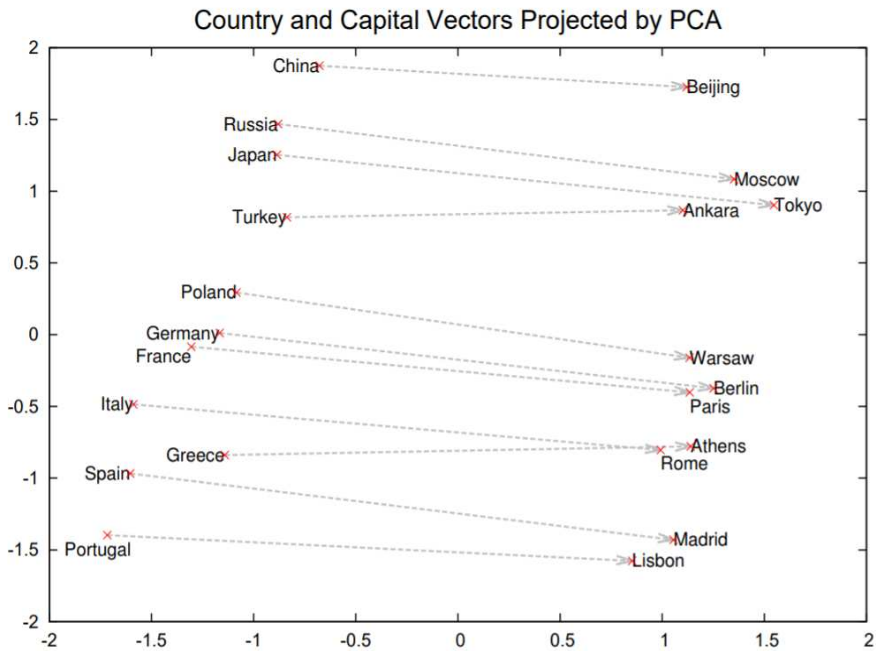

Figure 3.1 | *Two-dimensional PCA projection of the 1000-dimensional Skip-gram vectors of countries and their capital cities from Mikolov et al. (2013c).*

Mikolov et al. (2013c) showed that the analogical reasoning task can also be observed working for phrases. Furthermore, impressive results in semantic relatedness tasks were also proved by adding one vector to another (e.g. the first semantic related word of the element-wise addition of vector(Czech) and vector(currency) returns vector(koruna)) and for named entities (e.g. the first semantic related named entity of *Alan Bean* employing the Hierarchical Softmax is *moonwalker*).

3.3.1.1 The success of word embeddings

Browsing through the scientific literature produced so far, it is not hard to bump into studies leveraging the properties of word embeddings for the enhancement of several NLP tasks.

Systems developed for dealing with Word Sense Disambiguation (from here on WSD) tasks have widely benefited from the popularisation of these low-dense vector representations as several studies have shown (Taghipour and Ng, 2015; Rothe and Schütze, 2015).

Due to the large success, recently Iacobacci et al. (2016) have carried out a study where they propose a comparison of frameworks integrating different kind of embeddings (more precisely the ones generated by Word2Vec, C&W (Collobert and Weston, 2008) and the *retrofitted* approach (Faruqui et al., 2014)), evaluated on standard WSD tasks against multiple benchmarks. Authors show that an accurately designed WSD system taking advantage of word embeddings provides a significant improvement in results if compared to state-of-the-art WSD system equipped with several standard features.

In their study, Passos et al. (2014) showed that using a skip-gram model extended to include information from curated lexicons gathered from several resources not only improves the quality of the learnt phrase embeddings evaluated on three different tasks, but the high-quality word representations also play a key role in significantly boosting the performance of their Named Entity Recognition (from now on NER) system compared to the previous best performing framework using public data (Lin and Wu, 2009).

Unrelated to the enhancement of NLP tasks, Levy and Goldberg (2014) worked on the implementation of word embeddings that could include syntactic information. Instead of taking into account linear contexts, the authors generalise the skip-gram model with negative sampling using replacing the bag-of-words context with dependency-based contexts[10].

[10]The main tweak in the context is the collapsing of the prepositional relations into single arcs.

Compared to *classic* word embeddings, *dependency-based* embeddings produce different kinds of semantic relations similarity, favouring a more *local* context and functional similarity.

The most sound proof of the success of word embeddings and the popularisation of Word2Vec across the NLP community is probably represented by GloVe (Global Vectors) (Pennington et al., 2014). Starting from the claim that the gap between count-based and prediction-based methods is not as deep as it was thought, the authors developed the GloVe model to fully leverage the statistical information of global word co-occurrence counts as the Word2Vec's skip-gram model "poorly utilize the statistics of the corpus" (Pennington et al., 2014, p.1532). Indeed, GloVe makes explicit the encoding of meaning in the vector space performing the training on the *no-sparse* global word-word co-occurrence matrix representing the most valid alternative to Word2Vec's architectures.

Works on the computational modelling of metaphors involving the use of word embeddings techniques are to be discussed in Section 3.5.3.1.

3.4 Topic Modelling

As Bruni et al. (2014) point out, probabilistic topic models have been recently proposed as an alternative implementation of models of distributional semantics. Nonetheless, the main difference between the two models lies in the representation of the meaning provided: like DSMs, topic models are unsupervised algorithms that can process large amount of data; unlike DSMs, probabilistic topic models do not represent meaning in a geometric space model but infer hidden semantic structures in the form of cluster of words using statistical inference.

Topic models are actually a family of algorithms that allow to analyse unlabelled large collections of documents in order to discover the *latent* topics which they consist of. They were initially inspired and motivated by

Latent Semantic Analysis (henceforth LSA)[11] (Landauer and Dumais, 1997) and by its probabilistic alternative known as Probabilistic Latent Semantic Analysis (Hoffman et al., 2010) (henceforth pLSA[12]).

The prominence in the distributional semantic fields of LSA resides in its novel approach to dimensionality reduction of the word-space model (cf. Section 3.2.2). As Sahlgren (2006, p.39) points out, LSA was developed to deal with the synonymy issues that previous word-space models were unable to address in the field of information retrieval. Indeed, as the model learns low-dimensional vector representations by grouping together words co-occurring in similar contexts, it is possible to retrieve documents where synonym words of the *query-word* appear (ergo not needing the *query-word* to be present in the document).

Presented as an alternative to LSA, pLSA is categorised as a probabilistic variant to the original LSA's approach[13]. Indeed, instead of using linear algebra with "a somewhat *ad-hoc* [use]" (Hofmann, 2001, p.178) of SVD, pLSA provides instead a solid statistical foundation based on a generative probabilistic process. The aim of pLSA model is to discover the latent variables – namely the topics/themes/concepts in each document – associated with the observed variables, i.e. the observation of a word occurring in a particular document. Each document is finally represented as mixture of topics where each word is an expression of its probability given the particular topic.

In the next section, the particular topic modelling algorithm implemented in this dissertation for the task of metaphor recognition is described.

[11]LSA is also known as Latent Semantic Indexing (LSI) mainly in the field of information retrieval and the acronyms are often used interchangeably in literature. Here I decided to refer to the LSA model as presented by Landauer and Dumais in their landmark paper.

[12]Also known as *aspect model* and Probabilistic Latent Semantic Indexing (PLSI) mainly in the context of information retrieval.

[13]From which is largely inspired and influenced.

3.4.1 Latent Dirichlet Allocation

Amongst the algorithms proposed in literature, the most influential topic model is probably considered the latent Dirichlet Allocation (henceforth LDA)(Blei et al., 2003). LDA overcomes the main limitations posed by LSA and pLSA as both the models “do not embody generative probabilistic processes” (Blei and Lafferty, 2009, p.1). Instead, the generative probabilistic model adopted in LDA allows to perform better generalisations on new *unseen* documents.

The basic intuition behind the approach is that a document consists of multiple topics. The assumption of LDA is that each word in a document is associated to a single mixture of topics defined as bag-of-words sequence (hence not taking into account the order of their context of appearance). Each word in the topic is in turn associated to the probability of appearing in the identified topic. As Blei (2012, p.79) highlights, what differentiates LDA from other similar approaches is that “all the documents in the collection [i.e. the corpus] share the same set of topics, but each document exhibits those topics in different proportion”.

In a more intuitive view, given a document included in a collection (be it a Wikipedia page, an article in a newspaper and similar) LDA automatically discovers the topics *hidden* in the document and represents them as a cluster of semantically related words. The LDA process can be summed up as shown by Blei (2012, p.12). Topics are assumed to be specified before data have been generated. For each document in our collection, words are generated in two steps:

1. A distribution over the topics is randomly chosen.

2. For each word in the document:

    a) A topic is randomly chosen from the distribution over topics of step 1.

b) A word is randomly chosen from the corresponding distribution over the vocabulary (i.e. the words in our collection).

A note of caution is in order about the qualitative aspect of the discovered topics as these may not correspond to the actual topics present in the document itself. Indeed, in more practical terms, LDA defines the set of topics by analysing all the documents in the corpus and represent each document as a probability list of topics associated to it. It follows that a topic may be *activated* by some words in the document that do not actually *semantically* describe it. In this case, a key role is often played by the length of the document itself as it will be also discussed in Chapter 6.

Thus, LDA can be then defined as a generative probabilistic model that allows to infer latent topics in a collection of documents. The topic structure is the underlying *hidden* variable (Blei, 2012) to be discovered given the *observed* variables, i.e. documents' items from a fixed vocabulary, be them textual or not. LDA model defines indeed a joint probability distribution over both observed and hidden variables using it to compute the conditional distribution of the hidden variables given the observed ones Blei (2012).

3.4.1.1 The success of topic models

The rise of machine learning, the advantage of not requiring any prior labelling data (ergo unsupervised) and their successful application in several linguistic (but not only) tasks (Liu et al., 2016) has drawn the attention of NLP scholars and practitioners alike to topic models, with a special attention to LDA.

Boyd-Graber and Blei (2009) presented and evaluated a *syntactic topic model* which, as the name suggests, captures syntactically and thematically coherent topics. Unlike topic models, in their approach words are not treated as exchangeable units but are constrained by the syntactic structure of the tree.

Ó Séaghdha (2010) applied three topic models related to LDA for the task of selectional preferences induction and evaluated the corresponding outputs on a human plausibility judgements dataset. The author claim that the topic models perform competitively if compared to previous techniques proposed in literature, excelling in particular in the estimation of low-frequent predicate-argument pairs.

Li et al. (2010) dealt with the task of WSD on words and multi-word expressions proposing probabilistic frameworks based on topic models. Inferring the set of topics from a Wikipedia dump consisting of 320,000 articles, using WordNet (Miller, 1995) based sense paraphrases and experimenting different context settings, the three models choose the best sense by comparing the topic distribution of the instance under analysis with the corresponding sense paraphrases. Authors report a state-of-the-art performance for the proposed three models on the three evaluation tasks.

Topic modelling also found application as a means for historical study in (Hall et al., 2008). The authors apply LDA to the Association for Computational Linguistics (ACL) Anthology to explore the trends characterising the field and how they distribute across the main conferences and workshops (Empirical Methods in Natural Language Processing (EMNLP), Conference on Computational Linguistics (COLING) and so on).

The role played by topic modelling, and in particular by LDA in the metaphor recognition of metaphors is going to be discussed in Section 3.5.3.2.

3.5 Computational Modelling of Metaphors

3.5.1 Introduction

The wide range of approaches proposed in literature for the automatic processing of metaphors all fall under the umbrella (multi-)term of *computational modelling of metaphor*. Loosely speaking, when we talk of metaphor modelling we refer to the process of *endowing* the machine with the *knowl-*

edge necessary to automatically perform two tasks (or more often, just one of them): metaphor identification and metaphor interpretation.

Metaphor identification systems are most of the times developed to deal with the discrimination of the metaphorical or literal meaning of the target linguistic unit under analysis. The task is pretty much straightforward: it is typically considered as a binary classification[14] where the system must be able to discriminate the literalness/metaphoricity of the particular lexeme.

Even though much less common, some studies have focused instead on the identification of metaphors on a conceptual level: in this case, the task is to discover the metaphorical mapping between the source and target concepts identifying the linguistic item that ignites such association: e.g. "*light* with hope" reveals the conceptual metaphor FEELING IS FIRE (Shutova and Sun, 2013). In metaphor interpretation tasks the system is developed to detect and explain the particular trope in text. Up to now, different approaches have been proposed in literature, e.g. finding the corresponding conceptual metaphor (Fass, 1991; Martin, 1990) or paraphrasing the metaphorical expressions in a text-to-text form (Shutova, 2010a).

In order to carry out these tasks – no matter which one of these two – the first step in approaching the development of metaphor processing system is the definition of its design. As Shutova (2015, p.581) remarks, "[w]hen designing a metaphor processing system one faces a number of choices. Some stem from the linguistic and cognitive properties of metaphor, others concern the applicability and usefulness of the system in wider NLP context". The design process can be indeed thought as split in two dependent stages, where the first one informs the second one. Its complexity is reflected in the interplay of the multiple factors which it consists of, defining the final realisation of the model.

Thus, the metaphorical level of analysis and the corresponding linguistic aspects are first discussed, since not only defining the scope of interest and investigation, but also determining the computational techniques in-

[14] Be it unsupervised or supervised.

volved in the development of the model. In Section 3.5.3, the literature background on the use of unsupervised techniques for the metaphorical identification of metaphors in text is described in detail.

3.5.2 Level of analysis

When we decide to embark on the project of the computational investigation of metaphors, the first question we have to ask ourselves is what kind of metaphor we want to explore, detect and/or explain. This step is fundamental because it inevitably *constrains* our approach to the task, finally determining the implementation of the computational model itself.

We humans are unquestionably able to detect and understand metaphors when we encounter one, maybe sometimes with some difficulties than in most of the other cases. However, metaphors do not have one single clear-cut realisation in speech/text (especially if we think of them in terms of single linguistic tokens) and there is not one unique level of analysis.

Let us take as a way of example one of the most renowned lines in the realm of television entertainment:

(17) Winter is coming[15]

For those who are acquainted with the television show the phrase is extracted from, they know that this motto goes way beyond its literal meaning since standing for the difficulties and the menaces that lie ahead of us, prompting to a constant state of vigilance and warning. This metaphor displays different level of analysis. At a linguistic level the term *winter* does not stand (only) for the coldest season of the year, ergo being used in a metaphorical way. Thanks to this reasoning, it is possible to infer the hidden meaning of the entire phrase itself (not reflecting upon one single

[15]Drawn from the renowned Home Box Office (HBO) drama television series *Game of Thrones*

term), hence uncovering another level of analysis. In fact, it can be observed that a conceptual metaphor HARD TIMES ARE COMING as THE APPROACHING OF A SEASON seems to emerge. We are indeed led to experience these hard times associating only some relevant harsh aspects of the winter season.

Such a complex reasoning is often realised by humans in a matter of seconds, making it appear as a very simple task. However, this is not the case when it comes to the machines, where each single level of analysis needs to be treated as single process. Thus, in the design of a metaphor processing system, linguistics plays a major role in informing the metaphorical level of analysis on which the machine will operate on. Following the invaluable categorisation provided by Shutova (2015), four main levels of metaphorical analysis can be distinguished:

- **Conceptual metaphor**
- **Linguistic metaphor**
- **Extended metaphor**
- **Metaphorical inference**

Conceptual metaphor. As explained in (17), the conceptual metaphor refers to cognitive mechanism described by Lakoff and Johnson (1980) by which we process metaphorical expression and more precisely, the metaphorical mapping between concepts of the source and the target domain.

One of the most fascinating examples of this kind of metaphor ever encountered is the one described in (Lakoff and Johnson, 1980, p.144): PROBLEMS as a CHEMICAL SUBSTANCE. The authors explain to have found it out by listening to the word of an Iranian student that, unlike their western colleagues, perceived the problems “as things that never disappear utterly and that cannot be solved once and for all”. Indeed, when hearing the phrase “the solution of my problems”, the Iranian student thought about it as "a large volume of liquid, bubbling and smoking, containing all of your problems, either dissolved or in the form of precipitates, with catalysts constantly dissolving some problems (for the time being) and precipitating out others".

Although a system able to deal with conceptual metaphors would also have a clear advantage in the detection and *treatment* of linguistic metaphors, to this day the most challenging factor is represented by the building of such knowledge, with humans tending to disagree on the labelling and assignment of source and target categories (Shutova, 2015, p.584).

Extended metaphor. An extended metaphor is considered as such if its realisation can be observed at discourse level, via the use of a metaphorical lexicon (ergo, the target domain) that consistently draws upon the same source conceptual domain. Hillary Clinton's concession speech is exemplifies this kind of metaphor:

(18) Now, I know we have still not shattered that highest and hardest glass ceiling, but someday someone will – and hopefully sooner than we might think right now.[16]

In her words, her disappointment it is at the same time a call for action and project the achievement as a container which *physical boundaries* must be broken. This example is a proof of the importance of metaphor as a tool of communication and as an argumentative strategy, as also highlighted by several scholars (Musolf, 2000; Lakoff, 2008; Beigman Klebanov and Beigman, 2010; Cap, 2013).

Metaphorical inference. Shutova (2015, p.585) defines the metaphorical inference as "grounded in the source domain and [resulting] in the production of surface structures we observe in language as metaphorical

[16] From The Guardian website at `https://www.theguardian.com/us-news/2016/nov/09/hillary-clinton-concedes-election-donald-trump-speech`

expressions". In broader terms, it represents the process of inferring the *hidden* meaning of the conceptual metaphor in use.

Let us take as a way of example the extended metaphor uttered by Paul Broun - Representative for Georgia's 10th congressional district - stating that:

(19) [...] when someone is overextended and broke, they don't continue paying for expensive automobiles; they sell the expensive automobiles and buy a cheaper one. They don't continue paying for country club dues, they drop out of the country club.[17]

In this case, one needs to infer that Rep. Broun is comparing the financial situation of his country to the average U.S. family that needs to cut all the needless spending, "put this financial house back in order" and stop spending money on something they currently cannot afford.

Nonetheless, a system able to perform such a reasoning is today a challenging route not yet pursued. The *walls* to be overcome are indeed represented by the immense knowledge that this system should be endowed with, not to mention the complex association between different domains.

Linguistic metaphor. The linguistic metaphor represents the surface realisation of the underlying metaphorical cognitive reasoning. For example, let us have a look at the following statement from the American politician and current U.S. Secretary of Energy Rick Perry's:

(20) White House has been occupied by *giants* but from time to time it is sought by the small-minded [...]"[18]

[17]From the U.S. Huffington Post website at `http://www.huffingtonpost.com/2011/07/27/rep-paul-brounhighlights_n_911225.html`

[18]From the ABC News website at `http://abcnews.go.com/Politics/rick-perry-calls-donald-trump-cancer-conservatism/story?id=32622991`

In (20) we easily realise that the *giants* he is talking about are neither mythical creatures or very tall people. He is indeed making reference to the former U.S. Presidents.

Thus, the linguistic metaphor can be considered as the straightforward encounter in either speech or text of the metaphorical expression, both for humans and machines. Indeed, in the computational text-modelling of metaphors systems typically receive as input the metaphorical expression, and their basic requirement is to be able to detect and/or interpret the linguistic metaphor.

This level of metaphorical analysis has been widely investigated by the NLP community, as the works of Gedigian et al. (2006), Turney et al. (2011), Sporleder and Li (2009), Heintz et al. (2013), Shutova et al. (2016) have shown (just to mention a few). Nonetheless, although having in common the same level of analysis, each approach may present different shades of metaphorical investigation. Indeed, as Shutova (2015) points out, when designing a system exploring the metaphorical expression per se, one should take into account three fundamental linguistic aspects in the realisation of the linguistic metaphor.

- **Level of conventionality**. The level of conventionality ideally represents the metaphorical degree of metaphoricity of the linguistic metaphor. By nature, metaphors emerge in language as *novel* but as their use becomes widespread in everyday language, they tend to lose their aspect of novelty, turning into *conventionalised* metaphors (Nunberg, 1987). When this process takes place in language, it may result in a sense extension, augmenting the polysemy of the term itself. Furthermore, as Lakoff (1993, p.245) points out, "[t]he system of conventional conceptual metaphor is mostly unconscious, automatic, and is used with no noticeable effort [...]", bringing them closer to the use of literal terms: e.g. "She can *read* his mind". Different perspectives have emerged in literature about metaphoricity (Gibbs, 1984), with the most recent one represented by Dunn (2010) viewing it as gradient continuum and also proposing a computational scalar measurement (Dunn, 2014) for the degree of metaphorical *charge*.

- **Syntactic constructions**. To address the level of analysis of the linguistic metaphor is important to set in advance the syntactic construction that the system will investigate (e.g. verb-noun, adjective-noun and so forth). Selectional preferences (Wilks, 1975) have been indeed widely investigated, showing that metaphors can be detected as a violation of the semantic constraints posed by the predicate. Thus, syntax is a fundamental feature to be taken into account in the development of the system.

- **Lexical, relation or sentence level**. This linguistic aspect concerns the level of annotation of the metaphor. Three levels are categorised by Shutova (2015): the lexical or word level, where the source domain words are tagged; the relational level, where the source and target terms in a particular grammar relation are tagged; the sentence level, where the sentence is tagged according to the presence of metaphorical terms in it.

3.5.3 Metaphor identification: related work

As it will be discussed in Chapter 6, the metaphor processing systems developed in this dissertation deal with the task of metaphor identification, focusing on the level of analysis of the linguistic metaphor (cf. Section 6.3 for a detailed account of systems' features) and not relying on any lexical resource. Thus, according to the scope of the present work, previous approaches concerned with this specific task in the field of the computational modelling of metaphors are discussed in this section. Furthermore, following the title and the content of the current chapter, particular attention will be paid to works dealing with metaphor identification employing unsupervised techniques and limiting their recourse to lexical resources.

As described in Section 3.5.1, in the task of metaphor recognition the system must be able to discriminate between the literal and metaphorical use of the target unit. Thus, in a text-metaphor processing system, given the textual input, the developed framework typically returns a label for categorising the unit as either metaphorical or literal. The task has been widely investigated in literature, with the majority of the approaches focusing on the linguistic level of analysis.

A few words must be spent about the early methods since playing a relevant influence in the computational research on metaphors. First approaches to the task were characterised by the use of lexical resources and especially influenced by Wilk's selectional preferences (Wilks, 1975). Indeed, since selectional preferences can be seen as the semantic constraints the a word *forces* on the the other terms that are syntactically connected to it (Roberts and Egg, 2014), a selectional violation may indicate the a metaphor is in use (Wilks, 1978; Shutova and Sun, 2013). Renowned is Wilks (1978, p.199)'s example related to the act of drinking, "My car drinks gasoline", where the inanimate object 'car' *erroneously* fills the role of subject in its syntactic construction with the verb 'drink'.

The model that probably best exemplifies this early approach to the computational investigation of metaphors can be considered Fass (1991)'s met*. Indeed, working on selected examples, the system leverages selectional restrictions contained in a specific knowledge base for the detection of metaphors, subsequently drawing on a second knowledge structure for the discrimination of metaphors from anomalies. The main drawback of the system – as indicated by the author himself – is that selectional restrictions tend to detect any kind of anomalies in text, ergo not just the metaphors.

Nonetheless, selectional preferences have kept on playing a fundamental role in the computational investigation of metaphors. Indeed, since representing an "important source of semantic information about the properties of concepts" (Shutova, 2015, p.610), they have lead to the interpretation of metaphor as a violation of semantic norms, as a phenomenon breaking the flow of literal meaning in text (Hovy et al., 2013; Shutova et al., 2016).

3.5.3.1 Unsupervised approaches to metaphor recognition

During recent years, multiple factors have fostered the research on the computational modelling of metaphors towards the use of unsupervised techniques. The rise of machine learning, the implementation of new powerful techniques, the mounting need of trying to *do without* high-quality but at the same time *expensive* lexical resources, and the trend towards language-independent models are amongst the reasons that have led the NLP community towards this new path of research.

Although not having as final aim the recognition of metaphors but instead the discrimination of literal or non-literal usage of verbs, the TroFi (Trope Finder) system developed by Birke and Sarkar (2006) can be considered as one of the early approaches to the treatment of non-literal language employing *nearly* unsupervised techniques. TroFi performs the classification of non-literal language as a word-disambiguation task where literal and not-literal usages are considered as two different senses of the same word. The existing similarity-based WSD KE algorithm (Karov and Edelman, 1998) is adapted to the task of discriminating sentence clustering. TroFi employs seed sets annotated by their literal or non-literal sense. The target set consists of the '88-'89 Wall Street Journal (WSJ) Corpus and is tagged without human supervision (Ratnaparkhi et al. (1996)'s tagger and Bangalore and Joshi (1999)'s Super Tagger are used). The system computes the similarity between the sentences containing the target word and all the collections of seed sentences. The clustering of the sentences in the target set as literal or non-literal is performed according to the seed set to which they are attracted to. 25 verbs were chosen for evaluation, with a total of 1298 corresponding sentences extracted from the WSJ Corpus for literalness annotation. Inter-annotators agreement (in this case the annotators are the authors themselves) is measured at $k = 0.77$. TroFi reaches an average F-score of 53.8% on the hand-annotated sentences.

Not explicitly working on the task of *sheer* metaphor identification, Sporleder and Li (2009) proposed an unsupervised cohesion-based approach to the task of discrimination between literal usage of language and idiomatic expressions. Inspired by the work of Hirst et al. (1998) on malapropisms and grounding their approach on the notion of lexical cohesion and lexical chains – i.e. the sequences of semantically related words representing the lexical cohesion over a portion of text (be it a sentence, a paragraph and so forth) – the intuition of the authors is that idioms *break* the cohesion of the text: if a word in the linguistic expression does not take part in any lexical chain, the expression is then prone to have been used idiomatically. The authors compute semantic relatedness using the distributional approach of the Normalized Google Distance measuring semantic similarity between two words. Two classifiers were developed: a weakly supervised chain-based classifier and a fully unsupervised graph-based classifier where vertices are content words and their corresponding edges represent the semantic relatedness between them. 17 idioms were chosen for the evaluation step, extracting their occurrences from the Gigaword corpus (Graff and Cieri, 2003) with two paragraphs before and after the one including the idiom itself. Annotated as *literal* or *non-literal*, the authors report an inter-annotator agreement of $k = 0.7$. For comparison, the authors implemented an informed baseline classifier where expressions are tagged as literal if the noun is found in the context, as metaphorical otherwise, and a supervised classifier checking word overlap of the target expression with the literal and non-literal instances in the training set, classifying it accordingly. Results of the evaluation are reported in terms of accuracy, precision, recall and F-score. The lexical chain classifier globally optimised using an oracle outperforms the other unsupervised classifiers with a reported F-score of 60.53%. However, the graph-based classifier compares favourably, with an F-score of 59.02%. Furthermore, the authors state the the globally optimised chain classifier is "an upper bound for the lexical chain classifier that would not be obtained in a realistic scenario" Sporleder and Li (2009, p.760). Nonetheless, it is the supervised classifier to outperform the other classifiers, hinting at the importance of the context as a cue of idiomaticity.

The approach of Shutova et al. (2010) was the first one to employ unsupervised techniques for the task of automatic metaphor identification, as claimed by the authors. Their system uses word clustering techniques for detecting verb-subject and verb-object metaphorical constructions in unrestricted text. A small dataset consisting of annotated expressions where verbs are used metaphorically – as in the previous syntactic constructions – exemplifying source-domain mappings is used as *starter* seed set. Spectral clustering using syntactic and semantic features is employed to expand the original seed set clustering nouns, representing the target concepts, and verbs, representing source domains. Their hypothesis is that, by observing and learning the analogies in the seed set, the system is able to capture such regularities clustering together abstract concepts associated with the same source domain while target concepts are brought together by meaning similarity. The scope of the experiments is the whole British National Corpus (BNC) (Burnard, 2007). The system is evaluated against human judgements where an inter-annotator agreement of $k = 0.63$ is measured. A precision of 0.79 is reported. Since no large metaphor-annotated corpus was available, recall score is not provided. Furthermore, authors state that being the system seed-dependent, the recall would be questionable (although anyway stated to have harvested a total number of 4456 metaphors from BNC). The authors evaluate their method also against a WordNet baseline, where the synsets are used as source and target domains, reporting a higher coverage of the system in the retrieving of new metaphors and a significant improvement with respect to the 0.44 precision baseline score.

A fully unsupervised approach for the automatic identification of metaphors in unrestricted text was proposed by Shutova and Sun (2013), discovering both metaphorical associations and metaphorical expressions. Their method is based on a hierarchical clustering model. Indeed, the authors start from a hierarchical graph factorisation clustering of nouns returning a network of clusters with different levels of generality where the weights on the edges represent the associations between the clusters, hence indicative of the metaphorical associations. A dataset including 2000 most frequent nouns in the BNC corpus was used for clustering. Grammatical relations of verb lemmas with direct object, subject and indirect object nouns included in the dataset are used as features for clustering. After having obtained the graph of concepts, metaphorical associations are found using the weight connecting the clusters. After extracting the source-target domain mappings, a list of salient features for the metaphorically connected clusters is generated by ranking them according to the joint probability of a particular feature of occurring both with the input noun and the cluster. Finally, the system searches the BNC for the metaphorical expression describing the target domain using the verbs extracted from the set of salient features. The authors evaluate the performance of their model against human judgements and other two baselines represented by an agglomerative clustering baseline (AGG) and a supervised baseline built upon WordNet (WN). On the task of the identification of metaphorical expressions the system attains a precision of 0.69, beating the AGG and WN baseline but not the human ceiling one set at 0.80. Recall was measured against a gold-standard of manually annotated 63 mappings, resulting in 0.61 for the system developed by the authors. As for the recognition of metaphorical expressions, the hierarchical graph model was evaluated against human annotations of sampled sentences tagged as metaphorical by the system and the baselines, asked to mark the expression that were metaphorical in their opinion. The hierarchical graph reaches a precision of 0.65, outperforming the baselines AGG and WN but not the human ceiling measured at 0.79. Recall was not evaluated due to the lack of a large metaphor-annotated corpus available.

With the growing attention of the NLP community to word embeddings, a trend towards approaches drifting away from lexical resources has been observed. Knowledge built using word embeddings has proved to yield satisfactory results in many NLP applications.

Do Dinh and Gurevych (2016) have recently undertaken this line of research in the field of computational processing of metaphors. The authors rely solely on word embeddings as their knowledge-base, combining them with neural networks for the task of metaphor identification. A multilayer perceptrons (MLP) is chosen as feedforward neural network. The task is treated as a tagging problem at content-token level, hence extending the existing framework for NER of Reimers et al. (2014). The authors use pre-trained 300-dimensional word embeddings created with Word2Vec from the Google News dataset (Mikolov et al., 2013c) and take training and test data from the Vrije Universiteit Amsterdam Metaphor Corpus (henceforth VUAMC) (Steen et al., 2010). Search grid is performed on the validation sets for determining the best setting of the network, tuning it according to the best F-score. Their system is measured on precision, recall and F-score, reporting the corresponding scores, 0.58, 0.52 and 0.55, and beating the pseudo-baseline of labelling each token as metaphorical. The authors also experimented with the incorporation of 10-dimensional POS embeddings and concreteness ratings, improving the overall performance not significantly though.

Shutova et al. (2016) presented the first multi-modal method for metaphor recognition by combining linguistic and visual knowledge. Their approach is based on the cognitive findings claiming that meaning representation can be seen as a combination of multiple factors, "not merely a product of our linguistic exposure, but are also grounded in our perceptual system and sensori-motor experience [(Barsalou, 2008; Louwerse, 2011)]"(Shutova et al., 2016, p.162). Both text and visual knowledge are represented by word embeddings, which are learnt separately and then combined in the multi-modal system. Visual embeddings are learnt in a similar way to the work of Kiela and Bottou (2014), using the deep learning architecture Caffe (Jia et al., 2014) to extract image embeddings from a deep convolutional neural network trained on the ImageNet classification task (Russakovsky et al., 2015). Due to the scope of the present dissertation, I only discuss the linguistic modality of the system since of interest for the aim of research. The authors obtain linguistic representation using Mikolov et al. (2013a)'s skip-gram model and the linguistically pre-processed Wikipedia as corpus. Two different kind of embeddings are used for the experiments: 100-dimensional word-level emebeddings are learnt in a first step, using the standard skip-gram model with negative sampling; in the second step, 100-dimensional phrase embeddings are learnt rerunning the skip-gram model but keeping the same contex vectors of the word-level stage. The phrase embeddings are extracted according to the syntactic structure that are due to be analysed, namely verb-noun and adjective-noun phrases. Several arithmetical operations are proposed for measuring metaphoricity. The intuition is that if the two embeddings show a high degree of similarity, the phrase is then supposed to be literal since belonging to the same domain. On the contrary, the expression is considered metaphorical. For the classification of the single instances, an optimal threshold is determined for the proposed scoring methods maximising classification accuracy on a small annotated development set. The system is evaluated against the datasets of Mohammad et al. (2016) and Tsvetkov et al. (2014) annotated for metaphoricity. Authors report evaluation scores in terms of precision, recall and F-score. The MIXLATE method - combining the best linguistic and visual scoring strategies - outperforms the other scoring methods reporting the F-score of 0.75 for verbs and 0.79 for adjectives. However, the linguistic WORDCOS method, computing the cosine similarity between the $word_1$ and the $word_2$ in the phrase, compares favourably to the best performing method in both tasks.

3.5.3.2 Topic modelling metaphors

Among the new paths pursued by research on the computational modelling of metaphors, the topical structure of text has been gaining increasing attention in the community as a clue for the detection of metaphors.

Strzalkowski et al. (2013, p.69) hypothesised indeed that metaphors can be identified by looking at the words “ typically found outside the topical structure of the text”, similar to approaches grounded on the idea of lexical chains. On the same line, Beigman Klebanov et al. (2009) show that terms describing the topic of discussion are less likely to be used metaphorically. As pointed out by Shutova et al. (2016, p.610), the interpretation of the topical structure as a clue for metaphoricity “is somewhat similar to the idea of semantic norm violation as an indicator of metaphor”, although differing in "two crucial ways": the modelling of the source and target domains and the wide context of analysis.

Taking into account theories suggesting that the wide structure of appearance of the metaphor can provide rich understanding clues (Kittay, 1990), Beigman Klebanov et al. (2009) adopted a more *global* look at the investigation of metaphors (if compared to previous "localistic" approaches) hypothesising that words are less likely to be used metaphorically if they describe a common topic of discussion in a corpus of relevant documents (Beigman Klebanov et al., 2009, p.1). To test their hypothesis, the authors build a large corpus of articles about European Union Institutions from three British newspapers and evaluate the topical composition of the documents annotated for metaphoricity. LDA is applied to the corpus to identify discourse topics. In order to have a better description of the topic itself a parameter k was used to control topic assignment, where k represents the top most likely number of words for the particular topic. If the word falls in the top k, then it is assigned to the topic. Annotation data are extracted from Musolf (2000)'s work on the study of recurrent metaphors in European integration process since, as authors state, the choice of the data predated the construction of the corpus. From Musolff's list of source domains, 4 were chosen along with the 128 corresponding articles plus 23 articles from other source domains (for a total of 151 documents). 9 annotators worked on the data (8 undergraduate annotators and Musolff's original annotation). The k inter-annotator agreement for the source domains of LOVE and VEHICLE was 0.66, while for AUTHORITY and BUILD was respectively 0.39 and 0.43. The fit between the annotated documents and the corpus is measured using topical coverage (as the authors point out a large discrepancy in the length of texts between annotated documents and the corpus' ones is observed since annotated data are not actually a sample of the developed corpus). The authors sampled corpus' texts that were at least 343 words long and compared them to the annotated documents, resulting in $p > 0.37$ for every k. The results of the experiment confirmed authors' hypothesis: for $k = 25$, about 15% of the indexed words in the document are considered topical, capturing only 3% of metaphors; for $k = 400$, 22% of words are metaphorical in about 53% of topical indexed words. Although not consisting in a task of metaphor identification, the work of Beigman Klebanov et al. (2009) represents a first quantitative proof of the importance of the large *picture* for the study of extended metaphors.

Minimising the recourse to rich linguistic resources, Heintz et al. (2013) employed LDA technique for the task of metaphor recognition. As the authors state, their work is inspired by Bethard et al. (2009) where the topics generated by LDA are used as features for the Support Vector Machine (SVM) model to classify the target unit as either literal or metaphorical. The intuition behind their approach is that LDA's output can be used an approximation of conceptual domains. Heintz et al. (2013, p.59) follow this intuition hypothesising that words appearing in sentences containing both source and target domains vocabulary are likely to be used metaphorically, using LDA topics as "proxies for semantic concepts". Unlike Bethard et al. (2009), the authors minimise the supervision in their system using an extended large collection of potential source concepts and a small human-crafted list of seed words. Topics are inferred from the Wikipedia articles in the target language, aligned to concepts using the seed list, mapping the topic to at most one concept. The system selects sentences where words are strongly associated to source and target concepts, filtering sentences presenting a source concept common to the whole document (cf. Beigman Klebanov et al. (2009)) and sentences containing too few words not included in LDA stopwords. Evaluation data were collected from news websites and blogs focusing on the target concept of governance. Two kind of evaluations were performed: in the first one, the top five examples for each conceptual metaphor were selected and judged by two annotators ($k = 0.48$), reporting an average F-score of 0.59 (results are only reported for English due unavailability of Spanish annotators at the time). In the second evaluation for both English and Spanish language, the top 250 linguistic metaphors in the corpus were selected and judged for metaphoricity using Amazon Mechanical Turk. For English, a mean metaphoricity of target instances 0.41 (standard deviation (SD) = 0.33) and of 0.39 (SD = 0.26) for conceptual metaphors is reported. For Spanish, a mean metaphoricity of target instances of 0.33 (SD = 0.23) and of 0.31 (SD = 0.16) is instead returned. Authors comment on the results pointing out that many of the metaphors missed by the system are instances of primary metaphors, suggesting that these are "not well-characterized by word co-occurrence" (Heintz et al., 2013, p.64). Other major issues are to be found in frequent fixed phrases, the non-correspondence between most of the source concepts with LDA topics and the difficulties in the annotation process for judges.

Navarro-Colorado and Tomas (2015) proposed a fully unsupervised approach to the task of metaphor recognition. The main assumption behind their method is that metaphors are linguistic items having an unconventional referent or "[colligating] in an unconventional way (Goatly, 1977)" (Navarro-Colorado and Tomas, 2015, p.92). By leveraging topic models, the authors detect the unconventionality of the word by comparing its set of topics with those of the context. If the sets are not similar, the word is deemed to be used metaphorically. More precisely, if the word shares at least one topic of the context, then its meaning is labelled as literal. The context is considered as the words co-occurring with the target term having as boundaries the sentence containing them. LDA is run on Wikipedia, used as reference corpus, for the generation of the topics. Word-topic relations in Wikipedia are considered as conventional. Given a new target corpus, the system first extracts the sentence. Then, a vector of topics is created using those previously associated to each word in the Wikipedia corpus and the same step is performed for the target term. Finally, the system classifies the word as literal or metaphorical if it shares topics with its co-occurrences terms. The system was evaluated on a corpus of 100 Spanish sentences comprising only two target words (*desierto* and *oasis*), hence divided on two balanced subcorpora. The baseline set is to 50% accuracy since following a majority class approach. Eight experiments were carried out varying at each run the number of topics extracted and the number of representative keywords for each topic. Results are reported in terms of accuracy, precision, recall and F-score. For the first experiment (*desierto*), the best setting is registered with 1000 topics and 20 keywords reporting a 0.72 F-score. In the second experiment (*oasis*), the best F-score is of 0.68 when choosing 2500 topics and 50 keywords. The work indicates that using less keywords, the system performs better.

Due to the aim of the present dissertation, I decided to make a selection of the relevant literature accordingly, hence inevitably discarding some studies. Relevant works leveraging topical structure for the investigation of metaphors as those of Strzalkowski et al. (2013) and Beigman Klebanov et al. (2014) were not mentioned, since either supervised approaches or methods drawing on several rich linguistic resources. Nonetheless, studies focusing on this line of research have been enriching the literature produced so far, with recent works being strongly influenced by these approaches (Jang et al., 2015; Haagsma and Bjerva, 2016).

3.5.4 Evaluation of the metaphor processing system

As we have seen in the previous sections, the lack of a common framework forging the definition of annotation, task and evaluation strategies – among others things – in the field of the computational modelling of metaphors, has generated a fragmented picture. Taking a closer look at the evaluation strategies proposed so far, it is possible to observe how varied in literature the approaches are, due to the lack of shared dataset crucial for the development of the field in the NLP panorama. As Shutova (2015, p.613) points out, "[t]he most desirable type of evaluation is that conducted against an [...] naturally occurring, continuous text [corpus] manually annotated for metaphor [...] open-domain and representative of a range of genres".

Nonetheless, thus far studies having followed this invitation can be count on one hand (cf. Shutova et al. (2010) in 3.5.3.1), with the work of Dunn (2013c) standing out for conducting an evaluation of four systems leveraging different methods for metaphor identification (i.e. Sporleder and Li (2009), Turney et al. (2011), the approach developed by Shutova et al. in the works of Shutova (2010b), Shutova and Teufel (2010), Shutova et al. (2010) and Shutova et al. (2013), and the author's system developed in Dunn (2013a)), comparing their performance on the VUAMC (Steen et al., 2010) by genre and sub-class of metaphor. The study provides important information on the *desirable* metaphorical properties of the system for its successful performance, working on two versions of the same dataset, one pre-processed for NER and the other without this step. The author had previously conducted the same system comparison in (Dunn, 2013a), this time using as evaluation dataset the Contemporary Corpus of American English (from now on CoCA) (Davies, 2009) (although not representing a metaphorical annotated corpus, in this study the author annotates a sample using the labels *metaphorical*, *literal* and *humorous*), showing that different performance results with regard to work of Dunn (2013c) might be explained by the characteristics of the corpus used for evaluation. Indeed, the author states that results show that metaphor discriminating linguistic properties may be not applicable to every genre as there is a significant difference between each type of communication. Furthermore, different kind of metaphors have different linguistic properties.

Notwithstanding the works of Dunn (2013c,a), to this day it is still hard to find a common shared dataset for performance comparison, due also to the diversity of the approaches characterising the research on metaphor. Recently, the VUAMC seems to have emerged as a common standard of comparison in many studies present in literature focusing on the task of metaphor identification (Beigman Klebanov et al., 2014; Haagsma and Bjerva, 2016; Do Dinh and Gurevych, 2016). The VUAMC is indeed today the largest available corpus annotated for linguistic metaphors using the MIPVU procedure (Steen et al., 2010), an extension of the MIP (Pragglejaz Group, 2007) framework for empirical annotation. The corpus is a subset of the BNC's Baby Corpus, from which texts from four registers, consisting of approximately 188,000 lexical units, were randomly extracted. The VUAMC represents a pioneering work in the field of research on metaphors, not only for its coverage and since being freely available, but also for the coding of sub-classes of metaphors.

However, as it has been shown in Section 3.5.3, "the majority of approaches created their own test sets, making the results not directly comparable" (Shutova, 2015, p.613). If this fragmented picture is not favourable for an effective growth of the field of research – making it hard to look for a silver lining – the hope rests in the definition of a common line of research as prompted by the remarkable study of Shutova (2015) but also in future works that may unlock the treasure chest of resources being developed so far.

PART II

Introducing the data

CHAPTER 4

Building and development of an atypical political corpus

4.1 Overview of the chapter

This chapter illustrates the data at the heart of this work and the linguistic resource developed in this dissertation. In Section 4.2, the background about the White House Press Briefings is provided since representing the data on which the metaphor recognition systems will be tested on. The structure of the U.S. Press Briefings is first described and their properties as institutional genre are then discussed. Section 4.3 provides a background about political corpora in both (corpus) linguistics and NLP research fields. Finally, in Section 4.4 the corpus developed for the present dissertation is described in detail. After providing an overview of corpus data, its computational construction and characteristics are discussed.

4.2 Background

4.2.1 Some premises

The resource developed in this work and being used as dataset of analysis is the CompWHoB (Computational White House press Briefings) corpus (Esposito et al., 2015). As the acronym goes, this corpus collects the transcriptions of the White House Press Briefings, namely the daily meetings held by the White House Press Secretary for the national and international news media. The CompWHoB is the computational development of the pre-existing White House Press Briefings Corpus, built by Marco Venuti at the University of Naples Federico II, annotated with XML mark-up following the TEI Guidelines (Burnard and Sperberg-McQueen, 2006) and mainly used as resource in corpus-linguistics studies (Venuti et al., 2012; Spinzi and Venuti, 2013).

4.2.2 The White House Press Briefings

The White House Press Briefings (from now on WHoBs) represent one of the most important and symbolic strongholds in the recent history of U.S. political communication. Reading Perloff (1998, p.98), their success does not come as a surprise as U.S. have been dedicating much more time and resources to the communication process rather than to the actual stages of decision-making (at least, until the second and last presidency of Trump's predecessor, Barack Obama). The importance of the WHoBs has significantly grown along the three last decades leaving an indelible mark also on the spreading of the institutional communication across the web. Indeed, since the first term of Bill Clinton's presidency, WHoBs have been transcribed and uploaded to the web, making them available worldwide to an heterogeneous public.

The pivotal role played by WHoBs in U.S. communication strategies has been underlined by scholars and White House media professionals alike. Quoted in the work of Kumar (2010, p.55), Jim Kennedy – communications director for the White House Counsel's Office during Clinton's presidency, appointed to "focus exclusively on the communications aspects of scandal" – points out that the press briefing is like a *duel*. In fact, since being also televised, Press Secretaries have to be ready to answer questions provocatively posed by the press corps which are in turn "looking for an on-camera response".

Jim Kennedy is not the only one to look at these conferences as more than just a confrontation between two opponents. The military metaphor is indeed recurring across political and linguistic literature, viewing the press briefings as "the battlefield for the press secretary" (Spinzi and Venuti, 2013, p.183) where "[a] civilized, rule-bound war is fought"(Partington, 2003, p.111) (in some way, reinforicing the conceptual metaphor *argument is war* introduced by Lakoff and Johnson (1980)). This permanent state of tension, a soft *cold war* between press secretary and press corps, is stressed by Partington (2003) when he defines the relationship between the two starring characters as *rules of engagement*, drawing the phrase from the work of the BBC political correspondent Nicholas Jones (1996). It is no coincidence that the 46th Vice President of the United States and Congressman, Dick Cheney, argued that "You don't let the press set the agenda. They like to decide what's important and what isn't important. But if you let them do that, they're re going to trash your presidency" (Maltese, 1994, p.2).

Thus, the role played by the press secretary turns out to be critical in addressing and interacting with the media corps, as probably best summarised by the words of Marlin Fitzwater, longest serving press secretary under Reagan and H.W. Bush presidencies, stating that "the press secretary always fights with one arm behind his back, trying to serve two masters" (Nelson, 2000, p.1), namely the U.S. administration and the media.

In each press briefing, everybody seems to win. Undoubtedly, press secretary's main tasks are those to deliver official information about the President's daily schedule, explain administration's policies and decisions, provide commentary on current events and answer the questions coming from the journalists (Kumar, 2010). However, as Spinzi and Venuti (2013) highlight, WHoBs have more than an informative function and carry with them a two-sided gain-effect: on the one hand, the press has the chance to daily *test* White House corps and getting the information they need; on the other one, the White House can spread the *desired* information and at the same time "get an immediate feedback on the success or failure of their policies and communication strategies"(Spinzi and Venuti, M., 2016, p.92). At the end, their relationship can be seen as one of interdependence and cooperation (Venuti et al., 2012, p.67).

The main features of each WHoB and its well-defined structure were described by Partington (2003, pp.34-36). In the following lines, the most important characteristics for the aims of the present dissertation are summarised:

- **Setting**. WHoBs take usually place in the office of the White House Press Secretary, inside the White House itself.
- **Channel**. It is mainly spontaneous speech but sometimes the podium reads a prepared statement, also known as *readout*.
- **Topics**. They may range widely. They range from updates about the President's schedule to culture and education, from internal politics to foreign affairs.
- **Participants**. The podium and the press media corps.

As for the discourse structure of the single briefing, I still follow Partington (2003, pp.34-36)'s outline. Each WHoB can be indeed split into three main blocks:

- **Introduction**. The podium either reads out or summarises a prepared statement on the latest White House business.
- **Question and response**. The podium answers the questions coming from the press corps. This part of the briefing usually makes up the longest part of the event.

- **Closing**. It can be optional since there is not any fixed routine. Sometimes the week's future events are outlined and leave is taken.

4.2.3 White House Press Briefings as an atypical institutional genre

Due to the peculiar features that make them stand out from the wide panorama of political press conferences, WHoBs have drawn the attention of many scholars so far. WHoBs have been regarded in literature as an atypical kind of institutional talk (Spinzi and Venuti, 2013). Quoting Partington (2003, p.30), a press briefing is more pragmatically defined as a "talk between professionals and lay people [...], between two groups of professionals with an audience of lay persons (the TV and Internet audience)". Indeed, the task of the press secretary is not only to address the press corps present in the room, but also "those [people] whom the messages reach in the end" (Bhatia, 2006, p.176) – i.e. the media-users – making the press briefing both a political and media discourse.

Being WHoBs daily conferences, the setting of the briefing can be categorised as rather informal: participants know each other very well indeed, although asymmetry is still in place being the podium the one leading the interaction (Venuti et al., 2012, p.67-68). Due to the non-formal setting of the briefings, the conversation often switches from one social register to another, from informality to formality (and viceversa). In fact, quoted by Kumar (2010, p.235), Mike McCurry – White House Press Secretary for Clinton's administration – stated that "[t]he problem with the format and the problem with the job is that you have to wear different hats at different moments".

Thus, WHoBs can be summarised as an atypical institutional genre where the briefing takes place in a non-formal setting in which the podium – being the one leading the conversation – talks at the same time to the press and to the world. Furthermore, the podium shifts between the use of a *transactional* language, "the optimally efficient transmission of information" (Partington, 2003, p.154), and *interactional* language, "[which] primary goal [is] the establishment and maintenance of social relationships" (Partington, 2003, p.155). In particular, this last peculiar trait of the genre represents a fascinating aspect of investigation under a linguistic perspective. Indeed, not only do the press briefings display the use of language as a tool of communication but also as a device of manipulation, enriching in this way the repertoire of linguistic features displayed by the podium of every discourse.

The present dissertation proceeds along its investigation path taking into account the above-described features of the genre under analysis.

4.3 Political corpora across fields

The search for specialised data represents an endless quest across the NLP community where genre has proved to play a great importance. In the last two decades, academic literature has indeed shown that political corpora in particular have been gaining a growing importance not only in NLP but also across a wide spectrum of research areas. The interest shown by scholars for this specific genre can be probably explained by the paramount role played by the political dimension in every society around the globe and by the ever-increasing attention paid by media to the political speech per se. The interplay between different research fields once far away has also generated a great deal of interest in the building of political resources. This is the case of the PolMine-Project[1], where NLP and the social sciences come together to turn pre-existing politically relevant texts into corpora to be investigated using computational techniques.

Already in 2005 Cousins and McIntosh (2005) advocated the necessity for the social sciences of harnessing and applying the information technologies to the political research in order to augment human intellect (Engelbart, 1962). The authors provide an insightful and detailed outline on

[1] http://polmine.sowi.uni-due.de/polmine/

the integration of computing power in the process of collection and analysis of the data of increasing scope and scale, discussing the potential gains coming from Information Technology (from now on IT) but also the limitations of its tools, with a view to a greater efficiency and transparency.

When it comes to political corpora, it is always hard to draw a straight line that could set apart one research field from the other. Thus, in this section an overview of the political resources built as results of works in wide interdisciplinary fields only marginally involving NLP methods is first provided. Corpora and approaches to the analysis of political discourses developed using computational linguistics[2] techniques are then discussed.

4.3.1 Political Resources

The unstoppable expansion of the web over the last decades has made publicly available online data once restricted only to small domains and/or with a very limited access. The interest on these data and the desire to build properly structured resources have represented a significant boost in the rise of political corpora in literature. Designed and usually constructed in a semi-automatic way and looking at different phenomena of investigation, many political corpora are indeed today directly available for use on web platforms, often results of works in the field of corpus linguistics.

The Honk Kong Baptist University Corpus of Political Speeches[3] is an online database that collects political speeches from around the world and also features parallel corpora of English and Chinese Hong Kong Policy Address. It has represented a source of investigation for many and different studies, also focusing on the investigation of metaphors (Ahrens, 2009).

Bevitori (2007) presented a corpus of discourses of the two Houses of the Parliament of the United Kingdom having as topic the war in Iraq. The resource is annotated with socio-linguistic features for the investigation

[2]Note of terminology: in this work the terms Natural Language Processing and Computational Linguistics and their corresponding acronyms are used interchangeably.

[3]`http://digital.lib.hkbu.edu.hk/corpus/search.php`

of semantic patterns resulting from the study of gender in the collected events.

The Hansard corpus[4] is a semantically tagged resource that contains the British Parliament speeches from 1803 to 2005 and allows to find entire classes of related words based on their context scope.

Finally, a recent work from Merz et al. (2016) has presented an open-access, multilingual and annotated corpus of electoral programs based on the collection of the Manifesto Project (Volkens et al., 2015), which includes the largest hand-annotated text corpus of electoral programs.

4.3.2 Political Corpora in NLP

If the building of large political resources has raised the interest of the scientific community during the last decades, it has only been recently that the NLP community has applied computational linguistics techniques to both the development and the analysis of political corpora.

Although mainly developed to aid research on Statistical Machine Translation, the Europarl Parallel Corpus (Koehn, 2005) can be probably considered one of the most relevant works in this panorama. The Europarl is a diachronic parallel corpus collecting the proceedings of the European Parliament and including eleven languages. It can be consulted and investigated for different aims of research, ranging from the linguistic to the most political ones, using the CQPweb(Hardie, 2012) platform[5] or via the corpus management system Sketch Engine[6] (Kilgarriff et al., 2014).

Barbaresi (2012) developed the German Corpus of Political Speeches crawling the online archive of the German Presidency[7]. The corpus was tokenised and POS-tagged using TreeTagger and it was released in XML and Unicode format.

[4]`http://www.hansard-corpus.org`
[5]`https://cqpweb.lancs.ac.uk`
[6]`https://www.sketchengine.co.uk`
[7]`https://www.bundespraesident.de`

Osenova and Simov (2012) built the Political Speech Corpus of Bulgarian providing a detailed annotation for topics, both at *document* and *sentence* level, and for speakers, finally performing sentiment analysis annotation considering subjective and objective statements.

Guerini et al. (2008, 2013) developed the Corpus of Political tagged speeches (CORPS), a corpus collecting political speeches from the Web and tagged with specific audience reactions – such as *laughter*, *booing*, etc. – used to identify markers of persuasion.

As for the automatic analysis of political communication in text corpora, literature keeps on thriving day by day. Thomas et al. (2006) created a corpus extracting all the available transcripts of U.S. floor debates in the House of Representatives from 2005, together with voting records for all the roll-call votes. The authors used the data to predict the support or the opposition to a piece of legislation, also detecting agreement between utterances in a discussion.

Sim et al. (2013) built a corpus of contemporary political writings from books and magazines manually annotated by a political science domain expert with coarse and fine-grained ideology labels. The authors inferred ideological cues from the annotated corpus and then applied a bayesan Hidden Markov Model (henceforth HMM) to infer the proportion of ideologies U.S. Presidential candidates used in their campaigns in 2008 and 2012.

Prabhakaran et al. (2014) analysed the 2012 Republican presidential primary debates to study the dynamics of interaction in the political debates observing topic shift features and how they relate to the notion of power (Prabhakaran et al., 2013).

Recently, Brigadir et al. (2015) employed Distributional Semantic Models (from now on, DMSs) and a Critical Discourse Analysis' (henceforth CDA) framework for the analysis of opposing ideologies tweets from the the 2014 Scottish Independence Referendum and the 2014 U.S. Midterm Elections.

In the same vein, working on the CompWHoB corpus, Esposito et al.

(2017) proposed the use of the Temporal Random Indexing (Basile et al., 2015) – a specific DSM framework that can take into account word meaning change over time – in conjunction with CDA theories to investigate the politically choices made by the podium during the so-called *crisis communication management* moments (Coombs, 2007).

With regard to the automatic classification and annotation of political texts, working on the corpus of public statements given by Margaret Thatcher (Collins, 1999), Beigman Klebanov et al. (2008) proposed three methods for the automatic annotation of political texts. They found that Latent Dirichlet Allocation as unsupervised word clustering is useful for tracing topics while dictionary-based methods using the statistical tool WMatrix (Rayson, 2003) are more effective for comparative studies. The authors also proposed the lexical cohesion analyser (Beigman Klebanov, 2007) for semantic representation trained on the experimental data of Beigman Klebanov and Shamir (2007).

Purpura and Hillard (2006) focused on sorting process of the Congressional Bill Project proposing a topic spotting classification algorithm based on SVM techniques, with a two-phase hierarchical approach in training to greatly reduce the computational sorting costs. Working on the classification of political emails according to the sending party during the 2004 U.S. presidential election, Purpura et al. (2006) proved the effectiveness of binary differentiation between *Republican* and *Democratic* sources.

4.4 The CompWHoB corpus

4.4.1 Before the CompWHoB: the WHoB Corpus

In the beginning, there was the White House Press Briefings corpus (henceforth WHoB corpus). Developed by Marco Venuti at the University of Naples Federico II, the WHoB corpus (Spinzi and Venuti, 2013) is a specialised corpus covering nearly eighteen years of U.S. Press Briefings – from January 1998 to June 2011 – and five presidencies. The corpus was manually annotated using XML markup and following the TEI Guidelines outlined in Burnard and Sperberg-McQueen (2006). The main information encoded in each XML tag concerns the role and the name of the podium (if applicable), details about the briefing venue and extra-linguistic phenomena present in the transcriptions: e.g. *pause*, *laughter*, etc.. The corpus was indexed and investigated using the Xaira (XML Aware Indexing and Retrievable Architecture) package (Xiao, 2006), an open-source tool developed at the Oxford University used to explore any corpus marked up in XML, since supporting TEI out of the box and any other XML schema.

The WHoB Corpus has been mainly used as resource for studies in the field of corpus linguistics so far. Venuti et al. (2009) used a 4-word key-cluster analysis on the WHoB Corpus to investigate the main features resulting from the evolution of the press briefings during Clinton and G.W. Bush presidencies. Among other results, the study reveals the tendency of Bush first presidency to highlight the role of the president by putting him into the foreground.

In Venuti et al. (2012), the authors investigated the role of the podium in his attitude towards the press in a diachronic perspective. As shown in Section 4.2.3, due to the non-formal setting of the briefing itself, one of the peculiar trait emerging from the genre is the *switching* role of the podium according to the *transactional* or *interactional* use of the language made. Focusing on the phrase *I don't know the answer* uttered by the podium across the corpus, authors show that their analysis confirms the shifting role of the podium as an avoidance strategy in order to *preserve the face*, as stated by Partington (2003). Furthermore, the analysis highlights that the communication strategies adopted by the podium diverge according to the different presidency taken into account.

A phraseological study was carried out by Spinzi and Venuti (2013) to explore the discourse aspects of the press briefings as an institutional genre. Focusing on five 4-word clusters, their analysis on Clinton, Bush and Obama's presidencies claims to demonstrate the correlation between power of persuasion and the shifting role ability of the podium. Furthermore, the authors point out the high frequency of mental verbs – e.g. *think*, *believe*, *understand* – which confirms that the conversation genre is one of most relevant in U.S. press briefings.

4.4.2 The CompWHoB corpus: motivations

As mentioned in Section 4.2.1, the CompWHoB corpus is the computational *upgrade* of the pre-exisiting WHoB corpus. The decision to develop this project was mainly prompted by the necessity to automatise the process of construction and annotation of the corpus. Indeed, semi-automatic annotation comes at a cost: not only is it time-consuming but also prone to inevitably oversight errors that can affect the final result. Furthermore, relying only on specialised tools for the retrieval of the desired information from the Web represented an obvious limitation for a more defined customisation of the corpus. The final objective was to employ NLP and IT techniques to build a resource that could be used as a future reference in a wide spectrum of research fields, ranging from political and social sciences to computational linguistics, just to name a few.

4.4.3 Corpus overview

The CompWHoB corpus is a diachronic corpus collecting the transcripts of the daily U.S. press briefings extracted from the American Presidency Project (henceforth APP) website [8], where the *Press Briefings* section archive can be freely consulted. As stated on the website's homepage, the APP – a non-partisan and non-profit project – "is the leading source of presidential documents on the internet". The APP is a result of the collaboration between J.T. Wooley and G. Peters (Woolley and Peters, 2008) and it is actually hosted at the University of California, Santa Barbara.

The CompWHoB corpus covers a time-span of nearly twenty-five years, ranging from 1993 to 2017, and it is automatically updated month by month. Thus, the corpus includes the three full presidencies of William

[8] `www.presidency.ucsb.edu`

J. Clinton, George W. Bush, Barack H. Obama, plus the early days of the first term of the incumbent U.S. president, Donald J. Trump. The latest press briefing collected goes back to May 31, 2017.

Following the original work of Venuti, data are collected and formatted into a standardised XML encoding, according to the TEI Guidelines. As shown in Table 4.1, at the time of writing the CompWHoB corpus consists of a total of 5,900 briefings (indicated in Table 4.1 as *texts*), comprising 33,124,918 tokens and 191,268 types. Given the dialogical characteristics of the briefing, 583,688 turn-takings[9] were computed. Along the investigated time-span, 1,164 is the number of the individually identified speakers taking on the role of *podium* (identified in Table 4.1 in the *WHos* column).

CompWHoB corpus							
Presidency	**texts**	**tokens**	**tokens-mean**	**types**	**TTR**	**turn-takings**	**WHos**
Clinton_1	1,071	6,828,446	63.75	30,775	11.70	116,626	280
Clinton_2	1,066	4,664,096	43.77	31,302	14.49	102,061	303
Bush_1	774	3,662,691	47.32	25,681	13.41	78,745	85
Bush_2	1049	4,533,101	43.21	28,435	13.35	82,409	155
Obama_1	943	5,318,019	56.39	30,486	13.21	101,017	191
Obama_2	902	7,517,856	83.34	31,379	11.44	92,920	127
Trump_1*	95*	600,709*	63.23*	13,210*	17.04*	9,910*	23*
TOTAL	**5,900**	**33,124,918**		**191,268**		**583,688**	**1,164**

Table 4.1 | *Composition of the CompWHoB corpus at its current stage (May 2017).*

In Table 4.1, first term and second term of each presidency are correspondingly signalled by the use of _1 and _2. TTR stands for Type-Token Ratio, computed using Guiraud's Guiraud (1954) index of lexical richness. The row corresponding to the presidency of Donald J. Trump is marked

[9]With *turn-takings* I refer to the total number of utterances pronounced by the podium and the journalists in each briefing.

with an asterisk as data are still incomplete (only the briefings of his early days of presidency have been collected).

4.4.4 Corpus construction and annotation

The first step in the construction of the CompWHoB corpus is represented by the extraction of the press briefings data from the APP website. Press briefings transcripts come in a loose standardised format. Each conference presents a well-defined structure where every turn-taking between speakers is signalled by the use of capital letters. Two roles can be categorised in the transcripts. The first is the podium – be them the press secretary or any other administration official taking on its role – identified by the use of the speaker's surname preceded by the corresponding honorific or job title. The second is the press corps, always identified by the letter *Q*. Both information about podium and journalists come in capital letter in the transcripts.

The original information contained in the transcripts was retained, including the date of each briefing. After extracting the data, the resulting texts were encoded in an XML format following the TEI Guidelines. Transcripts were mapped to XML files according to a calendar year division and metatextual information was kept in order to enrich the corpus. Thus, the resource is diachronically structured and each year file reproduces the same timeline that can found on the APP's *Press Briefings* archive section.

The CompWHoB corpus shows the following XML structure. A *div1* tag is created to mark the beginning and the end of the single briefing, while its attribute value displays the date of the specific event in a `yyyy-mm-dd` format. Every *div1* includes the dialogical structure of the briefing, where each speaker is identified via a *u* tag. All the press conferences contained in a single year are included and can be retrieved using the *text* tag.

More formally, the CompWHoB corpus can be defined as follows: being C the corpus, Y the set of the years contained in it and B_y the collection of briefings in each year, the dataset can be formalised as:

$$C = \bigcup_{y \in Y} B_y$$

In turn, each collection of briefings B_y can be seen as:

$$B_y = \{b | year(b) = y, y \in Y\}$$

Finally, considering b as the single instance of B, the dialogical structure characterising the corpus C is represented as:

$$b \in C : b =< u_1, u_2, ..., u_n >$$

u is the utterance pronounced either by the podium or the journalist and n is the total length of b. Referring to u as document, u_i can be considered as the i-th document consisting of a number of tokens in a range $\{1, t\}$, where t is the total length of u_i.

In order to provide extralinguistic information about the speakers involved in each briefing, every *u* tag consists of self-explanatory multiple attributes: *role*, *job*, *gender*, *age* and *who*. As in the transcripts the press corps is only recognised via the use of the capital letter *Q*, it was not possible to extract information about them. Thus, every attribute values is filled with the label *journalist* except in the case of *age*, instead filled with *u*, namely *unknown*. With regard to the podium, press secretaries are not the only ones recognised as such. Since many are the White House members involved in the conferences, be them administration officials or personnel related to the White House, the value *podium* is assigned to the attribute *role* for all the speakers interacting with the journalists.

As the original transcripts also include meta-textual information about non verbal event descriptions (e.g. *laughter*, *pause*, *off-the-record*), self-closing tags were created to store this valuable information as shown in Table 4.2.

Tag
<event type=‘‘laughter’’/>
<event type=‘‘applause’’/>
<event type=‘‘off_the_record’’/>

Table 4.2 | *Meta-textual speech events tags.*

4.4.4.1 Automatising the structural annotation

In Esposito et al. (2015), the authors first presented a semi-automatic approach for the structural annotation of the corpus based on the use of regular expressions and manual lookup on Wikipedia. This decision was motivated by the several inconsistencies detected in the transcripts when it comes to the identification of the speakers. Indeed, the use of different honorifics and the sometimes *optional* punctuation were the main cause for the incorrect detection of the speakers. Due to this reasons, rules were devised using regular expressions together with manual checks.

In Cimmino et al. (2016), the authors fully automatised the annotation process leveraging the structured information contained in the publicly available databases DBpedia[10](Lehmann et al., 2015) and Wikidata[11](Vrandečić and Krötzsch, 2014). More precisely, each podium is first identified using the information about the presidency in which it is included, their surname and the corresponding honorific to confirm their gender. Then, the system starts looking for the podium’s first name in the text of the particular briefing. If it fails, the system then looks for the name in the White House official website[12] and finally queries DBpedia and Wikidata to retrieve the necessary information. System’s performance was evaluated computing the error rate in the failed recognition of the podium for each year, according to the following formula:

[10] https://wiki.dbpedia.org
[11] https://www.wikidata.org
[12] https://www.whitehouse.gov

$$Error\ rate = \frac{\sum_i P_{un_i}}{\sum_i P_{all_i}} \tag{4.1}$$

P_{un_i} (*unrecognised podium*) is identified as the ith event where the system was not able to retrieve specific information about the podium from the DBpedia and Wikidata databases, hence constructing the XML tag using only the information about the last name of the speaker contained in the original transcripts of the APP website. P_{all_i} is the total number of the podium XML tags produced by the system in the year under consideration. Table 4.3 describes the performance of the system on each year of the corpus.

As it can be observed, the annotation process produces an error rate of 0.38%, which can be considered a very small percentage in relation to the total number of podiums correctly identified. Furthermore, as the briefing collection stops at the early days of Trump's presidency, a high error rate is computed in 2017 as the number of briefings is still far from the total count.

Thus, not only did the automatisation of the structural annotation process remove the need of time-consuming manual checking but it did also enrich the corpus, making it easily navigable for the user. Indeed, since retrieving essential information from DBpedia and Wikidata, it is now possible to discriminate between the different roles of the podium, making accessible precious extralinguistic data which can better define the panorama under analysis.

4.4.4.2 NLP annotation

As regards the NLP aspect, the CompWHoB corpus was initially linguistically annotated using the Python Natural Language ToolKit (henceforth NLTK) library (Bird et al., 2009). The pipeline consisted of four steps: sentence segmentation, word tokenisation, POS-tagging and lemmatisation. The POS-tagging was performed using the Penn Treebank (Marcus et al.,

Year	Podium All	Podium Unrecognised	Podium Correct	Error Rate
1993	20,349	79	20,270	0.4%
1994	10,628	43	10,585	0.4%
1995	13,518	59	13,459	0.4%
1996	12,755	13	12,742	0.1%
1997	15,947	43	15,904	0.3%
1998	16,482	197	16,285	1.2%
1999	11,926	39	11,887	0.3%
2000	8,189	86	8,103	1.1%
2001	10,523	49	10,474	0.5%
2002	8,160	11	8,149	0.1%
2003	12,260	34	12,226	0,3%
2004	8,513	10	8,503	0.1%
2005	9,946	125	9,821	1.3%
2006	11,833	53	11,780	0.4%
2007	11,912	124	11,788	1%
2008	8,420	19	8,401	0.2%
2009	15,695	10	15,685	0.1%
2010	14,107	7	14,100	0.05%
2011	12,505	10	12,495	0.1%
2012	9,245	6	9,239	0.1%
2013	11,154	10	11,144	0.1%
2014	12,255	25	12,230	0.2%
2015	13,084	12	13,072	0.1%
2016	10,603	28	10,573	0.2%
2017*	5,782*	48*	5,735*	0.8%*
Total	295,791	1,149	294,650	
Total Error Rate 0.38%				

Table 4.3 | *Podium automatic annotation process performance.*

1993) tag-set trained on the Treebank Corpus. The system was evaluated via confusion matrix with a human-labelled gold standard test set consisting of 24 sections randomly selected from the corpus (for a total amount of 595 tokens), achieving 92% accuracy. The highest error rate was observed in the erroneous POS-tag of interjections, as they tend to have a higher occurrence in the corpus due to the informal register adopted by the participants to the interaction.

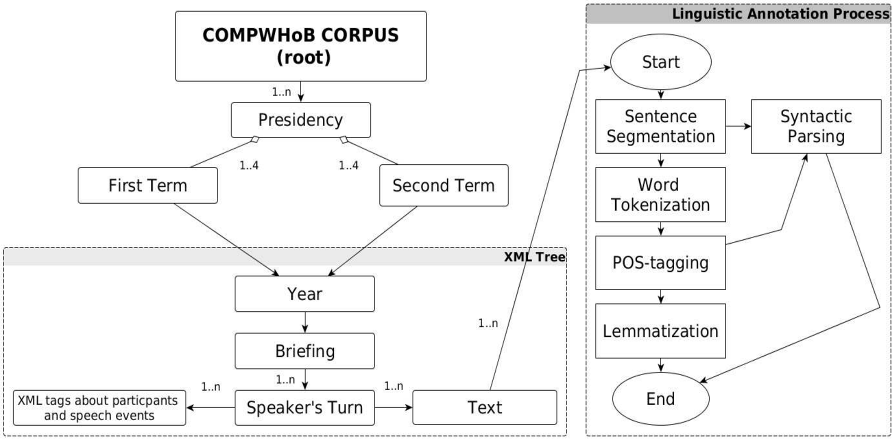

Figure 4.1 | *CompWHoB structural and linguistic annotation process.*

In order to improve system's performance, the NLP pipeline was re-implemented using SpaCy[13], an industrial-strength natural language processing library written in Python. The original pipeline was indeed further extended to include also a dependency parser. The choice behind the use of SpaCy was motivated not only by the wide range of functionalities that SpaCy freely offers, but especially by its fast syntactic parser performance with a high rate of accuracy if compared to the current state-of-the-art parsing technologies (Choi et al., 2015; Honnibal and Johnson, 2015). SpaCy parser is indeed based on MaltParser (Nivre et al., 2006) which is in turn trained on the UPenn Treebank. Furthermore, even if provided with the OntoNotes5 release of the Penn Tree Bank tag-set[14], SpaCy POS-tagger al-

[13] `https://spacy.io`
[14] `https://catalog.ldc.upenn.edu/LDC2013T19`

lows the use of the more general Google Universal tag-set[15] that makes easier the retrieval of wide grammatical categories for the purposes of the present research.

SpaCy POS-tagger was evaluated via confusion matrix on the same gold standard set on which the NLTK POS-tagger was tested. SpaCy POS-tagger reached the 94% accuracy, improving the performance of the system, in particular being able to discriminate the use of interjections.

In Figure 4.1 the structural annotation described in Section 4.4.3 and 4.4.4.1 and the implemented NLP pipeline are represented. The XML tree of the CompWHoB corpus is also described.

[15]Although SpaCy POS-tag list is further extended with more tags.

CHAPTER 5

Customising the corpus for the task

5.1 Overview of the chapter

In this chapter the lexical units under analysis and their relevance for the purposes of the present thesis are discussed. Section 5.2 presents the investigated linguistic units, namely the verbs of motion. After describing the motivations that led to the choice of this Levin's verb class, the criteria for the selection of the motion verbs to be investigated are discussed and their distribution in the CompWHoB corpus is described. Statistical measures are also provided to prove the representativeness of the selected verbs. In Section 5.3, the procedure for the annotation of the verbs of motion for the final task of metaphor recognition is illustrated. Section 5.3.1 explores the metaphoricity of the target units discussing the results of the first annotation task. In Section 5.3.2 the procedure for the annotation of the test set to be used for the evaluation of the developed systems is explained and described in detail.

5.2 The lexical units under analysis: the verbs of motion

As previously described in Section 2.6, in the present thesis the linguistic units under investigation for metaphoricity are represented by the terms belonging to the class n°51 of *verbs of motion* type delineated by Levin in her pioneering work "English Verb Classes and Alternations: A Preliminary Investigation"(Levin, 1993).

As shown in Section 2.5, the primary reason leading to the choice of this category stems from the relevance of the these lexical items in the communication strategies deployed by U.S. administration in public discourse. However, this does not represent the only motivation behind this choice. Looking at a more linguistic aspect, it needs to be stressed the role played by the verbs of motion as metaphorical *carriers*, as several studies have shown so far.

Investigating metaphors in educational discourse, Cameron (2003) carried out a corpus study on all parts of speech, finding out that about 50% of metaphors detected in text were represented by verbs.

Proposing an annotating scheme for the annotation of conceptual mappings, Shutova and Teufel (2010) focused on the detection of metaphorical or literal use of verbs on a sub-corpus of the British National Corpus (Burnard, 2007) divided by genre. Two interesting aspects emerge from their study: 164 of the total 241 metaphors identified are indeed verbs and the most frequent source domain is the one of MOTION. Furthermore – with a particular view to the present dissertation – authors point out that the source domain of MOTION is mainly associated to the target concepts of CHANGE, PROGRESS, CAREER and SUCCESS, showing how these metaphors shape our cognitive mappings.

Gedigian et al. (2006) also proved the pervasiveness of verb metaphors. Indeed, investigating verbal units from a subset of the Wall Street Journal (WSJ) corpus, the authors report that the 93% of the target frames of spatial motion, manipulation and health are annotated as metaphorical.

Finally, as previously discussed in Section 2.5, Cap (2014) indicates the crucial role played by verbs (phrases) of motion and directionality as markers of spatial movement – be them physical or not – in the cognitive-pragmatic paradigm of *proximization*, where these linguistic units are shown to form a significant part of the discursive strategies applied by the speakers in the realm of political discourse.

Thus, interesting aspects emerge from the above-described brief panorama: not only do verbs tend to be used most of the times in a metaphorical way (even though not very surprising as genres seem to play a big part in this picture) but the component of motion is relevant in building metaphorical associations that forge our vision of reality. Furthermore, verb phrases of motion stand out in the political context as tools in the hands of the speaker to manipulate bystanders' attention and perception of the unfolding situation around them.

Due to the atypical political nature of the corpus developed as part of the present thesis, the role played by motion verbs undoubtedly becomes of great interest from a research point of view both under a quantitative and a qualitative analysis of their interplay in this specific genre.

5.2.1 Motion verbs identification

Under a computational-linguistic perspective, in the present dissertation motion verbs are considered as all the instances of word-forms POS-tagged as verbs – hence any verb forms – encountered in the CompWHoB corpus and which lemmas can be found in Levin's class n°51 of verbs of motion. Thus, given the following toy sentence:

(21) The Russian military **is advancing** their borders posts in the Georgian territory.

the verb *advance* in his present continuous form is selected since belonging to the class 51.1 of the *inherently directed motion* verbs.

The motion verbs to be selected are contained in the *Verb Index*, a file including the alphabetical listings of the verbs referred to in Part I and II of "English Verb Classes and Alternations: A Preliminary Investigation" and made available by Beth Levin[1]. The *Verb Index* is organised in a dictionary data-format fashion where each key represents the verb and the corresponding value is its semantic classes. In the present work, verbs are extracted from the reverse-engineered file developed by John M. Lawler[2] where for each verb class a list of the verbs belonging to it is provided. The full list of the *verbs of motion* identified by Levin can be found in Appendix B, detailed according to the section they belong to as reported in her *Verb Index*.

Levin divides the verbs of motion in nine classes semantically defined. As it can be observed, some of the verbs may be found in more than just one class (e.g. drift, climb), making in some cases the semantic distinction between classes a little bit fuzzy. However, each class is considered as a semantically coherent set of verbs which share a similar syntactic linguistic behaviour.

Flicking through the Verb Index of the class n°51, the sub-class of *run* verbs belonging to the more extended *manner of motion* class, stands out as the most numerous one, indeed consisting of 124 units. The *manner of motion* class as a whole collects a total of 142 verbs. On the contrary, the class of *leave* verbs identifies only three linguistic items. Particular interest raises the class of *waltz* verbs, describing a very specific lexicon that however is likely to be encountered only in specialised contexts and genres.

Nonetheless, the difference between classes (and sub-classes) in the verbs of motion does not play a relevant role for the aims of the present dissertation. The nine classes are indeed categorised and treated as a whole set describing the verbs of motion. The candidate verbs of investigation drawn from the class n°51 and the selection of procedure are to be discussed in the next sections.

[1] Actually, it is John M. Lawler to make it available at his University of Michigan webpage: `http://www-personal.umich.edu/~jlawler/levin.html`

[2] `http://www-personal.umich.edu/~jlawler/levin.verbs`

5.2.2 Verbs of motion classes in the corpus

Identification of motion verbs was performed parsing the CompWHoB corpus and looking for tokens POS-tagged as verbs[3] and included in Levin's verbs of motion classes. This process returned 149 verbs out of the total of 247 included in the verbs of motion classes. Table 5.1 illustrates the list of the motion verbs found in the CompWHoB corpus and their corresponding verb type classes.

Although not crucial to the aims of the present dissertation, it is of interest that each of the n°51 verb classes is represented in the corpus. As expected, the *run* verb class stands out for its cardinality also in the CompWHoB corpus, featuring 77 verbs out of the 124 identified by Levin. Unexpectedly, also the *waltz* class gets represented in the corpus, even though including only the 35% of verbs of the *original* class. In Figure 5.1[4], the horizontal bar chart illustrates the cardinality of verbs of motion sub-classes.

However, as it can be observed in Figure 5.2[5], despite the lower number of verbs if compared to the *run* sub-class, it is the class of *inherently directed motion* to cover the lion's share in terms of frequency in the corpus. Indeed, this class includes some of the most recurring verbs in the CompWHoB corpus, as nine of them occur more than 1000 times. In this case, it must be highlighted that this class also features high-occurrence terms such as *go* and *come* which are ubiquitous in many verbal syntactic constructions in language (e.g. phrasal verbs, auxiliary roles in expressions of futurity and so forth).

[3]More details about the parsing process and POS tag-set used are to be discussed in Chapter 6.

[4]It is possible to interactively navigate the graph at the following link: **https://plot.ly/~fabrex/677/cardinality-of-verbs-of-motion-sub-classes/**. This option is made available for every graph presented in this work.

[5]The interactive graph is available at **https://plot.ly/~fabrex/673/frequencies-of-motion-verb-sub-classes-across-the-corpus/**

Class	Verb Type	Verbs
51.1	Inherently Directed Motion	Advance, Arrive, Ascend, Climb, Come, Cross, Depart, Descend, Enter, Escape, Exit, Fall, Flee, Go, Leave, Plunge, Recede, Return, Rise, Tumble
51.2	Leave Verb	Abandon, Desert, Leave
51.3.1	Manner of Motion: Roll Verbs	Bounce, Drift, Drop, Float, Glide, Move, Revolve, Roll, Rotate, Slide, Spin, Swing, Turn, Twist, Wind
51.3.2	Manner of Motion: Run Verbs	Bolt, Bounce, Bound, Bowl, Cavort, Charge, Climb, Coast, Crawl, Creep, Dash, Drift, File, Float, Fly, Glide, Hasten, Hike, Hobble, Hurry, Hurtle, Inch, Jog, Journey, Jump, Leap, Limp, Lurch, March, Meander, Mince, Nip, Pad, Parade, Race, Roam, Roll, Rove, Run, Rush, Scamper, Scoot, Scram, Scramble, Scud, Scurry, Scuttle, Shuffle, Skip, Slide, Slink, Slither, Sneak, Speed, Stagger, Stomp, Stray, Stroll, Strut, Stumble, Stump, Sweep, Swim, Tack, Tear, Tiptoe, Tramp, Travel, Trek, Troop, Trot, Trudge, Wade, Walk, Wander, Whiz, Zoom
51.4.1	Manner of Motion using a Vehicle: Vehicle Name Verbs	Balloon, Bicycle, Bike, Bus, Canoe, Coach, Cycle, Ferry, Motor, Parachute, Punt, Rocket, Skate, Ski, Sled, Taxi
51.4.2	Manner of Motion using a Vehicle: Verbs not associated with Vehicle Name Verbs	Cruise, Drive, Fly, Paddle, Pedal, Ride, Row, Sail, Tack
51.5	Waltz Verbs	Boogie, Bop, Clog, Dance, Jive, Shuffle, Waltz
51.6	Chase Verbs	Chase, Follow, Pursue, Shadow, Tail, Track, Trail
51.7	Accompany Verbs	Accompany, Conduct, Escort, Guide, Lead, Shepherd

Table 5.1 | *Distribution of verbs of motion in the CompWHoB corpus according to their corresponding classes.*

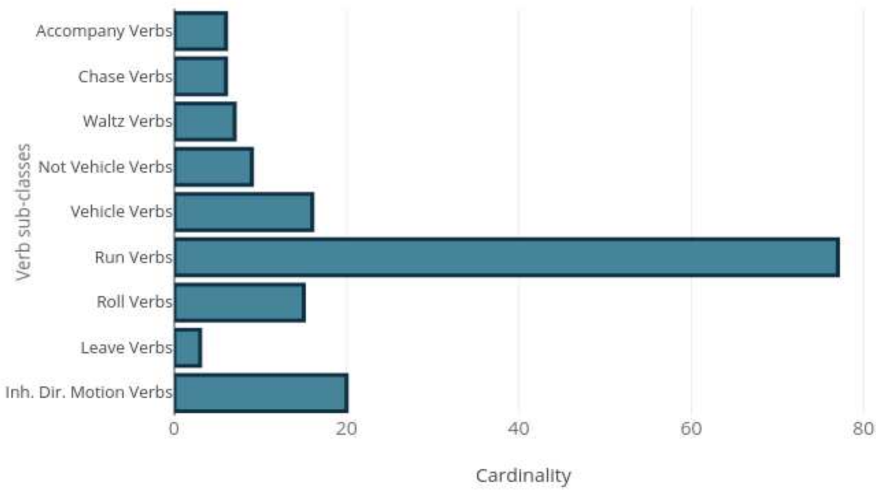

Figure 5.1 | *Cardinality of verbs of motion sub-classes.*

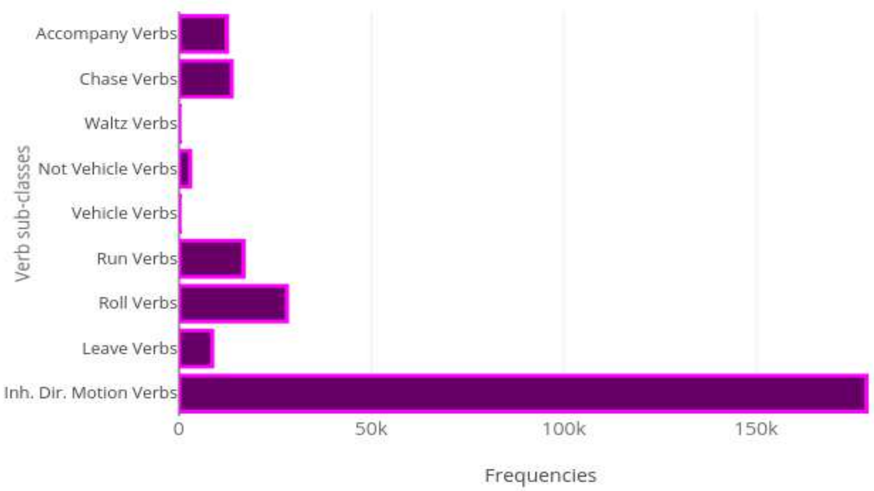

Figure 5.2 | *Frequencies of motion verb sub-classes across the corpus.*

5.2.3 The motion verbs of the corpus

As described in 5.2.2, the identification process returned 149 motion verbs out of Levin's *original* 247. Interpreting the corpus as a long series of question-and-answer between journalists and podium, motion verbs are drawn only from podium's utterances. This choice is based both on research interests and on the features used by the developed systems for metaphor detection. Indeed, on the one hand, as highlighted by the proximization theory (cf. Section 2.5), not only does the podium represent the main character of the press briefings, but it is in their words that the use of motion verbs might play an important role in U.S. political communication strategies. On the other hand, in order to detect the literal or metaphorical use of the motion verb, the systems needs to leverage its wide context of occurrence, hence leading to discard all journalists' questions. Indeed, the average word-length of journalists' utterance is of 30.94 tokens, hence not providing enough useful information for the approach to the task at hand, while for podium's documents[6] is instead reported at 87.96 words. Thus, each motion verb was retrieved along with the utterance in which it is included.

The pie chart of figure 5.3 graphically represents the overall frequency distribution of motion verbs across the corpus[7]. Despite the absence of labels due to the plot space restrictions, it can be observed that about 71% of the frequencies are condensed in the three verbs *go*, *come* and *move*, respectively represented by the orange, light green and dark green slices.

The *inherently directed motion* verb *go* predominates, occurring more than 117000 times in podium utterances and representing the 46.8% the verbs of motion present in the CompWHoB corpus. Added up, *come* and

[6] *Document* and *utterance* are used interchangeably when discussing of the CompWHoB corpus properties.

[7] The graph represented in the dissertation does not show labels since these would cross the borders of the picture. Nevertheless, it is possible to interactively *navigate* the graph at the following link: **https://plot.ly/~fabrex/527/distribution-of-motion-verbs-across-the-compwhob-corpus/**. Colours of the interactive pie chart are different from the ones illustrated here since randomly generated at each run.

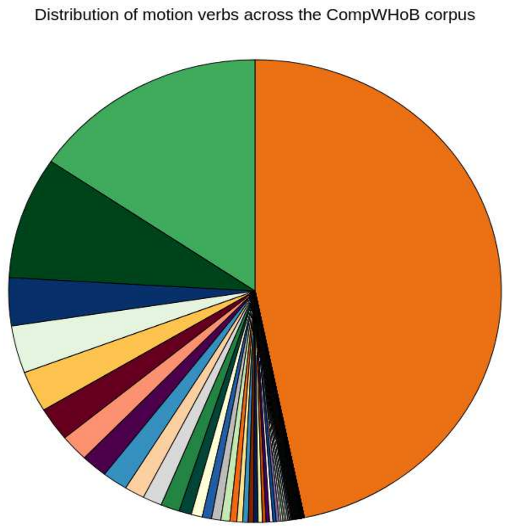

Figure 5.3 | *Motion verbs frequencies distribution in the CompWHoB corpus.*

move reach 60308 occurrences, amounting to about the 24% of the total frequencies. As previously said, it must be pointed out that the prevalence of these verbs is to be connected to their presence in several syntactic constructions, where they play the role of auxiliary verbs (e.g. futuricity expressions) and/or form part of compound verbs with prepositions and/or particles (i.e. phrasal verbs), often *losing* their contemporary meaning. Looking at the remaining nearly 30% of the pie chart described in Figure 5.3, 146 motion verbs are covered, with 16 of them reporting occurrences above the thousand frequencies. Thus, it emerges a picture of 19 verbs *dominating* the scene of the verbs of motion type, occurring in the range of the thousand times and amounting to around the 96% of the total frequencies.

5.2.4 Selecting motion verbs for the task

A selection of the initial list of 149 identified motion verbs was made in order to have a more informative representation of the data. Verbs falling into the range between the 30 and 10000 occurrences were selected for the metaphorical recognition task. High frequency verbs such as *go*, *come* and *move* were discarded since suggesting a wide use in syntactic constructions and in conventional verbal phrases not providing enough information about their metaphorical potential. A low frequency boundary was set at 30 occurrences – hence excluding verbs such as *bowl*, *pedal* and *trail* – since data provided by verbs falling below this threshold were deemed not sufficient for the following stage of training and testing of the system, making the building of an accurate word representation not an easy task (cf. Chapter 6).

Although it might be argued that this threshold is still too low for this learning step, I made this decision based on my research curiosity towards those motion verbs which potential metaphorical use could be of interest for the present thesis. Furthermore, raising the low frequency bar would have drastically reduced the number of verbs selected for the metaphor recognition task.

Applying the range between 30 and 10000 occurrences, 91 verbs were discarded. More precisely, 88 were the motion verbs occurring less than 30 times in the whole corpus, while those which frequency was higher than 10000 were represented by only three verbs, namely *go*, *come* and *move*. Thus, the final selection returned 58 motion verbs, which distribution in the corpus can be observed in figure 5.4[8]. The bar chart represents on the x axis the lemma of each verb and on the y axis their corresponding occurrences.

[8]The interactive graph is available at **https://plot.ly/~fabrex/539/motion-verbs-selection/**

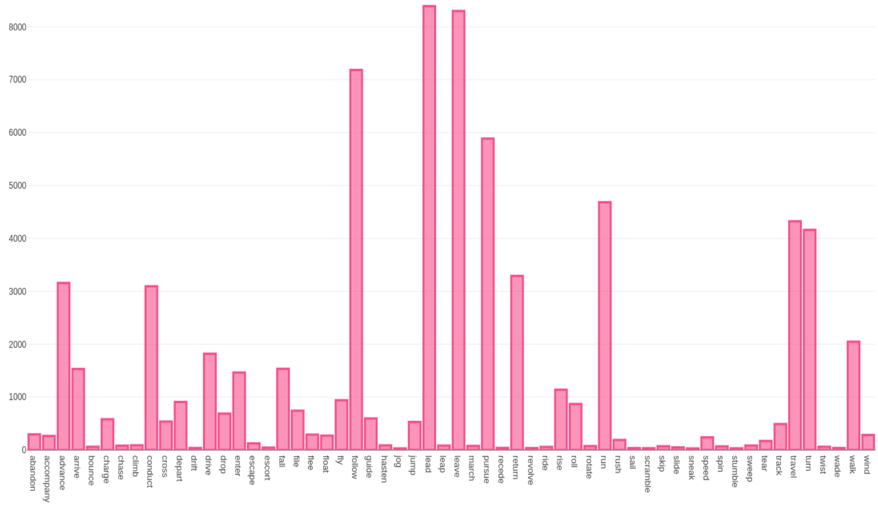

Figure 5.4 | *Distribution of the selected motion verbs in the CompWHoB corpus.*

Around 27% of the verbs are in the thousand units, with *lead*, *leave* and *follow* occurring more than 7000 times. About 33% is instead the slice of verbs within the hundreds range. The remaining 23 verbs display less than 100 occurrences, ergo representing the left 40% of the selection.

The visual representation of the general syntactic-semantic characteristics of the 58 motion verbs resulting from the selection previously described is shown in Figure 5.5[9]. The bubblechart shows on the x axis the frequency of the number of WordNet synsets for each selected verb and on the y axis the frequency of their verb classes retrieved from VerbNet (Karin Kipper, 2006). The size of the bubble is proportional to the number of occurrences of the verb. Hence, the higher its frequencies, the larger its bubble.

[9]The interactive graph is available at **https://plot.ly/~fabrex/557/cwhob-selected-motion-verbs-frequencies-vn-classes-and-wn-synsets/**

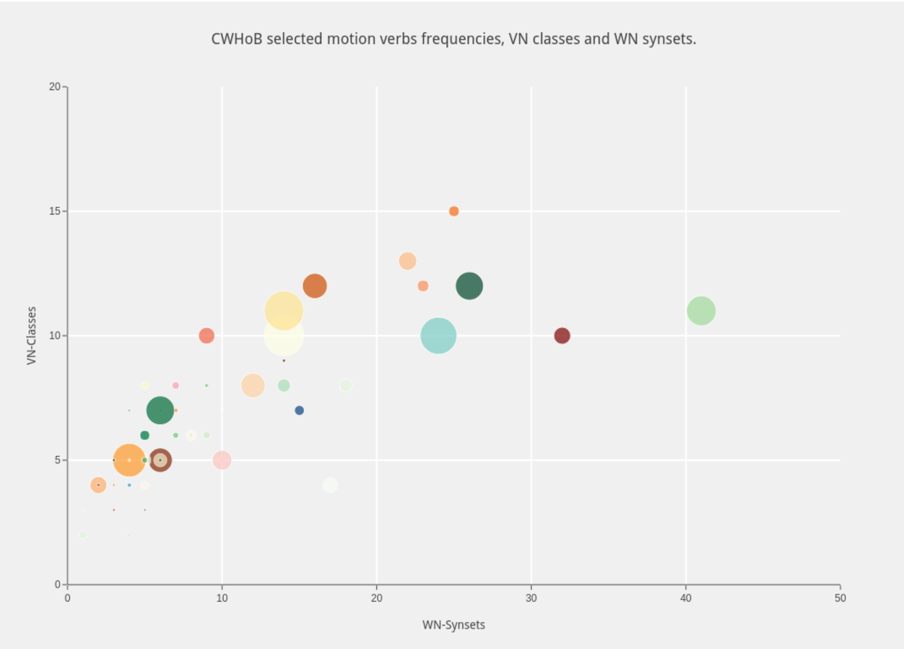

Figure 5.5 | *Semantic and syntactic general properties of selected motion verbs.*

5.2.5 Distribution of the selected motion verbs in the corpus

Politically speaking, the CompWHoB corpus can be ideally divided into four time spans corresponding to the last four U.S. presidencies, i.e. William J. Clinton, George W. Bush, Barack H. Obama and Donald J. Trump (in ascending chronological order). Each presidency is characterised by different political communication strategies which are reflected – among other things – also in the linguistic choices made by the podium during each press briefing.

As described in Section 4.4.3, a total number of 33,124,918 tokens is present in the whole corpus[10] with about 15% of them represented by verbs, reporting a total of 5,014,210 frequencies across the four presiden-

[10] At the time of collection of May 2017.

cies. Making reference to Levin's verbs of motion class, 250,529 are the occurrences of motion verbs collected from the corpus, ergo representing around the 5% of the total of tokens POS-tagged by the SpaCy parser as verbs. It follows that a motion verb is encountered in the CompWHoB corpus every 0.04 verbs or more in general every 0.007 tokens.

In the light of the selection described in Section 5.2.4, let us have now a look at the statistics of the motion verbs *chosen* for the metaphorical recognition task. The frequencies of the verbs falling into the pre-set range add up to 72,351 occurrences across the whole corpus. The *selected* motion verbs amount to the 28.87% of Levin's verbs of motion classes and to the 1.44% of the 5,014,210 verb occurrences in the four presidencies. Thus, a *selected* motion verb can be found every 0.28 verbs of motion identified by Levin and every 0.014 verbs in the corpus. The bar chart in Figure 5.6[11] describes the distribution of the selected motion verbs per each presidency term (e.g. Clinton_1, Clinton_2, Bush_1 and so forth) compared to the one of the whole Levin's class n°51 represented in the corpus.

The grouped bar chart shows on the x axis the three presidencies covered by the CompWHoB corpus[12] where the light blue and the less opaque bars indicate respectively the selected motion verbs and overall class n°51 for each presidency term. Each bar also displays the total frequency count for the corresponding presidency term. The y axis shows the frequencies of the selected motion verbs. As it can be observed, the selected motion verbs distribute quite similarly in Clinton and Bush presidency terms, with a slight increase in their use in both presidential second terms. Comparing the frequencies of the two distributions, it can be observed a quite similar trend in each term as the verbs of motion represent around the 26-30% of the overall instances of Levin's class n°51. However, two outliers can be detected: the first one is represented by the second term of Bush presidency where, despite the higher use of verbs of motion, their percentage

[11]The interactive graph is available at **https://plot.ly/~fabrex/637/distribution-comparison-of-selected-and-non-selected-motion-verbs-in-the-corpus/**

[12]Trump's presidency is excluded from the chart since briefings collection stops at its the early days, hence not providing enough data for comparison.

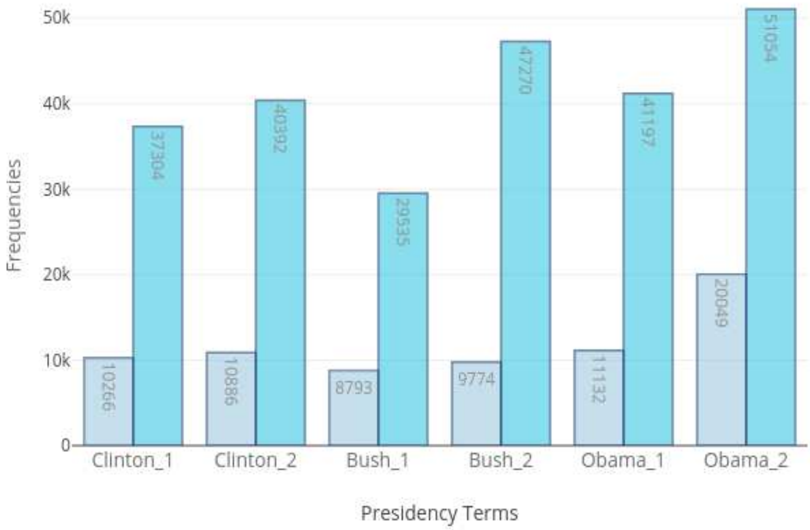

Figure 5.6 | *Distribution of selected and non-selected motion verbs across presidencies.*

decreases at the 20.67%. The second outlier is starkly visible as motion verbs occurrences in the second term of Obama's presidency are almost doubled compared to the first one. In this case, the verbs of motion represent the 39.27% of the overall instances of the class n°51 being used by the podium. The percentage difference recorded at 12.25% between the two terms of Obama's presidency is the higher recorded in this chart (the difference between *Bush_1* and *Bush_2* is indeed recorded at 9.1%). Going back to Table 4.1 of Section 4.4.3, it can be observed that this trend is also reflected in the number of tokens being used during each presidencies, with a significant increase in both Obama and Bush administrations.

5.2.6 Motion verbs representativeness

In order to test the representativeness of the selected motion verbs from the CompWHoB corpus, the CoCA corpus was used as reference corpus for comparison. The CoCA is claimed to be the largest freely available and balanced corpus of (American) English. It is evenly divided into five genres, comprising 220,225 texts and composed of more than 520 million words (I refer here to the latest update of December 2015). Each year of the corpus consists of five genres – i.e. spoken, fiction, popular magazines, newspapers and academic journals – which texts are collected from a variety of sources.

As the CompWHoB corpus only collects dialogical speech discourse in the form of press briefings, the *spoken* section of the CoCA was chosen for data comparison. Collecting the transcripts of unscripted conversations from television and radio programmes (e.g. *All thing considered* (NPR), *Today Show* (NBC), *Good Morning America* (ABC), etc.), this part of the CoCA corpus shares some significant characteristics with the CompWHoB corpus. Indeed:

- data from both corpora are transcriptions of actual spoken conversations;
- although conversations are claimed to be unscripted, a very small percentage is represented by scripted material often used to as catalyst for interaction between speakers;
- as discussed by Davies in the corpus characteristics section of the website `https://corpus.byu.edu/coca/`, these transcriptions cannot be considered as completely natural conversations. *Naturalness* is indeed altered by the fact that speakers know to be at the centre of the stage, being broadcast on (inter-)national television channels and/or radio programmes. This factor inevitably alters their use of language – one need only think of, for example, profanity. Furthermore, in the case of the U.S. press briefings, podium is trained for

dealing with the journalists, often employing precise communication strategies and a selected lexicon.

In the light of the above-mentioned main properties of the two corpora[13], it came the decision of using the *spoken* CoCA as a reference corpus. Looking at the data of the *spoken* CoCA, 109,391,643 is the total number of words in this genre section. Although freely available for access on its online interface[14], the full-text download on one's own machine of CoCA's data is only possible accepting to purchase the data. Thus, the free option download that gives access to a sample of the corpus was exercised.

The *spoken* CoCA sample consists of about 1.7 million tokens (punctuation is excluded). 73,996 are the frequencies of the verbs detected by the SpaCy parser in the data sample, with 4,716 of these represented by those included in Levin's motion verbs classification. With reference to the motion verbs selection discussed in the previous sections of the current chapter, 55 are the verbs being identified in the *spoken* section (*exit*, *jog* and *stumble* are not present in the sample). The frequencies of the 55 *spoken* motion verbs sample were then updated to their proportional estimate, given the total number of words of the whole population (109,391,643 tokens). The filled area line chart of Figure 5.7[15] graphically describes the distribution of the identified *common* verbs from the CompWHoB and CoCA corpora.

The x axis of the plot represents the selected motion verbs in common between the two corpora and the y axis their corresponding frequencies. The blue trace describes the frequencies of the CompWHoB corpus motion verbs while the orange one those included in the CoCA selection. To test the representiveness of the CompWHoB corpus motion verbs, the t-test for two independent sample of scores was performed on the null hypothesis

[13] The spoken genre is obviously a sub-corpus of CoCA but it is regarded here as a corpus on its own due to its characteristics and dimensions.

[14] `https://corpus.byu.edu/coca/`

[15] The interactive graph is available at **https://plot.ly/~fabrex/545/distribution-of-common-selected-motion-verbs-of-compwhob-and-coca-corpora/**

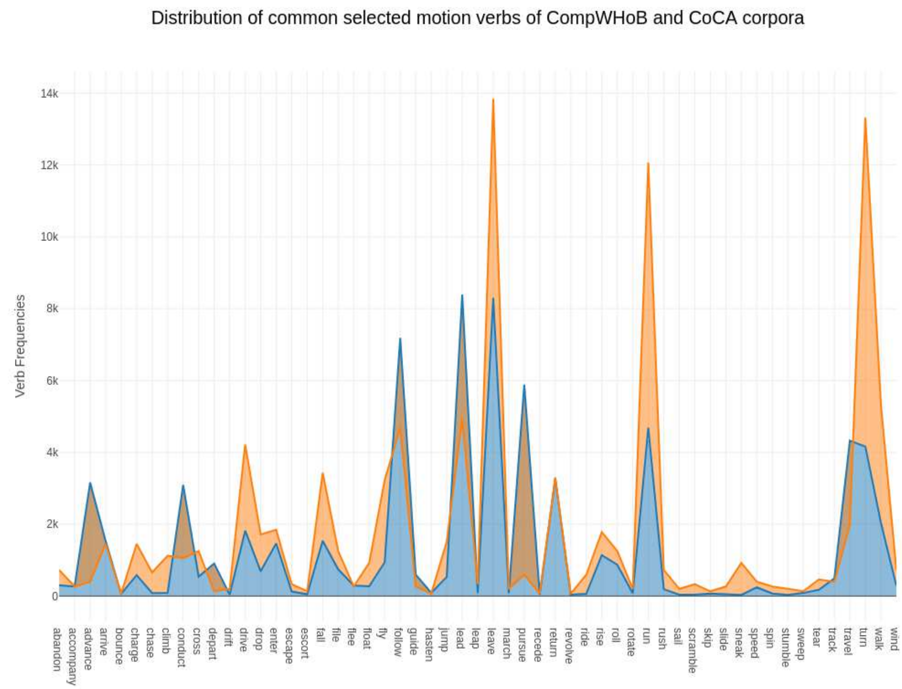

Figure 5.7 | *Distribution of the common selected motion verbs in the CompWHoB and CoCA corpora.*

that there is no significant difference in the average values across samples. The two sided test yielded $p > 0.05$, hence not rejecting the null hypothesis and confirming the representativeness of selected motion verbs with comparison to the larger population of the *spoken* CoCa ones.

5.3 Annotating data for the task

In the field of metaphor computational modelling, annotation for metaphoricity has been mainly playing a double role: exploration of the data (be them later processed by the system or not) and evaluation step of the computational model.

As regards the use of annotated data as evaluation strategy, the reason for that must be traced back to the lack of a common framework for system development and testing in the research on the computational modelling of metaphors, as discussed in Section 3.5.4. Indeed, *missing* guidelines in the field have lead researchers to develop their own evaluation strategies based on the data they were provided with.

So far, studies have described a variegated picture when it comes to the evaluation stage. In Shutova et al. (2013) human annotators were asked to express their judgements on the output of the metaphor processing systems. Mason (2004) carried out one of the two evaluations leveraging the mappings of the Master Metaphor List (Lakoff et al., 1991), to this day the largest catalogue for conceptual metaphors (arising from the works of Lakoff and Johnson (1980) and Reddy (1979)) and possibly resembling one of the few standard benchmarks in the field.

In the present dissertation, annotation for metaphoricity has played the double role of exploration of CompWHoB corpus data and test set for the evaluation of the developed metaphor processing systems. The creation of an *ad-hoc* test set – hence not resorting to any of those already available in literature – was necessary due to the lexical units under analysis and the peculiar characteristics of the texts of the CompWHoB corpus, which are leveraged in the approach proposed in the present work (cf. Chapter 6 for the detailed discussion of the systems). In particular, it is the context of appearance of the verb of motion – namely the whole utterance – to represent an important feature for the computational processing of metaphors. Furthermore, being the CompWHoB a *spoken* corpus, its genre inevitably restricted the possible range of potential data for evaluation.

As previously motivated in Section 5.2, the focus of metaphorical investigation is represented by the verbs of motion only uttered by the podium. Data for annotation come in the form of the whole podium's utterance, i.e. the sentence including the motion verb and possibly the n sentences preceding and/or following it. This choice is motivated by the characteristics of the press briefings: being a sequence of question-answer, it is often necessary to resort to the whole context of appearance of the motion verb to understand its meaning in text. However, in some cases it is not even possible to retrieve it without going back to the question posed by journalists.

5.3.1 Annotation for data exploration

The aim of the first round of annotation was to explore a sample of the corpus data in order to get a picture of how selected motion verbs were interpreted by the annotators in the atypical political context of the U.S. press briefings.

Following the verbs of motion selection described in the current chapter, 193 verb instances were randomly extracted from the corpus, for a total of 97 documents. Three human annotators were chosen for this task according to the following criteria: certified advanced level of English language knowledge, having lived abroad (in an English speaking environment, e.g. English as a second language also in the workplace) for at least 5 months and having a basic background in linguistics. Two annotators were Italian native speakers while the other one was bilingual in English and Italian. Annotators were instructed either in person or via video chat services.

The annotation procedure used as blueprint for its framework the metaphor identification procedure (MIP) proposed by Steen and his colleagues (Pragglejaz Group, 2007; Steen et al., 2010) for the identification of metaphorical words in discourse. The task they were presented to was the identification of the literal or metaphorical use of the motion verbs in context. With the term 'literal', the basic meaning of the verb was intended, i.e. the more concrete and tangible, often the one closest to its etymological one. If the meaning of the verb in the particular context contrasted its basic one but it could be understood in the terms of it, then the verb was to be regarded as 'metaphorical'. The annotators were provided with three options for the labelling of the verbs of motion: `MTP` for metaphorical use of the verb, `LTR` for the literal one and `UNK` whereby it was not possible to identify the reading as either literal or metaphorical. The detailed procedure is illustrated in Appendix A.

Thus, in this first round the focus of the task was the annotation of metaphors at a linguistic level distinguishing between literal and metaphorical use of motion verbs. Fleiss' kappa (Fleiss, 1971) was used to measure agreement reliability between the three annotators, returning $\kappa = 0.35$ ($n = 3, N = 193, k = 3$), where n is the number of annotators, N the total number of motion verbs to be annotated and k the labels available for annotation. The result can be interpreted using Landis and Koch (1977)'s criteria for the classification of the strength of agreement between raters, although the "divisions are clearly arbitrary [but] they do provide useful "benchmarks" "(Landis and Koch, 1977, p.165). According to these intervals, a κ statistic at 0.35 can be regarded as a *fair* agreement between annotators. In order to further assess the inter-annotator reliability, I report here also the Krippendorff's α, since this measure computes the disagreement between raters. On this task, α is measured at 0.35 (with values ranging between 0 and 1).

The two inter-annotator measures reproduce similar values, showing that the task was not as trivial as one would expect. Even if a fair strength of agreement is indicated, the low κ and α scores provide significant information about the very different interpretation of the metaphorical and literal meaning of motion verbs.

Annotators tend to agree more on the metaphorical use of verbs of motion. 55 are the cases on which the three annotators agree on metaphoricity (28.48% of the total motion verbs), against the 44 on which they agree for the literal meaning (22.79% given the total of 193 verbs). Hypotheses for discordance might be varied, although two are the main reasons that seem to emerge from the data: polysemy and possible *dead/frozen*[16] metaphors (Nunberg, 1987). Indeed, high polysemous verbs such as *run* and *turn* (interactively looking at Figure 5.5 it is possible to observe the WordNet synsets of each verb) prove to cause much trouble in annotation, e.g. *follow*

[16]The two terms are used interchangeably in this context.

bringing agreement in only one case. *Dead* metaphors – i.e. those expressions so ingrained in our contemporary language to be assimilated into *literal* use (Nunberg, 1987, p.198) – might represent a cause of discordance between annotators, as verbs such as *advance* and *conduct* showed, with the latter one never being agreed upon by annotators.

Although it is obvious that the dataset must to be considered too small to draw definitive conclusions, these data still provide indicative information about the meaning and the use of motion verbs in the specific atypical political genre of the CompWHoB corpus. Furthermore, they indicate that *dead* metaphors might represent a cause of discordance in the interpretation of the metaphorical expression.

5.3.2 Test-set annotation

In the second round of annotation, data to be used as test set for the evaluation of the computational metaphor modelling systems were annotated. According to the criteria shown in 5.3.1, two were the human annotators selected for the task. Both of them were Italian native speakers, instructed in person by the author of the present work.

The target focus of this second round of annotation reflects the final scope of the present work. Annotators were asked again to identify the literal or metaphorical use of motion verbs in text but this time at different level of metaphorical conventionality. Indeed, the interest of this task – and of this research – was the identification of the novel/unconventional[17] metaphors uttered by the podium in the realm of U.S. press briefings. This metaphorical focus was mainly motivated by three aspects:

- **Metaphoricity of motion verbs**. Motion verbs tend to be used metaphorically to a greater extent, as previous studies have shown (Cameron, 2003; Gedigian et al., 2006). This is particularly true when the historical aspects of metaphors are taken into account in the annotation procedure.

[17] The terms are used interchangeably in this context.

- **Research interest**. Bearing in mind the context of application of metaphors – i.e. the atypical political realm of U.S. Press Briefings – novel metaphors may be indeed revealing of political communication strategies, as the cognitive-pragmatics findings of Cap (2013) have shown.

- ***Frozen/Dead* metaphors issue**. As Nunberg (1987) highlighted, metaphors evolve in time with some of them *losing* their original metaphorical charge, as Section 5.3.1 also seems to suggest. Since they become so entrenched in everyday language to lose their metaphorical *status*, I pursue here the path of exclusion of these verb instances from the range of metaphoricity.

Thus, the description of the metaphorical level of conventionality for the second annotation task is provided in Paragraph 5.3.2.0.1. The characteristics of the test set and the corresponding annotation results are to be discussed in Chapter 7 at Section 7.2.1, when describing the evaluation of the computational metaphorical modelling system.

5.3.2.0.1 Unconventional/Novel metaphors With the terminology *novel* or *unconventional* metaphors, here I refer to those metaphorical expressions that are readily interpreted as such by humans because they "are not systematically used within a language system" (Gelo and Mergenthaler, 2012, p.160). Indeed, their understanding is not automatic since characterised by "marked rhetorical effects, whose comprehension requires a special imaginative leap" (Nunberg, 1987, p.198). These metaphors have the power to break our *logical* reasoning when we encounter them as they bring to light an a-systematic association between two seemingly unrelated domains of experience (e.g. *She was the sun warming my winter days*).

We can figure metaphors' *life* as a trajectory: metaphors are born as unconventional, result of "the idiosyncratic creative process of the speaker" (Gelo and Mergenthaler, 2012, p.160) and intuitively recognised as such by the members of a linguistic community. As metaphors take root in the language system as an integral part of it, their meaning becomes so well-established that “their comprehension becomes more automatic, and their rhetorical effect is dulled” (Nunberg, 1987, p.198). When metaphors reach this stage of their life, they are termed as *conventional*. Speakers process them without any effort, and most of the times without even realising that they are dealing with this figure of speech. Conventional metaphors are “generally established as a mode of thought among the members of a linguistic community” (Lakoff and Turner, 1989, p.55).

Let us take as way of example the following headline from a well-known newspaper webpage:

(22) Time is running out for Madagascar - evolution’s last, and greatest, laboratory[18].

The metaphorical expression *time is running out* is so entrenched in everyday use of English language that some native speakers (but not only them) may find hard to recall its *original* metaphorical nature and its corresponding hidden concept TIME IS AS A LIMITED RESOURCE.

Nunberg (1987, p.198) defines as *dead* or *frozen* those metaphors which common usage lead them not to be different from those terms literally used. Describing the “Career of Metaphor”, Bowdle and Gentner (2005, p.199) suggest that a computational distinction could be defined between *novel* metaphors, i.e. base terms referring to domain-specific concept not yet associated with a domain-general category (e.g. *Science is a glacier*), and *conventional* ones, i.e. base terms referring both to a literal concept and an associated metaphoric category (e.g. *A gene is a blueprint*). Working in the field of political science, Drulák (2005) suggests a more fine-grained division for metaphors adding the *sedimented* layer to the *conventional* and *unconventional* metaphors. The *sedimented* metaphors are interpreted as

[18] `https://www.theguardian.com/world/2017/may/13/madagascar-mass-extinction-plants-kew-gardens`

those communicated as having a literal meaning, also referred to as *dead* metaphors. The *conventional* ones are those processed without any effort as time wears them out, but still holding a metaphorical charge (e.g. *The European Union has three pillars*). *Unconventional* metaphors are those providing insights and interpreted as incongruous, novel, "strong metaphors" (Cameron, 1999, p.131). These works show that the lines of demarcation becomes fuzzy when it comes to try to draw a clear-cut distinction between different degrees of metaphorical usage.

Thus, keeping in mind what said so far, in the present thesis the identification of novel metaphors is realised according to the following binary distinction: under the umbrella term *literal*, instances of motion verbs being used either literally (ergo displaying their basic meaning) or as conventional metaphors are included. Using the label *metaphorical*, only the novel metaphorical usage of motion verbs is recognised.

In the field of the computational modelling of metaphors, Krishnakumaran and Zhu (2007) also pursued the identification of novel metaphors (defined as *live* as opposed to *dead*). However, they do not report the procedure followed in the manual annotation. Literature is very scarce in this respect, with most of the studies presenting detailed annotation procedures focusing on the discrimination between literal and (conventional) metaphorical uses of linguistic items and/or expressions. Thus, here I use as a blueprint for the annotation procedure the manual for Metaphor Analysis in Psychotherapy (MAP) (Gelo, 2008). In fact, MAP provides an operationalised procedure for the identification of metaphors as built on the works in cognitive linguistics of Kovecses (2010) for the understanding of conventional and unconventional metaphors. The detailed procedure presented to the two annotators is described in Appendix 1.

PART III

The Task

CHAPTER 6

A Lexical *Resource-less* Metaphor Recognition System

6.1 Overview of the chapter

In this chapter the systems developed for the detection of the literal or metaphorical use of motion verbs are described. In Section 6.2, the main reasons behind the choice of implementing metaphor processing models employing unsupervised techniques and without resorting to task-specific hand-coded linguistic resources are explained. Section 6.3 explores the design of the systems following the guidelines provided by Shutova (2015). The main characteristics of the data on which the systems are trained and tested and the limitations that arise in dealing with them are also discussed. The approaches at the basis of the systems and the main intuitions guiding their development are described in Section 6.4. Section 6.5 focuses on how the unsupervised methodologies are implemented and used for metaphor detection. Section 6.6 concludes the chapter describing in detail the the algorithms of the three models.

6.2 Why a lexical *resource-less* approach

As previously shown in Chapter 2, metaphors represent a complex conceptual system involving different layers of reasoning. Being the metaphors a reflection of our conceptual system, speakers are not only supposed to share a strong enough linguistic background to properly receive and process the communicated information but they also need to draw on a similar cultural knowledge (Lakoff and Johnson, 1980, p.15).

In order to represent such a complex mechanism, during the last few decades the NLP community have designed computational systems that have more than often relied on the availability of large manually-annotated resources (being them handcrafted lexical resources, ontologies, etc.). Yet these resources – fundamental as they are – also show some downsides. On the one hand, they are inevitably time-consuming. The building and development of such highly-curated and structured repositories entails a serious commitment that often turns their creation into a long-term project. On the other hand, most of these resources represent an exclusive advantage of a very handful of languages around the world. Thus, even if some intuitions may be applicable to different-speaking contexts, the language itself constrains the approach in its restricted realm.

Discarding precious information contained in general-domain lexical databases – such as WordNet and FrameNet among the others – might not represent a wise choice, as they undoubtedly provide precious information to better represent the complex mechanism under analysis. However, as Shutova (2015, p.587) remarks, "it would be an advantage if no such resource is required and the system can dynamically induce meanings in context".

Thus, being able to develop an approach that could as far as possible overcome the limitations posed by the use of general-domain lexical resources has been one of the principal aims of the present work.

6.3 Design of the metaphor recognition system

Working on the computational modelling of metaphors, during recent years Shutova (2015) has stressed the necessity of building a unifying landscape that could define the task at hand, the features of the developed system, the evaluation standards and the linguistic considerations involved in the analysis of metaphors.

Her work represents an invaluable step forward in this field of research as she did not only review the recent approaches proposed so far in literature but also delineated the principal properties that a metaphor processing system should possess. Furthermore, as the same author underlines, for a model to be successful "[t]he design of the metaphor processing task should thus be informed by the possible applications" (Shutova, 2015, p.617).

In the light of what said so far, the design of the systems presented in this thesis and their main characteristics are described below by keeping in mind the guidelines illustrated by Shutova (2015):

- **Task**. The task at hand is the one of metaphor identification. The system identifies every single instance of the verbs of motion as either literal or metaphorical.

- **Linguistic considerations and level of analysis**. As motivated in Chapter 5, the target lexical units under analysis are the motion verbs and the level of analysis is the one defined as *linguistic metaphor*, namely the surface realisation of the metaphor in the text.

- **Level of conventionality**. The system discriminates between a literal and metaphorical use of the motion verb in text. As previously stated in 5.3.2, under the label *literal* the basic contemporary meaning of the target lexical unit or its conventional metaphorical meaning is included. With the label *metaphorical* I refer instead to the unconventional metaphorical use of the target lexical unit.

- **Context of application**. The system operates on continuous, dialogical, naturally occurring texts from real-world data. More specifically, the texts under analysis are represented by the opening statements and the replies to journalists' questions of the podium.

- **Genre and topics**. The genre of the texts is constrained by the nature of the White House Press Briefings, an atypical kind of political discourse. However, as shown in Chapter 4 and in Appendix 1, the topics discussed during each briefing may vary from private matters of the incumbent president to more formal issues.

- **Knowledge**. As motivated in Section 6.2, the system does not rely on either hand-coded rules or general-domain lexical resources. As it will be shown in Section 6.5, the system acquires its knowledge directly from the corpus itself.

- **Word classes and syntactic constructions**. As previously shown in Chapter 5, the system deals with the verb word class and more specifically with a selection of motion verbs drawn from Levin's classification. As for the syntactic construction, the system is primarily evaluated on motion verbs-direct object pairs (verbs of motion in this particular syntactic relation are present at least once in each document). As the utterance may include more than one motion verb (not necessarily followed by a direct object), the system is also evaluated on instances presenting other syntactic constructions different from the one above-described (cf. Section 7.2.1 for more details).

6.3.1 A note on the syntactic constructions

The utterances to be used for the evaluation stage where chosen based on the syntactic criterion described in the previous section. Indeed, only documents including at least one motion verb followed by its direct object were selected. The instances were found by looking for *dobj* arcs with verb (*VB*) head and noun (*NN*), proper noun (*PROPN*) or pronoun (*PRON*) dependent.

The decision to focus on this particular syntactic construction was motivated by the approach to the identification of the literal or metaphorical usage of motion verbs. As it will described more in detail in the next sections, based on the intuition of Shutova et al. (2016) that the linguistic word embeddings of the verb-direct object pairs can actually capture information about the source and target domains, the degree of similarity between verb and direct object becomes a tool for metaphor detection, hence acting in a certain way as a proxy for selectional preference violation.

Although the initial intention was to also include subject-motion verb pairs in order to provide more information for the metaphorical detection and extend the number of data to be investigated, this option was discarded. Indeed, due to the dialogical nature of the press briefings, speakers' utterances are characterised by a high number of anaphoras that makes hard their resolution, since their reference may sometimes lie in the journalists' questions. Furthermore, due to the presence of a high number of complex syntactic structures such as open clausal complements and relative clause modifiers, the retrieval of the correct subject would have represented a task on its own, since most of the time requiring the resolution of its reference. Figure 6.1[1] describes the dependency relations of the selected motion verbs across the CompWHoB corpus. On the y axis the dependencies relations detected are shown. On the x axis their frequencies are displayed.

[1]The interactive graph is available at **https://plot.ly/~fabrex/474/dependency-relations-of-podium-motion-verbs-range-30-10000/** . Please double-click at the centre of the chart to display the full graph.

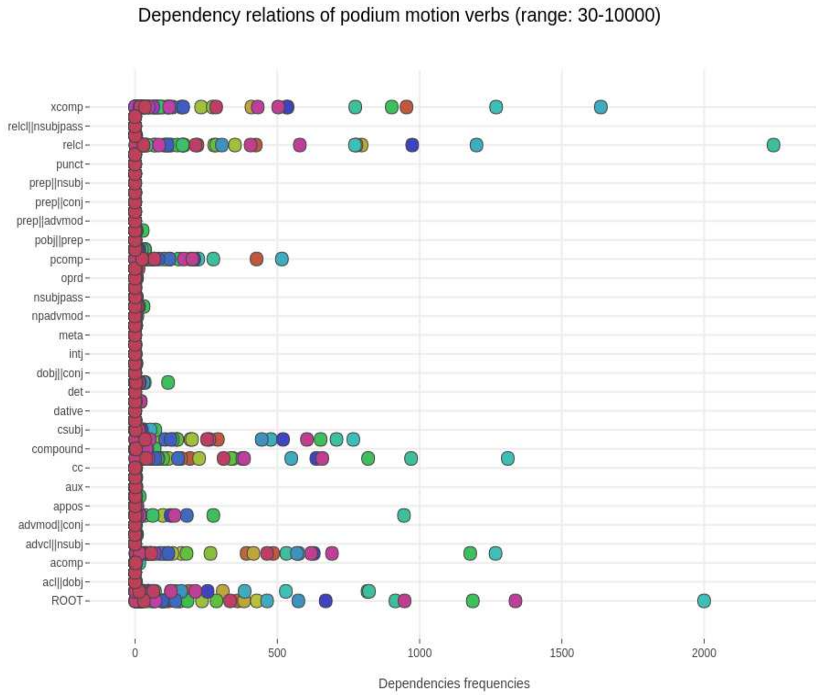

Figure 6.1 | *Distribution of dependencies relations across motion verbs.*

6.4 The Approaches

The approaches adopted in the development of the metaphor processing system presented in this dissertation are inspired by the works of Sporleder and Li (2009), Beigman Klebanov et al. (2009), Shutova et al. (2016) and grounded in the pioneering study on cohesion by Halliday and Hasan (1976). The two approaches are described in Paragraph 6.4.0.0.1 and 6.4.0.0.2.

6.4.0.0.1 First approach The first approach is defined as *global* and can be described as *syntax-agnostic* since only the context of appearance of the target motion verb is leveraged as feature for the detection of its literal or metaphorical use, hence without using any syntactic information. This approach is guided by the work of Sporleder and Li (2009, p.754) since suggesting that the potential metaphorical meaning of an expression can be inferred from the discourse context (although the approach of Sporleder and Li (2009) is more detailed focusing on cohesive ties with the surrounding discourse and not talking of metaphoricity but of literal or non-literal meaning). More precisely, supposing a text to be lexically coherent, the target motion verb should figuratively break this coherence when used metaphorically. When literally used instead, the lexical coherence of the text should be preserved. Using the words of Halliday and Hasan (1976, p.18) and Morris and Hirst (1991, p.21), the notion of cohesion is defined here as "the set of possibilities that exist in language for making the text hang together", since it "is not a guarantee of unity in text but rather a device for creating it". A group of sentences and phrases acting together as whole constitutes a cohesive text; if the group of sentences and phrases also describe the same topic then the text can be considered coherent too. The *lexical cohesion* is then "the cohesion that arises from semantic relationships between words" (Morris and Hirst, 1991) and also a strong clue of coherence in text. In a nutshell, using this first approach, a potential metaphorical expression should be detected as such if not semantically related with the words describing the topic(s) of conversation.

Previous approaches to the detection of literal or non-literal expressions have indeed considered lexical coherence as cue for metaphorical detection. In particular, as also discussed in Section 3.5.3, Sporleder and Li (2009) proposed a token-based unsupervised method identifying *lexical chains* as a cue for the detection of idiomaticity in text. Two cohesion-based classifiers were experimented, one computing lexical chains in text and the other one building a cohesion graph. Furthermore, working on a large corpus of media discourse documents, Beigman Klebanov et al. (2009) successfully hypothesised that words coherently representing the topic under discussion are less likely to be used metaphorically, then suggesting that the metaphors might be found in words "breaking" the lexical coherence of a text.

6.4.0.0.2 Second approach Inspired by the work of Shutova et al. (2016), the second approach takes a closer look to the syntactic information around the target motion verb, only when a verb-direct object relation is in place. For this reasons it is defined as *local*. The hypothesis is that a low semantic similarity observed between the verb and its direct object should hint at the presence of a metaphor in text. In this case, the verb and its direct object belong to two different conceptual domains. The choice of selecting only the verb-direct object relations is motivated by the inherent characteristics of the context of application, as discussed in Section 6.3.1.

6.5 The Method

6.5.1 Intro

The hypothesis presented in this dissertation suggests that when the target motion verb is lexically incoherent with the surrounding context, it may represent an indicator of metaphoricity. The context of appearance of the target motion verb – i.e. the utterance of the podium – is considered as the whole text representing either a statement or a reply to journalists' questions by the podium.

Let us take a look at the example (23). It is 1996 and the incumbent president of U.S. is Bill Clinton. The topic at hand is foreign policy and more precisely the peace negotiations in the context of the Israeli-Palestinian conflict. White House Press Secretary Michael Curry has just handed over to a senior administrator official with a more in-depth knowledge of the recent developments in the meetings between Arafat and Netanyahu. Answering a question coming from a journalist regarding the role of the United States in the peace talks, the senior administrator official replies by saying:

(23) *"Well, first, they've got enough that they have to resolve between them in terms of the implementation, the remaining issues of the implementation of the Interim Agreement. And I feel that, in fact, once* <u>*you*</u> ***cross*** *the* <u>*threshold*</u> *of having worked out agreements you will begin to change the landscape between them. It will begin to further change their attitudes not only about each other, but about what their relationship [pause] I don't mean individually, I mean people [pause] is going to be."*

Detecting the metaphorical use of the verb *cross* in this excerpt seems to be a reasoning not hard at all for us humans, notwithstanding the complex linguistic structure of the sentence itself. For starters, the ability to to process the referential ambiguity of the pronoun *you* might be taken for granted but it is not trivial at all. Indeed, we can only understand that it refers to the potential interference of the U.S. in the negotiation talks only looking at the question uttered by the journalist. But then, how do we realise that the verb *cross* is used metaphorically? Maybe we recognise the degree of abstractness of the direct object *threshold* and understand that it does not refer to a physical object but to a concept "distant from immediate perception"(Turney et al., 2011). Maybe having recognised the abstract property of the term *threshold*, our lexical knowledge tells us that the verb *cross* is then used metaphorically since violating selectional restrictions and making *threshold* standing out as a semantically incompatible argument.

No matter which reasoning helps us process this information, we humans can successfully carry out such a complex linguistic task in a very limited short time – instantly and effectively – and the key ingredient of this process seems to lie in our linguistic knowledge. A question then arises from this depicted scenario:

How can we teach a machine to recognise metaphors when no lexical knowledge has been made available to it?

6.5.2 Modelling machine *knowledge* using word embeddings

As previously stated, the metaphor processing system developed in this thesis does not rely on any lexical resource. Indeed, the *linguistic knowledge* of the system comes directly from the corpus itself, in the shape of the so-called *word emebeddings*.

As first step, the words dense vector representations known as *word embeddings* were learnt employing Word2Vec model, since proved to successfully encode the semantic meaning of words in each vector (Mikolov et al., 2013b). To carry out this task, the Python implementation of Word2Vec provided by the Gensim (Rehurek and Sojka, 2010) library[2] was used.

The skip-gram model with negative sampling was chosen. This technique indeed learns fine-grained word representations when trained on large collection of data (Mikolov et al., 2013c). The number of 'noise words' to be drawn was fixed at 10. The threshold for random downsample of high-frequency words was set to $1e$-3 in order to improve word embeddings quality (Mikolov et al., 2013c). Furthermore, Goldberg and Levy (2014) point out that sub-sampling and rare-pruning seem to increase the effective size of the window, in this way making the similiarities more topical. The dimensionality of the feature vector was fixed at 100 while the size of the symmetric window to 5.

The embeddings were trained on a linguistic pre-processed version of the CompWHoB Corpus for 5 epochs. The texts were indeed pre-processed using SpaCy. Each token is represented by its lemmatised form plus its POS-tag according to the Google Universal POS tag set [3] in the format `lemma|POS`, in order to distinguish between homographs words with different POS (e.g. "I **run** home", "They are on the **run**").

Finally, the choice of training the system on the CompWHoB Corpus is motivated by the will to focus on the atypical genre, the specific dialogical nature of texts and the use of language made in the U.S. press briefings.

[2]https://radimrehurek.com/gensim/models/word2vec.html

[3]http://universaldependencies.org/u/pos/

6.5.3 Modelling context using LDA

As said in Section 6.4, the intuition behind the approach is to leverage the context of appearance of the target motion verb for the detection of the metaphorical or literal use. The context is defined by the sentence including the target motion verb and the sentences preceding and following it (if any). More formally, the target motion verb is defined as mot_verb. A document d is under analysis if it includes the sentence containing mot_verb, defined as s_t. The context consists of s_t and the N sentences preceding and following s_t. Coming back to the example (23), the context is in this case formally represented as $1N + s_t + 1N$.

Thus, the context of appearance of the target motion verb is modelled using topic modelling, in particular employing the LDA generative probabilistic model. This technique was also implemented in Python using the LDA module[4] provided by Gensim.

Prior to the training of the LDA model, a linguistic pre-processing step was carried out. In order to generate topics semantically meaningful, only lemmas tagged as nouns were kept, according to the format `lemma|POS` previously introduced in Section 6.5.2. Furthermore, lemmas that appeared in less than 1000 documents or more than 30% documents were discarded. The *ad-hoc* stoplist developed by Esposito et al. (2016) was used to further filter the data since taking into account the main features of the linguistic genre at hand. The stoplist includes indeed all personal and indefinite pronouns as well as the most commonly used honorifics since often used in addressing the podium. The first names of the press secretaries in office during the period of time covered by the corpus are also part of the stoplist as most of the time used only as nouns of address (Brown and Gilman, 1960).

After having completed the linguistic pre-processing step, the LDA model was run on the training corpus employing the online variational Bayes (VB) algorithm (Hoffman et al., 2010). In fact, being based on online

[4]`https://radimrehurek.com/gensim/models/ldamodel.html`

stochastic optimisation with a natural gradient step, the LDA online succeeds in converging to a local optimum of the VB objective function. The number of latent topics to be extracted from the training corpus was set to 50. The model was updated every 150 documents giving one pass over the corpus and setting at 70 the maximum number of iterations for topic convergence. Once the LDA model estimation stage was complete, topic distribution inference was run on the unseen data.

6.6 Discriminating between literal and metaphorical meaning

In this section, the metaphor recognition systems implemented for the the detection of the literal or metaphorical use of motion verbs are described. Three different models were developed to investigate the performance of the system on this task.

The first model leverages only the information contained in the context of appearance of the motion verb, i.e. the podium's utterance. The second model adds a new feature since taking also into account the syntactic information of the verb-direct object pairs. The third model combines the two previously described approaches to investigate if the LDA system can inform the syntactic pair relation and improve the final overall performance.

6.6.1 Literal or Metaphorical

The discrimination between literal and metaphorical use of the motion verbs is interpreted as an (unsupervised) binary classification task. In order to identify the behaviour of the target unit, two thresholds were determined for the scores of the *global* and *local* approaches respectively. Classification thresholds for the three models were optimised maximising the F-score on a small annotated development set. In the case of the models combining the *global* and *local* approaches, the thresholds were considered as dependent variables and determined accordingly. As for the final classification step, the target motion verbs with values above the set threshold were categorised as literal while those with values below the threshold were considered to be used metaphorically.

6.6.2 *Global* model

The first model developed has been defined as *global* since it only uses the information provided by the wide context of appearance of the motion verb, without leveraging the syntactic structure of the sentence. More precisely, the model employs the output of the topic inference on documents – i.e. a distribution of words – and the corresponding word embeddings generated by applying Word2Vec on the training corpus. Indeed, LDA describes each document d as a multinomial distribution over topics where each topic is defined as multinomial distribution over words in a fixed vocabulary. The hypothesis suggested here is that comparing the vector representation of the motion verb mot_verb with the embedding of each word forming part of the topic representing the document d, it is possible to recognise the literal or metaphorical use of mot_verb. The assumption is that the multinomial distribution over words describe a coherent topic. Ergo, if the motion verb embedding shows a low semantic relatedness with each of the vector representations of the words in the topic, its use is categorised as metaphoric since *breaking* the lexical cohesion of the document. The semantic relatedness between the individual word representations is measured using cosine similarity.

Let us define as T the list of topics returned by the topic-inference stage on the unseen document d and t_n as the nth word embedding generated by the Word2Vec model (a lemma POS-tagged as noun according to the `lemma|POS` format) in T, cosine similarity is then defined as:

$$cos\_sim(v_i, t_n) \tag{6.1}$$

where v_i is the ith vector representation of the target motion verb under analysis in d. Cosine similarity is formally computed as:

$$cos(x, y) = \frac{x \cdot y}{||x||||y||} \tag{6.2}$$

Algorithm 1 in Appendix C describes the procedure devised for the recognition of either a literal or metaphorical use of the target motion verb. It can be divided in two main blocks:

1. Given the word embedding of ith motion verb mot_verb in a document d, the algorithm first computes the cosine similarity between v_i and the t_n in T. Every similarity score is then appended to the list C_{v_i}.
2. Every similarity score in C_{v_i} is then compared to the threshold previously defined on the development set. If every similarity score in C_{v_i} is below the threshold of the model, then v_i is recognised as metaphorical. Otherwise, if just one of the similarity scores is above the threshold, then v_i is categorised as literal.

6.6.3 *Glo-cal* model

The second model is named after the combination of the previously described *global* approach and the focus on the *local* syntactic relation of the target motion verb with its direct object. As stated in Section 6.4, the attention on the local context is inspired by the work of Shutova et al. (2016) on the hypothesis that if two linguistic units belong to two different conceptual domains, a low semantic similarity should be observed and a metaphorical use should be detected.

Thus, the local context is considered as the syntactic relation between the target verb of motion and its direct object. As proposed by Shutova et al. (2016), cosine similarity is used to measure the semantic relatedness of the syntactic pair:

$$sim(v_i, dobj_i)$$

where v_i is the ith target motion verb word embedding under analysis in a document d and $dobj_i$ is the ith vector representation of the direct object whose head is v_i. As shown in Section 6.6.2, semantic relatedness is measured using Formula 6.2.

Thus, the *glo-cal* model separately combines the *global* method with the syntactic *local* approach. More precisely, if the target motion verb mot_verb_i does not have a direct object, then the *glo-cal* model operates using the *global* procedure. If a verb-direct object relation is instead detected, the system then computes the cosine similarity between v_i and $dobj_i$. If the similarity score is above the set threshold, then the meaning is labelled as literal. Otherwise, the use of mot_verb_i is recognised as metaphorical.

The Algorithm 2 in Appendix C shows how the *glo-cal* model identifies the literal or metaphorical meaning of the target motion verb. It can be divided into three steps:

1. The *FindDirObj* runs through the text looking for the motion of verb and its possible direct object in the document d. d is parsed using SpaCy dependency parser.

2. If *FindDirObj* finds the mot_verb_i without a direct object, then the system proceeds as in the *global* algorithm and measuring similarity according to the *global* threshold.

3. If *FindDirObj* successfully returns the verb-direct object pair, then the system measure the semantic relatedness between v_i and $dobj_i$. If the similarity score is above the *glo-cal* threshold, mot_verb_i is identified as literal. Otherwise, its meaning is recognised as metaphorical.

6.6.4 *Glo-cal* weighted model

The third and last model developed in this thesis combines again the *global* and *local* procedures in one algorithm. Differently from the *glo-cal* approach, this system uses the output generated from the *global* model to inform the *local* strategy. More precisely, the *glo-cal weighted* model applies the *global* approach to the instances of the target verbs of motion having a direct object. The model then moves on to the *local* procedure described in 6.6.3 and on the basis of the resulting label assigned by the *global* algorithm block, a weight is added or subtracted to the final semantic relatedness measurement.

The Algorithm 3 in Appendix C describes in detail the devised procedure. It consists of four main steps:

1. The *FindDirObj* runs through the text looking for the motion of verb and its possible direct object in the document d. d is parsed using SpaCy dependency parser.

2. If *FindDirObj* does not find a verb-direct object syntactic relation, mot_verb_i literal or metaphorical meaning is then identified using the *global* approach

3. If *FindDirObj* does find a verb-direct object syntactic relation, the *global* algorithm block is first applied

4. According to the resulting output of the *global* algorithm, a weight is added or subtracted in the subsequent semantic relatedness measurement of the v_i and $dobj_i$. If the similarity score is below the *glo-cal* threshold, the use of mot_verb_i is then categorised as metaphorical. On the contrary, if it is above the *glo-cal* threshold, the algorithm identifies the use as literal.

CHAPTER 7

Evaluation of the system: results and discussions

7.1 Overview of the chapter

In this chapter the evaluation of the metaphor recognition systems is presented. Section 7.2 illustrates the dataset used for the training and testing of the three models. Section 7.3 describes the evaluation metrics employed for measuring systems' effectiveness. The baseline set for models performance comparison is discussed in Section 7.4. The results of each of the three models are reported and analysed both quantitatively and qualitatively in Section 7.5. Finally, overall performance on metaphor detection task is discussed in Section 7.6.

7.2 Dataset

The dataset employed in the evaluation of the metaphor recognition systems was built from the CompWHoB corpus. As discussed in Section 5.2.3, data were randomly selected only from podium's utterances and split into training, development and test set.

The training set represents the largest part of the dataset and was used to learn the 100-dimensional word embeddings and train the LDA model subsequently employed for the topic inference step. *Unseen* data were used for the building of development and test sets. The development set is a small collection of podium's utterances annotated by the author of this work according to the literal or metaphorical distinction described in Section 5.3.2, employed for tuning the parameters of the systems (e.g. number of inferred topics for each document, negative and positive weights, etc.) and determining the optimal classification thresholds for the *global* and *local* approaches (cf. Section 6.6.1). In order to be representative of the test set, at least one instance of each motion verb included in the selection described in Section 5.2.5 is present in the development set.

7.2.1 Test set

The test set on which the system was evaluated is represented by the utterances used for the annotation task introduced in Section 5.3.2. Since the main syntactic focus of investigation is represented by motion verb-direct object pairs, documents collected for the building of the test set *primarily* meet this criterion. The reason I highlight here the adverb *primarily* is due to the nature of the document itself. In fact, when retrieving the whole utterance, more than one motion verb may be included in it, moreover possibly not showing the particular syntactic construction. Thus, although the initial intention was to equally distribute the number of instances of the selected motion verbs in the test set, some verbs tend inevitably to *prevail* over the others due to their frequent use in the atypical political genre of U.S. press briefings. Furthermore, the higher frequencies of some verbs can be explained by either the *reticence* of some motion verbs to take a direct object in this particular genre or their subcategorisation restrictions.

As described in Section 5.3.2, in annotating the podium's utterances of the test set, human judges were asked to identify the unconventional metaphorical use of the selected motion verbs, as opposed to their literal or conventional metaphorical use. They were provided with three labelling options: `MTP` for metaphorical use of the verb, `LTR` for the literal/conventional metaphorical one and `UNK` whereby it was not possible to identify the reading as either literal or metaphorical. Inter-annotator agreement was in this case measured using Cohen's kappa Cohen (1960), yielding $\kappa = 0.79$ ($n = 2, N = 1221, k = 3$), where n is the number of annotators, N the number of annotated instances and k the labels available for annotation.

One source of disagreement between annotators can be *tracked down* to the interpretation of metaphorical conventionality, as boundaries between novel and conventional metaphors may appear sometimes very fuzzy. Furthermore, one needs to take into account also the atypical political genre of the U.S. press briefings and the corresponding use of language made by its main characters, i.e. the podium(s). Differences in annotations stem also from the limited context in which some motion verbs are *placed*. Indeed, being it hard to fully comprehend the general meaning of the utterance – i.e. the topic of conversation – due to the restricted context (it may consist of just one sentence, the one including the target verb instance), the identification of the metaphorical or literal use of the motion verb may depend on this uncertain interpretation. Finally, a certain percentage of disagreement may lie in the very annotation procedure framework since leaving in the end a significant room for subjective judgement, notwithstanding the structured sequential instructions annotators were provided with.

The system was evaluated only on the instances which both the annotators could agree on. 1147 motion verbs were indeed included in the test set. From the total number, 40 were the instances discarded during evaluation since wrongly POS-tagged by the SpaCy parser (mostly adjectives labelled as verbs), leading to a final number of 1107 instances and 488 documents. The stacked bar chart in Figure 7.1[1] describes the distribution of

[1]The interactive graph is available at **https://plot.ly/~fabrex/585/literal-and-metaphorical-readings-across-test-set/**

the motion verbs and their literal and/or metaphorical reading.

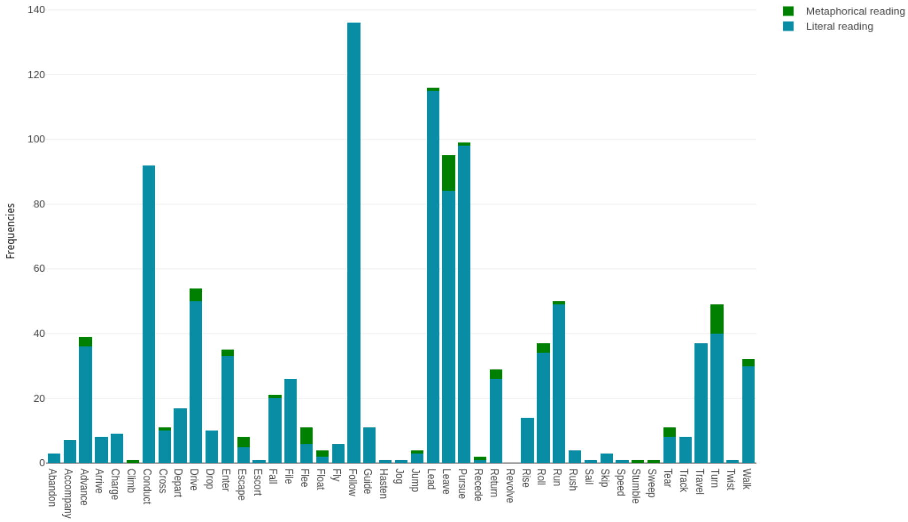

Figure 7.1 | *Distribution of motion verbs in the test set and the corresponding literal and/or metaphorical reading.*

As regards the annotations, the test set turned out to be highly unbalanced. Indeed, only 60 are the instances of motion verbs tagged as metaphorical, while the remaining 1047 were interpreted as either literal or conventional metaphorical. Metaphors are evenly distributed across the test set, showing this trend not to be determined by the random selection of podium's utterances. Although the test set is too small to draw definitive conclusions, this result might indicate that the sample represents a reflection of motion verbs metaphoricity across the CompWHoB corpus.

As regards the syntactic focus of analysis, 738 are the verb-direct object pairs investigated while 369 are the motion verbs in different syntactic relations. The largest part of metaphorical motion verbs distributes across those instances followed by a direct object, amounting to the total number of 55 occurrences. The Table 7.1 reports the reading distribution of the test set, where `N` stands for the frequency of the particular literal (LTR) or metaphorical (MTP) annotation in the test set.

Reading	N	%	Context	N	%
LTR	1047	94.58	Local	687	65.62
			Non-local	360	34.38
MTP	60	5.42	Local	51	85
			Non-local	9	15
Total	1107	100			

Table 7.1 | *Distribution of readings in the test set.*

7.3 Evaluation Metrics

The evaluation metrics reported in this work follow the measures indicated by Sebastiani (2002) for categorisation effectiveness. Considering the method proposed here as a two-class classifier, *macroaverage* is used for evaluation, in order to give equal weight to each class. *Microaverage* is not taken into account because yielding the same score as accuracy. The four metrics employed for the evaluation of the system effectiveness are the following:

- **Accuracy**: Accuracy is the number of the motion verbs correctly identified as either literal or metaphorical given the total number of instances.

- **Precision**: Precision is considered as the number of the correctly identified motion verbs given all the instances assigned by the model to that particular class. Macroprecision is computed as the average of the two precision scores (i.e. literal and metaphorical precisions). Hereafter, with the term 'precision' I refer to macroprecision.

- **Recall**: Recall is considered as the number of correctly identified instances given all the instances of that particular class present in the test set (i.e. the total number of gold-standard annotations for that specific class). Macrorecall is computed as the average of the two recall scores (i.e. literal and metaphorical recalls). Hereafter, with the term 'recall' I refer to macrorecall.

- **F-score**: F-score is the harmonic measure of the *globally* computed precision and recall. In this case, with F-score I refer to the macro-F-score measure, i.e. defined as:

$$2 \cdot \frac{macroprecision \cdot macrorecall}{macroprecision + macrorecall} \tag{7.1}$$

7.4 Evaluation baseline

Although previous approaches focused on verbal targets including also frames of motion (Gedigian et al., 2006), finding datasets manually annotated for metaphoricity featuring a high number of occurrences of motion verbs is not an easy task. Even when such resources are available (Steen et al., 2010), the large training required by the LDA model for the inference of optimal topics represents a serious limitation in the definition of an external baseline. Furthermore, to the best of my knowledge, the metaphor processing model introduced in this work is among the few ones to present unsupervised techniques for the identification of motion verbs metaphoricity in naturally-occurring, dialogical continuous texts from real-world data.

Thus, due to the lack of resources annotated for novel metaphors also providing large contexts for LDA topic inference and due to the highly unbalanced nature of the test set, the *global* model was chosen as baseline of evaluation since relying only on motion verbs context information when compared to the fully featured *glo-cal* and *glo-cal* weighted techniques.

The features used by each model are summed up in the following lines:

- *Global* model

    1. Topics for each document as a distribution of lemmatised word pos-tagged as nouns.

- *Glo-cal* model

    1. Topics for each document as a distribution of lemmatised word pos-tagged as nouns.
    2. Verb-direct object dependency relations.

- *Glo-cal* weighted model

    1. Topics for each document as a distribution of lemmatised word pos-tagged as nouns.
    2. Resulting output from the *global* approach applied as a positive or negative weight on the verb-direct object degree of similarity.
    3. Verb-direct object dependency relations.

7.5 Analysis of results

In this section, the results of the three models on the metaphor identification task are discussed. After showing the performance comparison of the three systems in Section 7.5.1, each model is then analysed separately. In order to discuss results both in quantitative and qualitative terms, 7 utterances including motion verb instances are analysed to evaluate how the model performs in recognising novel metaphorical use when the linguistic structure influences its interpretation. More precisely, 7 documents are randomly chosen based on Dunn (2013b)'s definition of *saturated* and *unsaturated* utterances (cf. Section 2.4) and labelled as metaphorical by annotators. An utterance is defined as unsaturated when it contains elements from both the source and target domains. On the contrary, it is categorised as saturated if it only includes elements from the target domain, making it reading either entirely metaphorical or non-metaphorical.

In some cases, the 7 selected utterances are reported here without their wide context, due to space limitations. The full paragraph can be found in Appendix D.

Saturated Utterances

(24) Well, go back to use Peter Welch as an example, which is that Peter Welch would agree that **leaving** the lights on at a federal building overnight with nobody there is bad government. So why spend our.

(25) [...] I know that the President has made it clear that this is the effort, this was the train that's **leaving** the station, and that he expects everyone [pause] you know, this is our opportunity. [...]

(26) [...] At the same time, the President acknowledges that he's **leaving** the national stage. [...]

Unsaturated Utterances

(27) [...] Not as fast as we would like; it certainly hasn't **turned** Syria into a Jeffersonian democracy that reflects the pluralism and diversity of that country.

(28) [...] And one important piece of context is simply that there was an historic wave that **entered** office at the end of 2008 and the beginning of 2009 of Democratic elected officials who benefitted from President Obama being at the top of the ballot in 2008. [...]

(29) [...] There is an opportunity for the North Korean government to **escape** the deep isolation that they currently face. [...]

(30) [...] I'll think he'll talk about how far we've come in shaping an architecture in the Asia Pacific for the United States to **lead** and to be at the table in forums like ASEAN and the East Asia Summit. [...]

Furthermore, in order to investigate how the system performs in evaluating motion verbs instances presenting a conventional metaphorical use and being labelled as literal by annotators, 3 more utterances are randomly chosen from the test set. The three examples present well-established idiomatic expression. Due to space limitations, utterances are here reported without their wide context (full documents can be found in Appendix D):

(31) [...] And it's also the retraining aspect of that [pause] that as people get older and certain industries start to **turn the corner** because of technology, that we're allowing people the opportunity for retraining to give them the skillset that they need to reenter the workforce and continue to be productive. [...]

(32) [...] This bill **falls far short of** that. [...]

(33) [...] It **turns out** Republicans and Democrats were able to **roll up their sleeves** and work together on one of the most challenging and complex issues of our time. [...]

After having analysed the performance of each model, the major challenging issues to be handled by the models are described in Section 7.5.5. Finally, the overall results are discussed in Section 7.6.

7.5.1 Performance models comparison

In this section, the performance of the each model evaluated against the manually annotated test data are presented. Table 7.2 shows the results of the system in recognising the literal or metaphorical use of motion verbs according to the metrics described in Section 7.3. Thus, precision, recall and F-score measured on both literal and metaphorical motion verbs are reported. The overall performance of the system is represented by the *macroaverage* of the three metrics. The accuracy of the model is also reported here, although its reliability is biased due to the unbalanced test set.

Model	A	Type	P	R	F1
Global	0.74	G_m	0.11	0.53	0.18
		G_l	0.96	0.76	0.85
		G_o	0.53	0.64	0.58
Glo-cal	0.60	G_m	0.10	0.78	0.17
		G_l	0.97	0.59	0.74
		G_o	0.54	0.69	0.60
Glo-cal weighted	**0.74**	G_m	0.12	0.63	0.21
		G_l	0.97	0.74	0.84
		G_o	**0.54**	**0.69**	**0.61**

Table 7.2 | *Summary of the metaphor recognition systems performance on the test set.*

The results in Table 7.2 are reported as follows: G_l refers to the performance of the model on recognising the literal use of motion verbs while

G_m stands for the detection rate of their metaphorical meaning. G_o is the overall performance of the model. The results of the best model are marked in bold.

7.5.2 *Global* model analysis

As described in Section 6.6.2, the *global* model uses as information for the detection of the literal or metaphorical use of motion verbs only their wide context of appearance, i.e. the whole utterance including it. In a nutshell, it is the degree of similarity between the LDA topic terms describing the document and the motion verb itself to define its reading.

Taking a look at the results in Table 7.2, the common thread running through the three models is represented by the higher recall when compared to the attained precision. This trend is also revealed in the *global* system.

As it can be observed, the reason for this significant difference is to be found in the performance of the model described by the G_m rates. Indeed, the *global* model successfully identifies more than half of the metaphorical instances in the test set (32 out of the annotated 60). However, the high number of metaphorical false positives results in an inevitably low metaphorical precision.

On the contrary, performance on the literal detection of motion verbs usages results in high scores in precision, with 798 occurrences out of the total 1047 correctly identified by the system. In Figure 7.2[2], the plot of the confusion matrix for the *global* system can be observed.

Having a closer look at the syntactic context of the motion verbs, the *global* system identifies the 54.90% of the instances presenting a verb-direct object dependency relation and the 44.44% of those not presenting the particular syntactic structure under analysis. When dealing with verb-object pairs, the model tends to return a major number false positives among metaphors ($FP = 141$) if compared to the instances not presenting the

[2]The interactive graph is available at **https://plot.ly/~fabrex/601/confusion-matrix-global-model/**

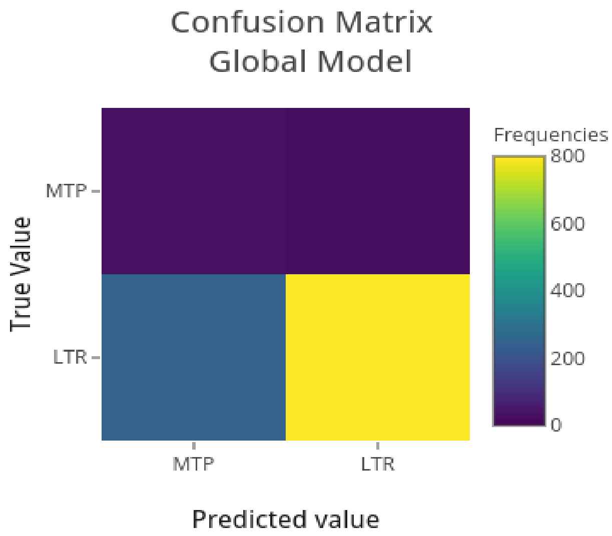

Figure 7.2 | *Confusion matrix plot for the global model.*

particular syntactic constructions ($FP = 108$). However, one must bear in mind that only 9 are the motion verbs tagged as metaphorical not having a specific syntactic focus, as Table 7.1 shows. Looking at the performance of the model on the literal detection, the *global* system correctly identifies the 79.47% of the transitive motion verbs and the 70% of the ones not having a direct object.

As regards the performance in recognising the metaphoricity of motion verbs in saturated utterances, the system correctly identifies as metaphorical the instance in (25), where the *train leaving the station* stands metaphorically for an opportunity not to be missed out by the country. Despite the wide context surrounding the motion verb in (26), the information provided by the topics assigned to the utterance does not lead the system to tag the instance as metaphorical. The metaphorical expression in (24) is also wrongly labelled as literal, but in this case the context of the motion verb is represented by just the sentence including it (if not considering the few not very informative words included in the following sentence), hence making the topic inference on the utterance less reliable.

When dealing with the unsaturated examples shown in Section 7.5, the *global* model correctly identifies the metaphorical value of 2 out of 4 instances. In both (27) and (29) foreign affairs topics are discussed by the podium and in both utterances the metaphoricity of the corresponding verbs – i.e. *turn* and *escape* – is detected by the system. The system fails in (28) probably mislead by the expression *enter office*, very common in the political genre. The same reason can be hypothesised to be the cause of the wrong classification in (30), since *lead* plays a major role in the lexicon being used by the podium in the various topics.

Interesting results are also reported in the evaluation of conventional metaphorical use of motion verbs. In working on idioms such as *turn the corner* and *fall far of (something)*, the system fails since labelling them as metaphorical in (31) and (32). On the contrary, in (33) the *global* model correctly detects as literal both the motion verbs, namely *roll* and *turn*. In these cases, it is hard to understand why the system returns such results. Table 7.3 summarises the performance of the system on the selected utterances.

***Global* model**					
Saturated		Unsaturated		Conventional	
(24)	✗	(27)	✓	(31)	✗
(25)	✓	(28)	✗	(32)	✗
(26)	✗	(29)	✓	$(33)_1$	✓
		(30)	✗	$(33)_2$	✓

Table 7.3 | *Performance of the global model on selected utterances.*

7.5.3 *Glo-cal* model analysis

As the name of the model goes, the *glo-cal* system combines the *global* and *local* approaches. Instances of motion verbs not followed by a direct object are treated *globally*, namely leveraging LDA topics for metaphor detection. Motion verb-direct object pairs are instead *handled* with a *local* approach, using the degree of similarity between verb and object to determine the instance metaphoricity.

Results displayed in Table 7.2 show that adding to the system *local* features actually improves its performance, hence confirming Shutova et al. (2016)'s intuition that linguistic embeddings can capture information of the target and source domains.

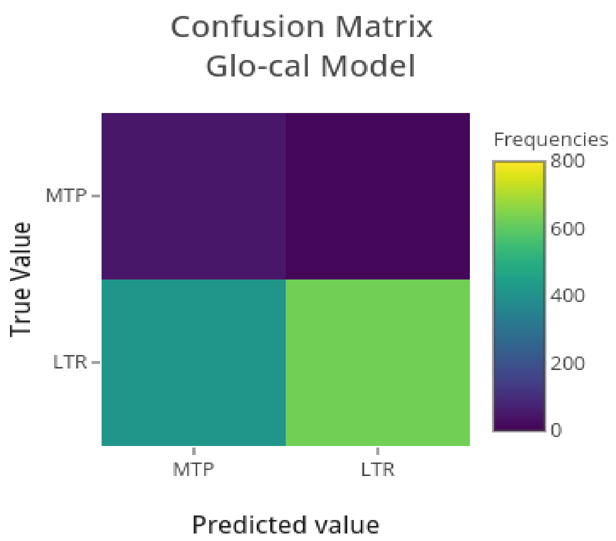

Figure 7.3 | *Confusion matrix plot for the glo-cal model.*

If precision does not change significantly compared with the *global* baseline, it is the recall that reports a substantial improvement (0.69 compared to the *global* 0.64), hence leading to a slightly higher F-score of 0.60. Indeed, taking a closer look at the confusion matrix plot in Figure 7.3[3], the

[3]The interactive graph is available at **https://plot.ly/~fabrex/611/confusion-matrix-glo-cal-model/**

glo-cal model correctly identifies 47 motion verbs as metaphorical.

The best performance is reached in the *local* syntactic context, where 44 are the metaphors detected out of the 51 followed by a direct object. Increasing the instances recognised as metaphorical, the *glo-cal* system inevitably brings down the recall on the literal detection, reaching on the contrary a high precision of 0.97. The model performs better in detecting the literal use of motion verbs not acting as transitively, identifying the 90.83% of the instances correctly, while only the 43.81% of those in the syntactic context under investigation.

Looking at the performance of the system in dealing with saturated utterances, the *glo-cal* model correctly identifies the metaphorical use of each instance in the examples here investigated. In (24), (25) and (26) the three motion verbs present the verb-direct object relation, namely the syntactic construction where the model reaches its highest performance. Metaphorical use is also correctly detected in three of the four saturated utterances. In (27), (28) and (29) the running syntactic common thread is indeed represented by the *local* dependency relation under analysis. Instead, in (30), where *lead* is treated *globally* due to the absence of the direct object, the model fails the correct detection.

***Glo-cal* model**					
Saturated		Unsaturated		Conventional	
(24)	✓	(27)	✓	(31)	✓
(25)	✓	(28)	✓	(32)	✓
(26)	✓	(29)	✓	$(33)_1$	✓
		(30)	✗	$(33)_2$	✓

Table 7.4 | *Performance of the glo-cal model on selected utterances.*

As regards the analysis of the conventional metaphorical use of motion verbs, the *glo-cal* model successfully identifies as literal the four instances present in the reported utterances. The model succeeds in correctly discriminating the two instances followed by a direct object – i.e. *turn* and *roll* in (31) and (33) respectively – and the two verbs in (32) and (33) not in the *local* syntactic construction. Table 7.4 summarises the performance of the system on the selected utterances.

7.5.4 *Glo-cal weighted* model analysis

The *glo-cal weighted* system uses the framework developed in the *glo-cal* model but this time leveraging the output returned from the *global* approach to inform the *local* one. In a nutshell, working on motion verb-direct object pairs, the *glo-cal weighted* model first measures the similarity between the verb instance and the utterance's topic and then uses its output as a weight applied to the *local* procedure for metaphorical detection. The hypothesis to be tested here is that the combination of the document's topic and local syntactic context can help in the successful detection of motion verb use.

Looking at the overall scores in Table 7.2, it can be observed that the performance of the *glo-cal weighted* model slightly improves compared to the *glo-cal* system, attaining an F-score of 0.61. Even if precision and recall turn out to be fixed at 0.54 and 0.69 respectively, attention must be paid to the G_m and G_l scores to fully comprehend the performance of the system.

The *glo-cal weighted* model successfully identifies 38 instances as metaphorical, almost ten less than the total number reached by the *glocal* system. On the sub-metaphorical task, recall is indeed driven down to 0.63 and precision slightly improved, yielding an F-score of 0.21, higher than those of the *global* and *glo-cal* models. This is explained by the considerable reduction in the number of false positives returned by the system. More precisely 262 are the instances labelled as metaphorical by the *glo-cal weighted* model, with the *glo-cal* system reaching the 419 units.

Again, the system attains an higher performance in recognising metaphors in the *local* syntactic context, where 35 are the instances correctly labelled out of the 51 presenting the particular dependency relation. However, this is not a surprise since both the *glocal* and the *glo-cal weighted* model employs the same threshold for the *global* approach. On the contrary, higher performance is reported in the successful detection of literal motion verbs. Indeed, looking at the confusion matrix plot in Figure 7.4[4], G_l scores display a significant jump in terms of recall and (inevitably) F-score in comparison with the *glo-cal* model, however not topping the highest score of the *global* system. In fact, 66.66% of the *local* motion verbs and, again, 90.83% of instances not followed by a direct object are correctly identified.

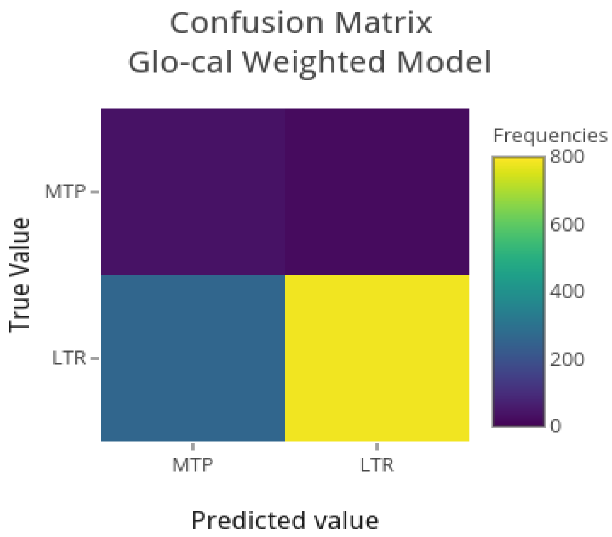

Figure 7.4 | *Confusion matrix plot for the glo-cal weighted model.*

As regards the investigation of saturated utterances, the *glo-cal weighted* model succeeds in detecting the metaphorical value of motion verbs in (24) and (25) but fails the correct identification of *leaving* in (26). Looking at unsaturated utterances, the *glo-cal weighted* models does not improve the performance of the *glo-cal* system. If in (27) and (29) motion verbs are success-

[4]The interactive graph is available at **https://plot.ly/~fabrex/597/confusion-matrix-glo-cal-weighted-model/**

fully tagged as metaphorical, in (30) the instance is still labelled as literal (since *globally* treated), while the model also fails to detect metaphoricity in (28).

Turning to the performance of the conventional metaphorical use of motion verbs reported in (31), (32) and (33), the model successfully labels as literal the four instances included in the utterances, confirming the positive results reached by the *glo-cal* model. Table 7.5 summarises the performance of the system on the selected utterances.

***Glo-cal weighted* model**					
Saturated		Unsaturated		Conventional	
(24)	✓	(27)	✓	(31)	✓
(25)	✓	(28)	✗	(32)	✓
(26)	✗	(29)	✓	$(33)_1$	✓
		(30)	✗	$(33)_2$	✓

Table 7.5 | *Performance of the glo-cal model on selected utterances.*

7.5.5 The hurdles in metaphor recognition

Unlike many of the previous works present in literature, characterised by the use of cherry-picked examples for the testing of the metaphorical recognition system at hand, in this thesis the utterances on which the three developed models are tested represent actual real-world data. Although the selection of podium's documents included in the test set is based on lexical and syntactic criteria, the *chosen* examples represent indeed real-world discourse, continuous texts characterised by a dialogical nature. In fact, each utterance is most of the times a reply to the question posed by the journalists, hence transcriptions of spoken dialogues where the podium shows quick thinking and confidence as well as hesitation and caution. The mood of the podium is revealed not only by their uttered words but also by the non-verbal events they display in conversation, such as the voluntary or involuntary pauses they take. The analysis of these utterances becomes inevitably harder as the syntactic *holes* in structure of a sentence may prove.

Thus, in this section the main issues that negatively influenced the detection of the literal or metaphorical use of motion verbs are discussed.

7.5.5.1 Restricted context

As already discussed in Section 2.4, sometimes the full metaphorical *power* of a linguistic expression can only be ascertained looking at the wide context of appearance. This is the case of saturated utterances, which can be either entirely metaphorical or non-metaphorical. One of the features used by the models to determine the literal or metaphorical usage of the instance under analysis are the topics inferred by the LDA model. For these topics to be reliable, it is necessary for the system to work on documents of large dimension. However, due to the nature of the press briefings, sometimes this is not possible. This is the case of (24), where the *global* approach is not able to correctly detect the use of the motion verb, unlikely the *local*-based ones. Thus, the absence of a verb-direct object pairs and the narrow context of appearance of the instance under analysis may seriously influence the successful outcome of the recognition.

7.5.5.2 Ambiguous pronouns

Unlike the *syntax-agnostic global* approach, in verb-direct object pairs pronouns play a major role for the correct identification of the literal or metaphorical use of the motion verb. Let us take as way of example the following sentence:

(34) And with that, I will **turn it** over to Secretary Johnson to begin.

In (34), we are not able to understand what the podium is turning over to Secretary Johnson, hence not even able to identify the use of *turn* as either metaphorical or literal. The only way to fully comprehend the utterance itself is to go back to the preceding sentences and find what the pronoun *it* is making reference to. We would then figure out that what the podium was turning over to the Secretary Johnson was actually a call, as the preceding sentence of (34) describes:

(35) We'll do this call on the record, but it will be embargoed until the conclusion, so we ask that you please not use this material until the call concludes.

However, due to the linguistic characteristics of the U.S. Press Briefings, the application of a well-performing anaphora resolution would have probably represented a separate task on its own. Furthermore, being the briefings actually a series of question-answer between journalists and podium, it is sometimes impossible to retrieve the reference of the pronoun in the podium's utterance without going back to the question posed by the journalists themselves.

7.5.5.3 Parser issues

As the *local* approach focuses on motion verbs in verb-direct object relations, the *glo-cal* and *glo-cal weighted* models are highly dependent on the accuracy reached by the parser. In this case, the main hurdles are represented by:

1. anomalous syntactic constructions;

2. the use of an informal register between speakers.

The first point is often caused by non-verbal events (e.g. pauses, laughter, etc.) and sudden disruptions in discourse that inevitably leads to *miss* the correct identification of the dependency relations of the instance under analysis. Point 2 usually leads to a wrongly POS-tag of the motion verb, due to the use of a lexicon not present in the parser database.

7.6 Summing up the results

In the previous sections, the models developed for the task of the metaphor recognition of motion verbs have been discussed. The baseline of evaluation is represented by the *global* model, since taking into account a smaller number of features compared with the other two systems. As it can be observed in Table 7.2, both the *glo-cal* and the *glo-cal weighted* models outperform the baseline system. In particular, the *glo-cal weighted* system attains the highest F-score of 0.61, which makes it the best-performing among the three strategies presented here.

The thread connecting the three models is represented by the presence of a modest high recall and a low precision, inevitably determined by the unbalanced test set. As regards the latter, the main reason is to be found in the very restricted number of novel metaphors in the test set and by the too high percentage of metaphor false positives returned by the systems. On the contrary, precision in the detection of literal use of motion verbs is reported in high scores that do not differ significantly between one model and the other.

The analysis of recall score has revealed different trends among the systems. Although the *glo-cal* and *glo-cal weighted* models attain similar overall scores, they do *behave* differently. Indeed, the *glo-cal* system outperforms the other models in terms of G_m recall since retrieving the largest number of metaphors. On the contrary, the *glo-cal weighted* and *global* models tend to favour recall in literal detection, with the latter yielding the highest number of true positives. However, it is the *glo-cal weighted* system to attain the highest F-score, hence showing more balance in metaphor detection.

As regards the particular syntactic constructions under investigation in this work, it can be observed that focusing specifically on verb-direct object pairs actually improves the performance of the models in the detection of the metaphorical instances. Indeed, the *glo-cal* model successfully identifies 44 metaphorical motion verbs showing the particular syntactic construction. The result is more modest in the *glo-cal weighted* system (35 out of the 38 correctly detected), while it decreases substantially when the syntactic construction is not taken into account, as in the *global* model (28 out of the 32 successfully recognised). However, it is interesting to note that looking at the overall accuracy in the detection of correct *local* instances, it is the *global* model to return the highest score, as shown in Figure 7.5[5]. In the *glo-cal* model a significant fall is indeed observed, while in the *glo-cal weighted* system the *local* accuracy decrease is less important. These systems evidently return more false positives among the instances followed by a direct object, therefore suffering reduction in terms of the *local* syntactic accuracy.

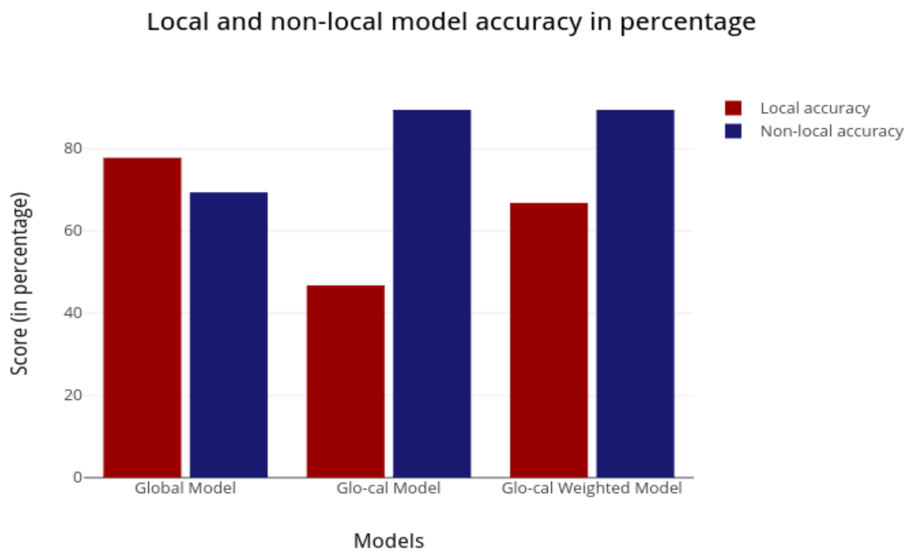

Figure 7.5 | *Local and non-local accuracy performance across models.*

Metaphorical motion verbs *a-syntactically* treated represent indeed the

[5]The interactive graph is available at **https://plot.ly/~fabrex/615/local-and-non-local-model-accuracy-in-percentage/**

Achilles' heel of the three systems, even though data are too exiguous to draw definitive conclusions. The *global* model succeeds in finding 4 of the 9 instances not acting as transitive verbs, while the others do worse. In fact, some of the annotated metaphors always escape the systems, as in the case of (36) where the similarity between verbs and word topics only yields a literal reading:

(36) [...] Last quarter, confidence among CEOs of U.S.-based companies **jumped** by 4.2 percent points in the YPO Global Pulse Survey, one of the single-largest quarter gains in history. [...]

Even if representing a restricted number, the investigation of saturated and unsaturated utterances seems to confirm the importance of necessarily taking into account the syntactic structure of the motion verb for its successful metaphorical detection. The *glo-cal* model indeed only fails the recognition of (30), where it is not possible for the parser to retrieve the object of the motion verb *lead*. Results of the analysis of conventional utterances in the two models outperforming the baseline show instead that a successful detection is achieved regardless of the specific syntactic structure. However, one must be careful in the interpretation of these results, as the number of examples is too exiguous and they were randomly selected. Furthermore, it is in the recognition of instances acting as transitive verbs that both the models suffer a low accuracy, due to the threshold set for the identification of metaphorical motion verbs.

Thus, summing up the above-described results, the following conclusions can be drawn. As regards the use of LDA topics alone as features of the system, it can be stated that the performance on the recognition of metaphorical instances is poor while returning a very large number of correct literal motion verbs. Adding verb-direct object relations to the features of the system improves the performance of the model, with a significant increase in the number of successful metaphorical detections. The *glo-cal* model is the best performing system in retrieving the highest number of metaphors, largely penalising the identification of literal motion verbs though. In terms of F-score, it is the *glo-cal weighted* model to represent the best performing system. The weights applied on the basis of the information provided by the *global* approach improves slightly the performance of the model if compared to the *glo-cal* strategy, leading indeed to a more balanced model. Although the number of metaphors decreases of almost ten units (*enter* in (28) is labelled literal just like the *global* model does), the *glo-cal weighted* model identifies a larger number of correct literal instances in return, hence able to distinguish the use of more conventional metaphors. Thus, information coming from LDA can slightly improve the performance of the system towards more balanced results. However, as it does not provide any progress in the identification of metaphors, the hypothesis presented in Section 7.5.4 must be refuted.

7.7 Related unsupervised metaphor processing systems performance

The fragmented picture of the field of research, the lack of shared datasets and the different experimental settings are among the several reasons that make the results of different metaphor processing systems not directly comparable. In the case of the present thesis, the specific genre taken into account, the choice of investigating a pre-definite sub-class of lexical items and the dialogical nature of the continuous texts make the things more complicated.

Even the largest publicly available resource annotated for metaphoricity – i.e. the VUAMC corpus – could not *fit* the experimental settings of the three systems proposed in this work. The main issue is represented by the large context *needed* by the LDA model for a reliable inference of topic on new unseen documents. Indeed, 362 is the total number of documents including motion verbs in the VUAMC. In this case, the average length of the utterance is of 27.46 words (punctuation excluded), hence a too small context for reliable topic representations. Furthermore, it must be borne in mind that the in the VUAMC the metaphorical procedure of annotation does not discriminate between conventional and unconventional metaphorical expressions.

Thus, being conscious that "[t]he linguistic properties which can distinguish metaphors in one genre may not apply to other genres" and that "[e]ach of the systems is based on a different theory of metaphor-in-language" (Dunn, 2013c, p.9), here I use the results of related studies employing unsupervised techniques for the identification of metaphors to indicatively draw a comparison that could shed more light on the performance of systems proposed in this thesis. Furthermore, due to the highly unbalanced test set used for evaluation in this thesis, results must be discussed with due care.

The works chosen for comparison are Shutova et al. (2016)[6], Shutova and Sun (2013) and Shutova et al. (2010) (cf. Section 3.5.3).

System	P	R	F1
WORDCOS	0.67	0.76	0.71
HGFC	0.65	–	–
CLUSTERING	0.79	–	–
GLO-CAL WEIGHTED	0.54	0.69	0.61

Table 7.6 | *Comparison of unsupervised metaphor processing systems.*

Table 7.6 describes the performance of each model as reported in their corresponding studies. WORDCOS stands for the work of Shutova et al. (2016), HGFC for the study of Shutova and Sun (2013), CLUSTERING for the system of Shutova et al. (2010).

[6]Results of the evaluation on the Mohammad et al. (2016) dataset are reported since consisting of verb-noun pairs.

It should be remembered that Shutova et al. (2016) work on cherry-picked examples extracted on the basis of precise syntactic relations. In the study of Shutova and Sun (2013) the system searches for metaphorical expressions in the BNC – hence working on unrestricted texts – evaluated against human judgements on the randomised selection of instances tagged as metaphorical from the unsupervised system and other two baselines. In the work of Shutova et al. (2010) unsupervised noun and verb clustering techniques are used for the identification verb-subject and verb-object metaphorical constructions in unrestricted text. As for the systems developed in this thesis, performance of the *glo-cal weighted* models is reported since the best performing in terms of F-score.

As previously stated, the comparison proposed here is only indicative due to the huge differences in the experimental settings. However, the results seem to show that the performance of *glo-cal weighted* model is still far from the those attained using unsupervised techniques in the field of the computational modelling of metaphors. Adopting the same intuition as Shutova et al. (2016) on using word embeddings similarity as proxy of conceptual domains, the *glo-cal weighted* model yields favourable results however not reaching a high rate of performance as the WORDCOS system proposed by the authors. A significant gap in performance is observed in the precision of the CLUSTERING model, pointing out the weakness of the *glo-cal weighted* model. This result seems to be also confirmed by HGFC, where the system attains a fair score in terms of precision, reaching the 0.65.

PART IV

Conclusions

CHAPTER 8

Summary and Conclusions

Metaphors are a fascinating product of human language. Their ubiquity in every form and in any mode of communication makes them a very productive phenomenon. Metaphors can be used to delight the reader and/or the speaker, they can enrich our communication but they can be also used as a tool of manipulation of the reality we live in. However, metaphors represent also a very complex phenomenon when it comes to its computational understanding. Several and different linguistic layers are involved in its detection and/or interpretation and this process is far for being defined as an easy task.

The present thesis has covered the topic of the computational modelling of metaphors. More precisely, in this work the automatic identification of metaphors in continuous, dialogical and naturally-occurring political data has been dealt with, proposing three algorithms that do not rely on task-specific hand-crafted resources and try to minimise as far as possible the need of labelled data. The focus of metaphorical investigation has been represented by a selection of verbs drawn from Levin's semantic class n°51 of verbs of motion. This choice was motivated by recent findings in cognitive-pragmatic studies in the realm of political discourse, unveiling the role played by these lexical items in the communication strategies deployed by public speakers. Employing an ESP corpus focusing on the genre of the atypical political speech of U.S. press briefings, the three developed models were tested on a random sample of utterances – coming from U.S. White House Press Secretaries or administration personnel discourse – where motion verbs were annotated for novel metaphoricity by human judges.

In Part I, I started off with the presentation of the theoretical background at the basis of the present thesis, also introducing the lexical items under metaphorical analysis and the theory motivating their choice. In Chapter 2, particular attention was dedicated to the influence of the linguistic structure in the reading and interpretation of metaphors. A binary distinction was indeed delineated by Dunn (2013b) based on conceptual metaphorical patterns – i.e. the presence of both target and source concepts in an utterance or of only elements from the target domain – to be later used for a qualitative analysis of systems' performance. In Chapter 3, an overview of the unsupervised methodologies used in this thesis for the metaphor recognition of the verbs of motion is provided. The field of research of the computational modelling of metaphors is then introduced, discussing its main features and the related works, with a particular focus on those studies leveraging the potential of the unsupervised techniques previously presented.

In Part II, the political data used for the metaphor recognition task are presented. Chapter 4 focuses on the description of the building of the CompWHoB corpus, its NLP features and the qualitative aspects of the collected in this resource.

One of the major contribution of the present thesis is indeed represented by the computational development of the pre-existing WHoB corpus – kindly donated to the candidate by Prof. Marco Venuti – since being the initial core of this work. The main specific characteristics of the corpus, in particular its genre and the related linguistic features have been taken into account before defining the aims of research. The CompWHoB corpus represents an important and necessary upgrade of the WHoB corpus under different perspectives. In terms of coverage, the time-span is not only widely extended, but it is constantly updated with the most recent press briefings published on the APP website. Under a computational linguistic perspective, the corpus is now equipped with a NLP pipeline including the main steps of linguistic pre-processing, in this way allowing to be independent from external tools for its (computational) linguistic investigation. Furthermore, under a more IT aspect, the structural annotation is now not only fully automatised but, extracting information directly from huge databases such as DBpedia and Wikidata, it is now possible to collect essential social characteristics about the identity of the institutional speakers in the briefing. All these aspects make the CompWHoB corpus an important resource for several research fields, ranging from social and political sciences to computational linguistics. Finally, to the best of my knowledge, this is the only corpus publicly available on request dealing with the atypical political genre of the U.S. press briefings.

Going back to Part II, in Chapter 5 the lexical items under metaphorical investigation – i.e. the verbs of motion – were quantitatively analysed and the criteria of their selection for the task were described. The results of an introductory task of annotation were used to explore the literalness/metaphoricity of motion verbs based on the level of metaphorical analysis set in the task. Finally, the novel procedure for the annotation of the unconventional metaphorical usage of verbs of motion in the corpus was introduced, which results were used as test set for the evaluation of the proposed algorithms.

In Part III, the three algorithms for the metaphor recognition of the selected verbs of motion in the context of U.S. press briefings were presented and evaluated. In Chapter 6, the intuition and the motivations at the basis of the approaches supporting the development of the three systems for metaphor identification were discussed. After describing in detail the three algorithms, in Chapter 7 the performance of the three models were presented and discussed at length. On the basis of these results, I am now ready to answer the three research questions.

Let us begin with the first one:

RQ1 Can we move towards a *syntax-agnostic* approach when dealing with the automatic recognition of metaphors employing broadly-based distributional semantics techniques such as word embeddings and topic models? More precisely:

(i) How does the use of these two unsupervised techniques perform on lexical items when no syntactic information is taken into consideration?

(ii) Does the information coming from topic modelling improve the performance of the system when syntactic information is taken into account?

The intuition at the basis of the *global* approach is that the motion verb being metaphorically used should be inconsistent with the utterance's topic(s) in which it is included, hence showing a low semantic relation with it. A clear answer to the Question (i) of RQ1 has been provided in Section 7.6. The model only based on the *global* approach yields indeed the lowest score among the three algorithms proposed in this thesis. The correct number of metaphors recognised by the model is barely above the half of the total number of instances found in the human-annotated gold-standard test set. Furthermore, the *global* approach represents the Achilles' heel also of the *glo-cal* and *glo-cal weighted* systems, as discussed in Section 7.6.

Turning to Question (ii) of RQ1, we are focusing in this case specifically on the performance of the *glo-cal weighted* model. Results show that using topic modelling's output to *inform* the *local* approach does not improve the performance of the model on the recognition of metaphoricity of motion verbs. However, the joint use of LDA and word embeddings leads to a more balanced model and brings an overall improvement in terms of F-score, hence defining the *glo-cal weighted* algorithm as the best performing one.

The reasons of the *syntax-agnostic* approach low performance in the recognition of metaphors may be hypothesised to come down to one or more of the following points:

- inaccuracy of the LDA inference process on too short documents;
- word distribution over the topic not sufficiently representative of the semantic content of the topic itself;
- *anonymous* semantic value of the lexical item under investigation when compared to the topic's word distribution, i.e. the motion verb is not distinctive enough in that semantic context (e.g. due to its high occurrence and its versatility across the corpus being too general).

Thus, the answer to both (i) and (ii) of this first research question must be considered negative at the moment. Although further research is needed to confirm this response, a *syntax-agnostic* approach combining LDA and word embeddings as implemented in the present thesis is still far from reaching significant results.

Let us move on to the second research question:

RQ2 On the basis of Dunn (2013b)' claim that the linguistic structure of an utterance influences its metaphorical reading leading to a binary distinction between *saturated*[1] and *unsaturated*[2] utterances:

(i) Can it be observed any influence of the particular linguistic structure in the final performance of the metaphor recognition system?

(ii) Does the *global* approach help improve the performance of the algorithm in the detection of metaphors in saturated utterances?

[1]Either an entirely metaphorical or non-metaphorical reading is possible.
[2]Only a metaphorical reading is possible.

The small number of instances analysed in Section 7 and the restricted number of saturated utterances does not allow to draw definitive conclusions from this second research question. However, if taken with the due caution, results can still represent indicative pointers of metaphorical investigation.

Addressing Question (i) of RQ2, the qualitative analysis of systems' performance on the 7 selected examples in Chapter 7 reveals that utterances including (metaphorical) verb-direct objects relations are more likely to be correctly tagged by the system if the particular syntactic structure is taken into consideration. The *glo-cal* algorithm – the best performing in the detection of metaphorically used verbs of motion – indeed only fails in recognising the metaphorical value of the verb not followed by a direct object. However, I reiterate that these conclusions can only be considered tentative at the moment.

This answer leads in turn to Question (ii) of RQ2, being also strictly connected to the RQ1. Only using a *global* approach does not seem to provide the algorithm with the information necessary for the correct recognition of metaphorical motion verbs in saturated utterances, where saturation is caused by *taken-for-granted* background knowledge and lexical ambiguity. This is evident in the results of the *global* model, as discussed in Section 7.5.2. Indeed, the system does fail the correct detection of the utterance (24) – basically consisting of just the sentence including the verb of motion – where context information is reduced to the minimum. The reasons for this *behaviour* must be found in the above-stated ones of RQ1.

Thus, going back to (i) of RQ2, the analysis carried out in Chapter 7 tends to indicate that, in the context of the present work, the difference between saturated and unsaturated utterances does not play a significant role. On the contrary, it is the focus on the syntactic structure to influence the final performance of the system.

Finally, let us turn to third and last research question:

RQ3 Avoiding the recourse to any task-specific hand-coded knowledge and labelled data, does the joint use of word embeddings and topic modelling compare favourably with metaphor processing systems based on unsupervised approaches present in literature?

Due to the extremely unbalanced test set used in the evaluation stage, results of the three models proposed in this thesis have been *handled* with due care. The answer to this research question has been addressed in Section 7.7, showing that only a negative answer can be provided. Even though taking into consideration the best performing system – i.e. the *glo-cal weighted* model – and the role played by the different experimental settings in other studies, results indicate the performance attained in this thesis are still far from those reached in the main relevant works employing unsupervised techniques for metaphor recognition. Some of the reasons that can explain these modest results must be found in the answers to the previous research questions. However, it should not be forgotten the role played by the conversational features of the utterances analysed in this work, and the resulting hurdles for computational investigation.

Thus, summing up, in the present thesis a novel approach combining broadly-based distributional semantics unsupervised techniques for the automatic recognition of metaphors in text has been proposed. This work has focused on the identification of novel metaphorical use of a selection of verbs of motion in atypical political conversational genre of White House press briefings. To the best of my knowledge, this is the first study to explicitly focus on the automatic recognition of metaphors in naturally-occurring, continuous and dialogical texts. Although biased by a highly unbalanced test set, models' performance shows that the a *syntax-agnostic* approach is still far from being accomplished combining topic models and word embeddings as implemented in this thesis. Even though performance can be only considered modest, promising results can be observed in the *local* approach, highlighting the importance of leveraging the syntactic context for metaphor recognition and at the same time representing an important stimulus for improvement on this path of research.

Thus, the present thesis shows that there is plenty of room for progress. One of the future directions is undoubtedly represented by the enhancement of the topic modelling process on conversational texts. Topic models are not to be ditched in the exploration of metaphors, but they should represent only one of the multiple features of a computational modelling system of metaphors. In order to achieve better results, a future step is represented by a more in-depth linguistic pre-processing of each utterance due to the characteristics of the dialogical texts. Anaphora resolution seems indeed necessary for focusing on a larger number of syntactic relations. Another future direction is to take into account the gradient nature of metaphors. Indeed, as we have seen, human judges identified only 60 novel metaphorical usages of motion verbs in more than 1000 podium's utterances. This figures lead to highlight the importance of taking into account different metaphorical levels, without marking their membership in only two particular classes. Thus, a future line of investigation is represented by the enhancement of word embeddings techniques as tools for the gradient representation of metaphoricity, as their promising results have shown.

Finally, since this thesis also contributes to the linguistic resources in NLP, as future project a new release of the CompWHoB corpus is scheduled, equipped with a richer NLP pipeline also including NER and coreference resolution steps. Furthermore, the plan is to make the corpus available in the near future on a web platform where a dedicated interface will allow to easily navigate the resource.

Appendices

APPENDIX A

Annotation Guidelines

A.1 Introduction

In this document the tasks you are kindly asked to carry out are explained and shown to you. More specifically, you will be dealing with the annotation of the metaphorical or literal use of motion verbs (do not worry, everything will be clear at the end of the document. Or I hope so). As introductory step, a very brief overview of the CompWHoB corpus is provided to you in order to better understand its content and structure.

A.1.1 The Corpus: Press Briefings and the CompWHoB Corpus

CompWHoB is an acronym that stands for Computational White House Press Briefings (corpus). This corpus is a diachronic collection of the transcriptions of the United States of America Press Briefings, the press conferences held by the White House Press Secretaries (and other administration personnel) for the national and international news media. The briefings are extracted from the American Presidency Project website `www.presidency.ucsb.edu`, where the Press Briefings document archive section can be freely consulted. The CompWHoB corpus covers a period time of twenty-four years and seven presidencies, from the first term of William J. Clinton till the first term of Donald J. Trump.

The U.S Press Briefings are considered as part of a wider category of political press conferences as the recipient is not only represented by the press present at the scene, but also by the audience at home. Moreover, being these briefings daily held by U.S. administration, both the speakers - i.e. the press secretaries and the journalists - know each other very well. This means that sometimes the topic may diverge from the institutional ones, with both speakers even joking about their private life.

The topics related to the U.S. governance discussed during the briefings have been summed up in the following macro-categories:

- Crime & Justice
- Culture & Education
- Economy & Welfare
- Foreign Affairs
- Greetings
- Health
- Internal Politics
- Legislation & Reforms
- Military & Defense
- President Updates
- Presidential News
- Press Issues

As said before, sometimes the topic discussed by journalists and administration cannot be included in some official categories and you may encounter utterances where the speakers talk about everything but politics.

A.1.2 The CompWHoB Structure

The data extracted from the American Presidency Project website are collected and encoded in an XML format, a markup language that allows documents to be easily readable for both humans and machines alike. Each briefing consists of a series of question-answer between journalists and the U.S administration personnel. For the sake of clarity, in the corpus the U.S. administration personnel is always referred to as *podium* while the news media members are referred to as *journalists*. As the original transcripts contained also meta-textual information enclosed in brackets about audience reactions and speech events descriptions (e.g. (Laughter), (Applause), etc.), self-closing tags are used to identify these events in the corpus (e.g. in the case of a pause in either podium or journalists utterances, you will encounter the following tag: `<pause/>`).

A.2 Annotation Tasks

In this section the annotation tasks you are kindly asked to perform are described (by the way, thank you very much in advance!). As you will deal with the annotation of the metaphorical or literal usage of motion verbs, in Section A.2.1 I first illustrate the concept of metaphor. In Section A.2.2 an overview of the lexical items you will be working on is provided. Finally, in Section A.2.3.2 and A.2.3.3 the annotation procedures are described.

A.2.1 Metaphors

According to the scientific literature, a metaphor is perceived in language when there is an association between distinct and seemingly unrelated concepts. Let's take a political domain example extracted from a scientific paper (Shutova, 2015):

(37) The President Obama is *rebuilding* the campaign *machinery* that *vaulted* him into office" (New York Times, 2011)

In the example (37), we can see that there is a metaphor in use. The political system is indeed viewed as a machinery, and terms are drawn from this specific domain (conventionally used with physical targets). In this example, we define the sentence as whole as a *conceptual metaphor*, since we can see that there is a mapping between two different domains of experience. More in detail, in (37) we can say that the verbs *rebuild* and *vault*, and the noun *machinery*, are used metaphorically as not making reference to concrete objects.

Recognising a metaphor is not an easy task though. Some metaphors are so rooted in our use of language that we hardly recognise them as such. Let's take a look at (38):

(38) A Metropolitan Police officer used the force's computer systems and colluded with a Croydon Council worker *to dig up* information to use in a dispute over a dodgy second-hand car, a court has heard. (Croydonguardian, 2016)

In everyday language, the use of the phrasal verb *to dig up* plus an abstract concept such as *information* has become widespread, making its metaphor detection not trivial at all. In this case, the use of the verb *to dig up* is metaphorical as the journalist refers to the information as a 'ground' material that can be brought to light. Nonetheless, as this use is becoming conventionalised over time, we consider it as a *conventional* metaphor. What this definition implies is that even a native speaker may overlook the metaphorical charge of this verb phrase. Indeed, one of the first clues that help humans in recognising a metaphor as such is that they appear to be expressions out of context.

Two excerpts from the CompWHoB corpus can help us in tracing a dividing line between the literal and metaphorical aspects that need to be taken into consideration in this task:

(39) I mean, we might or, we might start saying to the American people, you want to know how much money we're spending to *chase down these stray bits of information* because we've got some overzealous staff people up there who are trying to keep us from doing our work.

(40) May of this year: "I've talked about the idea of having a different force posture that would enable us to be there to help the Iraqis in a variety of ways, protect the border, *chase down al Qaeda*, embed and train the troops, provide security, psychological security of helping this new government."

In (39) the verb *chase down* is used metaphorically because the direct object of the verb - i.e. *these stray bits of information* - is treated as a physical object despite being an abstract entity, hence mapping together two different domains of experience. Nonetheless, again there is not a clear distinction between literal and metaphorical use.

In (40) the verb *chase down* is used literally. Since we usually chase down people and/or objects, *al Qaeda* complies with the linguistic conditions required by the verb itself. Even if in this case another figure of speech is present, more precisely the metonymy *al Qaeda*, you are not asked to identify them as well.

To sum up, to best detect the metaphor use of a word in this task, the definition provided by the online dictionary Merriam Webster[1] comes in handy:

"[Metaphor]: a figure of speech in which a word or phrase literally denoting one kind of object or idea is used in place of another to suggest a likeness or analogy between them (as in *drowning in money*)."

[1]`https://www.merriam-webster.com`

A.2.2 Your Data: Motion Verbs Utterances

The data you will be working on are a random sample of the podium utterances extracted from the CompWHoB corpus. Questions coming from the journalists are not included in these data. The utterances are selected according to a common criterion: they all contain one or more *motion verbs*. These verbs were selected according to the taxonomy provided by Levin (1993) in her work *English verb classes and alternations: A preliminary investigation*. As their pretty much self-explanatory definition states, these verbs are characterised for encoding in their meaning a specific kind of motion as in *"[...] they **fly out** of neighboring countries"* but also in *"[...] the President might **enter** in the fray himself"*.

Each verb is enclosed in an XML tag that allows the annotation to be performed by the user. The XML format chosen is the following:

```
<motverb usage=" "> motion verb </motverb>
```

A.2.3 Annotating the Data

A.2.3.1 What your data look like

As previously said, every utterance will contain one or more motion verbs. Then, in your data a sentence like (39) will look like as follows:

(41) I mean, we might or, we might start saying to the American people, you want to know how much money we're spending to `<motverb usage=" ">` chase `<motverb>` down these stray bits of information because we've got some overzealous staff people up there who are trying to keep us from doing our work.

As you can see, in this case only the verb itself is enclosed in the XML tag but not the particle accompanying it and forming the phrasal verb. This is due to a step in the linguistic pre-processing of the text that allows to recognise the verb but not its phrasal-verb value if present.

The utterances you will be dealing with may vary in length. They are actually paragraphs most of the time consisting of more than two sentences. The choice of extracting entire paragraphs is necessary for the understanding of the potential metaphorical charge of a verb. Indeed, some metaphors may not be detected until their context is inspected. To make it clear, Jonathan Dunn's examples (Dunn, 2013b) come to our aid:

(42) Mary demolished John's stronghold with her newly found evidence.

(43) Mary demolished John's stronghold with her newly found weapon.

If we can be sure that (42) is a metaphorical utterance, the same cannot be said of (43). In this case our reading is uncertain because we do not know what the context of the sentence is. Maybe Mary just bought a bazooka that destroyed John's castle. This is why the importance of the surrounding context will be emphasised in the following section.

A.2.3.2 Task 1: Annotation Procedure

As said in Section A.2.3.1, some paragraphs may increase the degree of certainty of the metaphoricity or literal judgement by inspecting the surrounding context. However, being these paragraphs most of the times answers to journalists' questions, they may be full of anaphoras that make it hard to spot the subject of the whole paragraph (and consequently also the detection of the metaphorical or literal use). Moreover, some utterances may consist of single sentences that will make the judgement even harder.

Bearing in mind the nature of the data just described, I illustrate here the procedure to annotate the metaphorical or literal charge of motion verbs:

1. Carefully read the whole utterance to understand the general meaning of the content.

2. Carefully read the context of the verb under analysis. If the utterance consists of more than just the sentence containing the verb itself, compare its content to the previous and following sentences.

3. Determine the meaning of the verb in the particular context in which it appears by paying attention to its close surroundings. If the words around the verb do not provide the necessary information, look at the wider context.

4. Determine the basic meaning of the verb. In this task, consider the basic meaning of the verb as the one that tends to be more concrete and tangible. Basic meaning can be often considered as the one closest to its etymological one and it is usually listed as the first entry in a dictionary (Beware! Some dictionaries use the highest occurrence collected in a corpus. It does not necessarily entail that the highest occurrence is the literal meaning of a word).

5. Compare now the basic meaning of the verb with its contextual meaning, i.e. the meaning of the verb in the particular context you are analysing. If its contextual meaning contrasts with the basic meaning of the verb but it can be understood in comparison with it (*original* properties of the verb emerge in the interpretation of its meaning), mark it as metaphorical, otherwise as literal. An example is provided here in order to make the procedure more clear.

   e.g. "Don't try to **twist** what I said into something else." Contextual meaning: In this context, 'twist' indicates to repeat someone's *original* message/words drastically changing the *original* meaning of what been said, often in a negative way.

   - Contextual meaning: In this context, 'twist' indicates to repeat someone's *original* message/words drastically changing the *original* meaning of what been said, often in a negative way.
   - Basic meaning: The basic meaning of 'twist' is to form something into a particular shape, often a distorted one.

- Contextual vs. basic meaning: The contextual meaning can be understood in the terms (better said, semantic aspects) of the basic meaning. We can understand the change of the *original* meaning of someone's message via the physical distortion conveyed by the verb 'twist'.

6. After having performed the previous steps, express your judgement:

   a) If in your opinion the verb is used metaphorically, insert '**MTP**' in the 'usage' field between double quotes.
   e.g. `<motverb usage="MTP">...</motverb>`

   b) If in your opinion the verb is used literally, insert '**LTR**' in the 'usage' field between double quotes.
   e.g. `<motverb usage="LTR">...</motverb>`

   c) If it is not possible to identify the reading as either metaphorical or literal, insert '**UNK**' in the *usage* field between double quotes.
   e.g. `<motverb usage="UNK">...</motverb>`

If you desire, you can consult any dictionary you prefer to ascertain the etymology of the verb, its primary literal meaning (if there is one) and its syntactic use and semantic requirements. Below I provide a list of online dictionaries that may come in handy:

- Longman Dictionary:
  `http://www.ldoceonline.com`

- Merriam-Webster Dictionary:
  `https://www.merriam-webster.com`

- Merriam-Webster Learners Dictionary:
  `http://www.learnersdictionary.com`

- Oxford Dictionaries:
  `https://en.oxforddictionaries.com`

Once the annotation process is complete, you can send your annotated file to the following email: `fabrizio.esposito3@unina.it`. In order to safeguard your privacy, please save the file under this text format `compwhob_motverbs_annotated_nutt.xml` and replace the string of text `nutt` with the number of utterances you annotated.

A word of caution. In reading your utterance, you may encounter some verbs that were mistakenly tagged as such by the system. For instance, in the following example, the noun *follow-up* is incorrectly tagged as a verb:

e.g. The President will speak only after a `<motverb usage=" ">` follow-up `</motverb>`

If you bump into one of them, I kindly ask you to notify us by inserting the label '**WVB**' in the `usage` field between double quotes.

e.g. `<motverb usage="WVB">...</motverb>`

Furthermore, the data you will be provided with will be presented as single tokens, separating each part-of-speech element from the other (punctuation included). Each one is separated from the other with a front- and end-whitespace. You may also encounter double whitespaces in your data. These are remainings of the meta-textual XML tags. They can be considered as a normal whitespace (no implicit meaning then).

A.2.3.3 Task 2: Annotation Procedure

In this second annotation task you are kindly asked to identify the novel (or unconventional) metaphorical usage of motion verbs as opposed to their literal or conventional metaphorical use. As the terminology might be a little bit confusing, let me first explain the difference between novel and conventional metaphors.

Try to figure out the *life* of a metaphor as a trajectory: metaphors are born *in nature* as novel (hence, just the way they are). We intuitively and readily recognise them either in speech or in text due to their metaphorical *charge*, as in a certain way they *break* our usual logical reasoning. Indeed, we are not able to interpret metaphors automatically as they are highly incongruous and they are not systematically used in our language. They associate two seemingly unrelated domains of experience, making us perceive them as *novel*. For example, in *She was warming my winter days*, we instantly realise the association between the unusual comparison, understanding the metaphorical nature of the expression. The target term *She* is used in this case to refer to something else, to the source topic HEAT/SUN.

However, when a metaphor starts to be used intensively and for a long time by its speakers, its meaning becomes so common that it *loses* its original aspect of novelty. Indeed, we are able to understand it almost automatically, sometimes without even realise that we are actually dealing with a metaphorical expression. Let us take as way of example the following headline from a well-known newspaper webpage: "Time is running out for Madagascar - evolution's last, and greatest, laboratory"[2]. The metaphorical expression *time is running out* is so entrenched in everyday use of English language that some native speakers (but not only them) may find hard to recall its *original* metaphorical nature and its corresponding hidden concept TIME IS AS A LIMITED RESOURCE.

Thus, in this task you are only asked to identify as *metaphorical* those instances of motion verbs that correspond to the description of novel metaphors above-provided. As opposed to the novel metaphors, under the umbrella term *literal*, the instances of motion verbs being used either literally (i.e. using their basic meaning) or in a metaphorical conventional way are included.

I illustrate here the procedure for the annotation of the metaphorical or literal use of motion verbs. Remember that a metaphorical expression is in place if a linguistic expression is used to refer to something else, comparing two different domains of experience/knowledge.

1. Carefully read the whole utterance to understand the general meaning of the content.

[2]https://www.theguardian.com/world/2017/may/13/madagascar-mass-extinction-plants-kew-gardens

2. Carefully read the context of the verb under analysis. If the utterance consists of more than just the sentence containing the verb itself, compare its content to the previous and following sentences.

3. Try to understand the topic of the utterance. The motion verb is likely to be used metaphorically (conventionally or unconventionally) if it is in discordance with the topic of discussion.

4. Determine the meaning of the verb in the particular context in which it appears by paying attention to its close surroundings. If the words around the verb do not provide the necessary information, look at the wider context.

5. If the basic meaning of the verb is in contrast with its contextual meaning, the verb is likely to be used metaphorically.

6. If in your opinion the verb is used metaphorically, determine if it is novel or conventional. Two are the main clues for the recognition of a metaphor as novel:

   - its meaning is not fixed (as in conventional metaphors) but rather *idiosyncratically* produced.
   - you may feel relatively unfamiliar with the metaphorical expression, hence requiring a certain effort to understand it.

7. After having performed the previous steps, express your judgement:

   a) If in your opinion the verb is used metaphorically, insert '**MTP**' in the 'usage' field between double quotes.
   e.g. `<motverb usage="MTP">...</motverb>`

   b) If in your opinion the verb is used literally, insert '**LTR**' in the 'usage' field between double quotes.
   e.g. `<motverb usage="LTR">...</motverb>`

c) If it is not possible to identify the reading as either metaphorical or literal, insert '**UNK**' in the *usage* field between double quotes.
e.g. `<motverb usage="UNK">...</motverb>`

Once the annotation process is complete, you can send your annotated file to the following email: `fabrizio.esposito3@unina.it`. In order to safeguard your privacy, please save the file under this text format `compwhob_motverbs_annotated_nutt.xml` and replace the string of text `nutt` with the number of utterances you annotated.

A word of caution. In reading your utterance, you may encounter some verbs that were mistakenly tagged as such by our system. For instance, in the following example, the noun *follow-up* is incorrectly tagged as a verb:

e.g. The President will speak only after a `<motverb usage=" ">` follow-up `</motverb>`

If you bump into one of them, I kindly ask you to notify us by inserting the label '**WVB**' in the `usage` field between double quotes.

e.g. `<motverb usage="WVB">...</motverb>`

Furthermore, the data you will be provided with will be presented as single tokens, separating each part-of-speech element from the other (punctuation included). Each one is separated from the other with a front- and end-whitespace. You may also encounter double whitespaces in your data. These are remainings of the meta-textual XML tags. They can be considered as a normal whitespace (no implicit meaning then).

A.2.3.4 System Requirements

The annotation procedure can be performed on your local machine and on any OS (Windows, Linux, Mac OS). To open and edit the XML files, any text editor would do. Anyway, I warmly advise you to use Sublime Text (*https://www.sublimetext.com/*) as text editor in this annotation task since natively supporting many markup languages.

A.2.4 Questions/Doubts & Answers

If you have any questions/doubts which answer is not covered in this document and/or if you want to let me know your opinion about the task itself, I warmly invite you to contact me using the following email: `fabrizio.esposito3@unina.it`.

APPENDIX B

Levin's Class N°51 of Verbs of Motion

51.1	Inherently directed motion	**Advance, Arrive, Ascend, Climb, Come, Cross, Depart, Descend, Enter, Escape, Exit, Fall, Flee, Go, Leave, Plunge, Recede, Return, Rise, Tumble**
51.2	Leave Verbs	**Abandon,Desert,Leave**
51.3.1	Manner of Motion: Roll Verbs	**Bounce,Coil,Drift,Drop,Float, Glide,Move,Revolve,Roll,Rotate, Slide,Spin,Swing,Turn,Twirl, Twist,Whirl,Wind**
51.3.2	Manner of Motion: Run Verbs	**Amble, Backpack, Bolt, Bounce, Bound, Bowl, Canter, Carom, Cavort, Charge, Clamber, Climb, Clump, Coast, Crawl, Creep, Dart, Dash, Dodder, Drift, File, Flit, Float, Fly, Frolic, Gallop, Gambol, Glide, Goosestep, Hasten, Hike, Hobble, Hop, Hurry, Hurtle, Inch, Jog, Journey, Jump, Leap, Limp, Lollop, Lope, Lumber, Lurch, March, Meander, Mince, Mosey, Nip, Pad, Parade, Perambulate, Plod, Prance, Promenade, Prowl, Race, Ramble, Roam, Roll, Romp, Rove, Run, Rush, Sashay, Saunter, Scamper, Scoot, Scram, Scramble, Scud, Scurry, Scutter, Scuttle, Shamble, Shuffle, Sidle, Skedaddle, Skip, Skitter, Skulk, Sleepwalk, Slide, Slink, Slither, Slog, Slouch, Sneak, Somersault, Speed, Stagger, Stomp, Stray, Streak, Stride, Stroll, Strut, Stumble, Stump, Swagger, Sweep, Swim, Tack, Tear, Tiptoe, Toddle, Totter, Traipse, Tramp, Travel, Trek, Troop, Trot, Trudge, Trundle, Vault, Waddle, Wade, Walk, Wander, Whiz, Zigzag, Zoom**

51.4.1	Manner of Motion using a Vehicle: Vehicle Name Verbs	**Balloon, Bicycle, Bike, Boat, Bobsled, Bus, Cab, Canoe, Caravan, Chariot, Coach, Dogsled, Ferry, Gondola, Helicopter, Jeep, Jet, Kayak, Moped, Motor, Motorbike, Motorcycle, Parachute, Punt, Raft, Rickshaw, Rocket, Skate, Skateboard, Ski, Sled, Sledge, Sleigh, Taxi, Toboggan, Tram Trolley, Yacht**
51.4.2	Manner of Motion using a Vehicle: Verbs not associated with Vehicles Names	**Cruise, Drive, Fly, Oar, Paddle, Pedal, Ride, Row, Sail, Tuck**
51.5	Waltz Verbs	**Boogie, Bop, Cancan, Clog, Conga, Dance, Foxtrot, Jig, Jitterbug, Jive, Pirouette, Polka, Quickstep, Rumba, Samba, Shuffle, Squaredance, Tango, Tapdance, Waltz**
51.6	Chase Verbs	**Chase, Follow, Pursue, Shadow, Tail, Track, Trail**
51.7	Accompany Verbs	**Accompany, Conduct, Escort, Guide, Lead, Shepherd**

APPENDIX C

Metaphor Recognition Algorithms

Algorithm 1 Pseudocode for the *Global* model algorithm.

Input: motion verb v_i
Output: Ctg, category as either metaphorical or literal

$C_{v_i} \leftarrow \emptyset$ ▷ set of similarity scores of each v_i
$sim \leftarrow 0$
threshold$_g$ ▷ threshold previously set for the *global* model
for each $t_n \in T$ **do**
    $sim \leftarrow$ cosine similarity$(embedding[v_i], embedding[t_n])$
    $C_{v_i} \leftarrow C_{v_i} \cup \{sim\}$
end for
$score \leftarrow 0$
for each $sim \in C_{v_i}$ **do**
    if $sim \geq$ threshold$_g$ **then**
        $score \leftarrow score + 1$
    end if
end for
if $score \geq 1$ **then**
    $Ctg \leftarrow$ literal
    return Ctg
else
    $Ctg \leftarrow$ metaphorical
    return Ctg
end if

Algorithm 2 Pseudocode for the *Glo-cal* algorithm

Input: $FindDirObj$ ▷ function that returns the motion verb and its direct object if true

Output: Ctg, category as either metaphorical or literal

$C_{v_i} \leftarrow \emptyset$ ▷ set of similarity scores of each v_i
$sim_s \leftarrow 0$ ▷ similarity score of *glo-cal* model
$sim_c \leftarrow 0$ ▷ similarity score of *global* model
threshold$_l$ ▷ threshold set for the *glo-cal* model
threshold$_g$ ▷ threshold set for the *global* model
if $FindDirObj$ **returns** True **then**
    $sim_s \leftarrow$ cosine similarity$(embedding[v_i], embedding[dobj_i])$
    if $sim_s \geq$ threshold$_l$ **then**
        $Ctg \leftarrow$ 'literal'
    else
        $Ctg \leftarrow$ 'metaphorical'
    end if
else
    for each $t_n \in T$ **do**
        $sim_c \leftarrow$ cosine similarity$(embedding[v_i], embedding[t_n])$
        $C_{v_i} \leftarrow C_{v_i} \cup \{sim_c\}$
    end for
    $score \leftarrow 0$
    for each $sim_c \in C_{v_i}$ **do**
        if $sim_c \geq$ threshold$_g$ **then**
            $score \leftarrow score + 1$
        end if
    end for
    if $score \geq 1$ **then**
        $Ctg \leftarrow$ 'literal'
        return Ctg
    else
        $Ctg \leftarrow$ 'metaphorical'
        return Ctg
    end if
end if

Algorithm 3 Pseudocode for the *Glo-cal* weighted model

Input: $FindDirObj$ ▷ function that returns the verb and its direct object if true

Output: Ctg, category as either metaphorical or literal

```
C_{v_i} ← ∅                                   ▷ set of similarity scores of each v_i
sim_s ← 0                                     ▷ similarity score of glo-cal model
sim_c ← 0                                     ▷ similarity score of global model
threshold_l                                   ▷ threshold set for the glo-cal model
threshold_g                                   ▷ threshold set for the global model
weight                                        ▷ weight returned from the global model
if FindDirObj returns True then
    for each t_n ∈ T do
        sim_s ← cosine similarity(embedding[v_i], embedding[t_n])
        C_{v_i} ← C_{v_i} ∪ {sim}
    end for
    score ← 0
    for each sim_s ∈ C_{v_i} do
        if sim_s ≥ threshold_l then
            score ← score + 1
        end if
    end for
    if score ≥ 1 then
        weightedsim ← 0
        sim_s ← cosine similarity(embedding[v_i], embedding[dobj_i])
        weightedsim ← sim_s + weight
        if weightedsim ≥ threshold_l then
            Ctg ← 'literal'
        else
            Ctg ← 'metaphorical'
        end if
    else
        weightedsim ← 0
        sim_s ← cosine similarity(embedding[v_i], embedding[dobj_i])
        weightedsim ← sim_s − weight
        if weightedsim ≥ threshold_l then
            Ctg ← 'literal'
        else
            Ctg ← 'metaphorical'
        end if
    end if
else
    for each t_n ∈ T do
        sim_c ← cosine similarity(embedding[v_i], embedding[t_n])
        C_{v_i} ← C_{v_i} ∪ {sim}
    end for
    score ← 0
    for each sim_c ∈ C_{v_i} do
        if sim_c ≥ threshold_g then
            score ← score + 1
        end if
    end for
    if score ≥ 1 then
        Ctg ← 'literal'
        return Ctg
    else
        Ctg ← 'metaphorical'
        return Ctg
    end if
end if
```

APPENDIX D

Utterances for Metaphorical Analysis

D.1 Saturated utterances

(19) Well, go back to use Peter Welch as an example, which is that Peter Welch would agree that leaving the lights on at a federal building overnight with nobody there is bad government. So why spend our

(20) I can't say that there'll never be and again, I'm not going to be fatalistic when we've got a vote at 3:30 p.m. I know that the President has made it clear that this is the effort, this was the train that's leaving the station, and that he expects everyone you know, this is our opportunity. And he's got a lot left on the agenda that he wants to get done, whether it's immigration, taxes, the border wall. There's so many other things that he wants to get done that we're not going to sit around and figure out this is the opportunity, this is the time, this is the opportunity for every member who has said that they want to repeal and replace Obamacare to put their vote in the "yes" column.

(21) Well, listen, “resigned” is not at all the word that I would use. I think it was evident from the Democrats in the room that the President feels quite passionate about all of these issues. And the President is confident that the kind of argument that Democrats can put forward is a winning one. So the President continues to be very confident, particularly on this issue of the Affordable Care Act, in the ability of Democrats to make the kind of argument that's going to resonate deeply with the American people. And there's already some evidence that Republicans are uneasy about this, both as and so the two pieces of evidence that I cited today are the op-ed from Senator Paul and the inability of one of the most articulate Republicans on Capitol Hill to explain why Republicans don't have their own replacement plan to put forward, even though he's the guy who's responsible for putting that plan forward. So I think that is an indication that Republicans are already starting to reckon with the challenge of keeping this promise. At the same time, the President acknowledges that he's leaving the national stage. That's what the Constitution requires. That's certainly consistent with his wife's preferences. And it's going to be time for somebody else to pick up the mantle. Does that mean that the President is any less committed to these issues than he was before? Of course not. But it does mean that the President expects to be in a position that he can observe the kinds of customs and courtesy, frankly, that was afforded to him by his predecessor.Now, the President has also been clear, and the President did discuss this in the meeting as well, that he's hopeful that this won't happen. But if there are basic, fundamental American values that are undermined by a specific policy proposal, then he may feel the need to speak out. But it is his hope, and I would say even his expectation, that that's not something that he will have to do. And I think the other thing I want to point out here and I think this is relevant to the entire context what I'm trying to lay out and describe to you is the President's plans for the first year or two that he's out of office.

And President Obama is obviously leaving this office at a young age he's just 55 and I think that there's he still has a lot of ambition and a lot more that he would like to do. Most of it he hopes he will be able to do behind the scenes in terms of continuing to stay true to his roots as a community organizer, and motivating and inspiring and even offering training to people who feel called in a similar direction. He wants to make sure that public servants, or people who aspire to public office are people who can get trained in the fundamentals of community organizing. He wants to make sure that young people around the world are exposed to the kinds of values and principles and norms and customs and traditions of the United States when it comes to democracy and citizen engagement and respect for all people, and even entrepreneurship. These are things that the President has talked about as a President and something that he hopes to continue in his post-presidency.So I don't want to leave you with the impression that there's still not a lot of important work for former President Obama to be engaged in there is. He recognizes that. And he's got a long to-do list. But that is different than being engaged in the same back-and-forth that he's responsible for engaging in as President.

D.2 Unsaturated utterances

(22) Well, I guess this is the point. So I'm glad you say that, because this, I think, is the point. The President is relying on his military advisors. They have not put forward a specific plan that would address the concerns that I've just raised. They acknowledge that. The President is relying on them for the good military advice that he's getting thus far that is having the tangible impact of applying additional pressure on ISIL. We have made progress, just looking inside of Syria, in terms of regaining more than 20 percent of the territory that ISIL previously controlled. And we've done that just by training forces

inside of Syria. So we are making progress in encircling Raqqa. We are applying significant pressure against ISIL in Syria. We are making progress in terms of shutting off the border they benefit from.So the point is the President is relying on the best military advice. He is following that military advice, and it's showing results. Not as fast as we would like; it certainly hasn't turned Syria into a Jeffersonian democracy that reflects the pluralism and diversity of that country. But we are making progress. And it's because the President is relying on the best advice that's out there.So it's important to remember that for all of the criticism about how the President's policy has not led to the kind of results that we'd all like to see inside of Syria, it's not because the President has failed to consider or implement an alternative proposal. There is nobody else that has put forward a specific idea with the possible exception of the safe zone that you referred to. I think the President has laid out in pretty clear terms why he doesn't think that's a good idea. And the truth is you don't hear a whole lot of people talking about that anymore, and I don't even know if that was mentioned in the speech today from, frankly, the President's most high-profile critic.

(23) Well, John, I think it's always important to evaluate the context of those numbers. And one important piece of context is simply that there was an historic wave that entered office at the end of 2008 and the beginning of 2009 of Democratic elected officials who benefitted from President Obama being at the top of the ballot in 2008. So when we're talking about those kinds of numbers, it's important to recognize that those numbers got built up in the first place because of President Obama's political success in winning the White House the first time.That said, the President is the leader of the Democratic Party. And he has been disappointed, particularly with regard to this most recent election, that a lot of good Democratic elected officials, public servants didn't succeed at the ballot box. And the President has expressed his view about why that is. It includes the need for

Democratic activists and Democratic voters to express their view persuasively in communities all across the country, and that certainly is part of the challenge that President Obama is going to spend some time thinking about as a former President. And this will certainly be the challenge that the incoming Democratic Party chairman will take on in taking office and making sure that Democrats are showing up and competing in communities all across the country. We've got the values right, we've got the policy prescriptions right, but we just need to go and make the argument. And the President is confident that if and when Democrats do that, there are important gains for the party and for the country that lie ahead.Let me just run through the week ahead real quick. On Saturday, the President will travel to Jacksonville, Florida to attend the wedding ceremony of a White House staffer. There will be no media coverage of the event. This is just a private event and the President is looking forward to it.On Monday, the President will attend meetings at the White House. On Tuesday, the President will travel to Chicago, Illinois, as we've discussed, to deliver his farewell address to the American people. In the address, he will thank his supporters, celebrate the ways the country has changed these past eight years, and offer some thoughts on where the country will go from here. The First Lady, the Vice President, and Dr. Biden will also attend.Through the rest of the week, the President intends to attend meetings at the White House and it should be an interesting week.Thanks, everybody, have a great weekend.

(24) Well, Jeff, we continue to be deeply concerned about the provocations and destabilizing activities that are mounted by the North Korean regime. The United States is strongly committed to denuclearizing the Korean Peninsula and standing should-to-shoulder with our allies in the Republic of Korea as they face the threat from North Korea. And that's not going to change.There is an opportunity for the North Korean government to escape the deep isolation that they currently

face. But it will require them to make a commitment to giving up their nuclear program and coming into compliance with the wide range of international obligations and U.N. Security Council resolutions that they currently ignore.So we've been quite clear about what we believe the North Korean government should do. Thus far, they have chosen the path of confrontation and provocation. As soon as they're ready to consider an alternate path toward reconciliation the international community will be ready to engage.Chris.

(25) Great. Thanks, Josh. Let me just start by giving a bit of an overview of the trip and the current expected schedule. And let me just begin by saying also that we see this trip as really bringing together a number of the President's top priorities really for the last seven and a half years.First of all, there will be a significant focus, particularly on the front end of the trip, on the efforts that we've been engaged in to confront climate change. There will also be a significant focus on the global economy through the G20. And then, of course, this is the President's 10th trip to the Asia Pacific region, and the rebalance of the Asia Pacific has been a centerpiece of our foreign policy. And the trip will give him an opportunity to once again make the case for America's focus on the Asia Pacific, to make the case for TPP as a centerpiece of our economic and strategic leadership in the region, and to address some of the very pressing issues that are going to be on the agenda. The summit is including maritime issues and the South China Sea.So, again, I think three big pieces of the presidency are going to be front and center here through climate change, the global economy and the Asia Pacific region. And I think the schedule will illustrate that.Brian will give you greater detail on the energy and climate events. As you know, we'll be departing on Wednesday morning and he'll have an initial stop in Nevada, where he will be speaking at the annual Lake Tahoe Summit with Senator Harry Reid there. And then that evening in Hawaii he will address leaders from the Pacific Island Conference of Leaders and the IUCN World Conser-

vation Congress. And this is the first time that the United States has hosted the World Conservation Congress an important opportunity to bring together not just Pacific island leaders who have been a motivating factor around the urgency of action against climate change, but also conservation advocates from around the world. And Brian will speak in greater detail to his plans there.Thursday, he will travel to Midway, where he will be able to speak to the latest marine national monument that I will leave it to Brian to pronounce and describe to you, but I think we will be bringing together through these three events both our domestic and international climate efforts and conservation efforts. The United States has been leading at home and we've been leading around the world in the pursuit of global action against climate change. Following the trip to Midway, he will depart on the morning of September 2nd for China. And then that afternoon and evening in China, he will be engaged in the bilateral program with President Xi Jinping. And this will be on Saturday, September 3rd, China time, because we will skip forward by a day. And we expect that he'll have both an extensive bilateral meeting, and then be hosted for a small dinner by President Xi Jinping as has been practiced at their previous meetings. And this will build on the work that we've done in our previous travel to Beijing, which included the historic breakthrough announcement on cooperation on climate change and also the engagements we've had here in Washington and Sunnylands with President Xi Jinping. I think we'll be reviewing all of the issues that have been front and center in the U.S.-China relationship for the last seven and a half years. On the positive side, we'll be able to review the progress we've made on the global economy, on climate change, our shared efforts to prevent the proliferation of nuclear weapons through the Iran deal, our shared concern about the situation on the Korean Peninsula. Of course, we'll also be addressing differences, as we always do with China, whether it relates to cyber issues, some of the economic practices that we have raised concerns about, some of the tensions around maritime issues in the South China Sea and, of course, our longstanding differences on human rights as well. But, again, I think this is going to be the last occasion of this sort for the President to spend several hours with his Chinese counterpart and to review the state of U.S.-China relations and to try to see where we can make progress, and working together on areas of common interest or bridging some of the differences that have been characteristic of the relationship. So that will be the program for Saturday, September 3rd. Then, on the morning of Sunday, September 4th, in advance of the G20 Summit, we expect that the President will be able to have some bilateral meetings. We certainly anticipate that one of those will be with President Erdogan of Turkey. President Obama will want to discuss obviously the circumstances in Turkey since the attempted coup, as well as our counter-ISIL campaign and our efforts to promote greater stability in Syria and to respond to the refugee crisis. We anticipate there will be additional bilateral meetings and we'll keep you posted as they are scheduled.Then the President will move into the G20 schedule, and I'll leave it to Wally to review the sessions and the agenda. That begins, again, on the afternoon of Sunday the 4th, and continues throughout the day on Monday the 5th. And it will conclude the President's time in China with his press conference at the end of the G20. And, again, we'll keep you updated on any additional bilateral meetings that are scheduled.On the evening of Monday, September 5th, the President will fly to Laos. A few words about the program in Laos. First of all, this is the first-ever U.S. presidential visit to Laos. It's a truly historic event for U.S.-Lao relationships and for the people and country of Laos. We obviously have a very difficult history with Laos, but given our increased focus on the Asia Pacific, given our attendance at the ASEAN and East Asia Summit meetings and also given this President's commitment to reach out to countries with whom we've had complicated histories, we see this as a real opportunity to advance the U.S.-Laos relationship, to begin to build a real working partnership that can benefit both of our peoples. And, again, I think the President's program in Laos will demonstrate that.We will begin on Tuesday, September 6th, with the bilateral program with the President of Laos. We also anticipate President Obama will have an opportunity to interact with the Prime Minister of Laos as well. Laos has recently gone through a leadership transition, but President Obama has been able to engage both of those leaders through his ASEAN meetings in the past.Again, I think the agenda with Laos will seek to identify areas where we can cooperate and we have an increasing development relationship focused on education and health and human capital; an increasing trade and investment relationship. We also have the substantial effort that we are ramping up to address the legacy of war in Laos. For our part, we have been steadily increasing our commitment to clearing unexploded ordnance in Laos, which has caused significant human suffering and been an impediment to development since the conclusion of the Vietnam War. We've been spending additional resources each year as it relates to clearing unexploded ordnance, and we anticipate the President will make this a focus of his visit.We also have a POW/MIA recovery effort in Laos that we're committed to continuing to pursue and, if necessary, take additional steps to ensure that we're doing everything we can to recover those who have been lost.After his bilateral program, the President will give a speech that we anticipate to be an opportunity for him to step back and review his Asia policy over the course of the last seven years. I'll think he'll talk about how far we've come in shaping an architecture in the Asia Pacific for the United States to lead and to be at the table in forums like ASEAN and the East Asia Summit. I think he'll speak to the fact that we've significantly upgraded our commercial and economic diplomacy in the region, our security presence in the partnerships that we're building, both with allies but also with emerging partners on issues like maritime security and disaster response. I think you'll hear the President give a forceful case for TPP and why it is essential to American economic and security interests for Congress to move forward with approval of TPP.Again, in this part of the world, which is the largest emerging market in the world, TPP is seen as a litmus test for U.S. leadership. TPP allows us to establish the rules of the road for trade and commerce. It's also seen as a demonstration of America's commitment to be a Pacific power. And we would be stepping back from that leadership role. We would be ceding the region to countries like China, who do not set the same types of high standards for trade agreements, were we to not follow through with TPP. So at each of his stops, including in this speech, I think you'll see the President make the case for TPP.He'll also have an opportunity to address both the enormous potential in the region for greater connectivity economically, for greater cooperation, and also some of the areas of recent difference, including the South China Sea and our approach to maritime issues. So following the speech, we anticipate the President will have time for a bilateral meeting with President Duterte of the Philippines. The Philippines is obviously a treaty ally of the United States, a party to the recent arbitral ruling in the South China Sea. I think we'll want to review the state of play as it relates to our treaty lines and the situation in the South China Sea in that dialogue with the new President of the Philippines. One Wednesday, September 7th, we anticipate the President will have an event devoted to unexploded ordnance, where he'll be able to discuss our efforts to support the people of Laos as they seek to clear unexploded ordnance, and have an opportunity to interact with some of the workers and survivors who have confronted this issue. And then the President will travel to Luang Prabang, which is a cultural capital of Laos, an historical capital of Laos, where he'll have an opportunity to do a town hall meeting with some of our Southeast Asian Leaders. For those of you who have traveled with us, you know that the YSEALI Initiative, like our African Leaders Initiative, has generated enormous enthusiasm in Southeast Asia. It's something the President is committed to, and he'll have an opportunity for a town hall. And we anticipate he'll have some cultural stops in Luang Prabang, as well. Then, that night is the gala dinner kicking off the ASEAN and East Asia summits hosted back in Vientiane.Then, on Thursday, September 8th, he will have the U.S.-ASEAN meeting and then the East Asia Summit. And again, we anticipate the potential for an additional bilateral meeting. And we'll keep you updated as we have any additions to the schedule.So again, I will stop there, and we can deal with some of the specific issues at each of the stops in questions. But let me turn it over to Brian to give you an overview of the climate and energy event.

D.3 Conventional Metaphors

(26) Look, I think the President has talked extensively about education during the primary, whether it's an Associate's Degree, a Bachelor's Degree, a PhD or vo-technical education that we've got to give students these days the options they need for the workforce; that a vo-tech education in some cases is what's in the interest of students in terms of their success, and giving them the skills to work on cars or become a computer engineer or whatever. But as we head into as we look towards the future, we've got to make sure that we're preparing our students to have the skillset that they need. And it's also the retraining aspect of that that as people get older and certain industries start to turn the corner because of technology, that we're allowing people the opportunity for retraining to give them the skillset that they need to reenter the workforce and continue to be productive.So I think as you will see look, we literally will swear in the Secretary of Education hopefully later this evening, which we will let you all know probably around the 5:00 or 6:00 hour. But that will be something that he's going to continue to have a conversation

with Secretary DeVos about. Something that she's made clear and it's unfortunate that we haven't been able to have this conversation sooner because it was held up for so long but I think that's something that Secretary DeVos will be speaking a lot about, about the education funding and skillset and opportunities that we give not just our children but people older in life who are looking to get back into the workforce through another avenue.

(27) Roberta, as you referenced the statement last night, we are disappointed that Republicans turned this into a political exercise. This is a conference report that doesn't look like it can even pass the United States Senate. But if it did, and the President was presented with the bill, he would veto it. And let me explain why.First of all, it's woefully inadequate. Our public health professionals estimate that the federal government needs $1.9 billion of funding to attack this emergency. This bill falls far short of that. This is a bill that would also steal money from other critically important public health priorities, including those funded by the Affordable Care Act and those funded by our effort to combat Ebola. So at the end of the day, this bill does not provide adequate funding.Third, this bill unfortunately includes an ideological rider blocking access to contraception for women in the U.S., including those in Puerto Rico, even though Zika is a sexually transmitted disease and even though it has been transmitted in Puerto Rico. It makes no sense. I'd also draw your attention to provision in this bill that guts some provisions of the Clean Water Act. So for those reasons, the President would veto this bill if it ever got to his desk. Again I haven't seen much analysis that suggests it could even pass the United States Senate. And for that reason, we urge Republicans to stop turning this into a political football, to actually get to work, and come up with a proposal that's going to serve the American people.

(28) Well, Julie, I quibble with some parts of your question, namely that

there wasn't much to show for the legislative effort. If you recall, there was a lot of independent reporting and analysis that getting the United States Senate to pass comprehensive immigration reform in a bipartisan way was a far-fetched prediction. It turns out Republicans and Democrats were able to roll up their sleeves and work together on one of the most challenging and complex issues of our time. So I do think that that was an important milestone to show that comprehensive immigration reform isn't just common sense, it isn't just the right thing to do consistent with American values, it's not just widely supported by wide swaths of the American people, but it's also legislatively doable. Unfortunately, House Republicans, for reasons still unclear, refuse to call that bill up for even a vote.So you're right, legislatively there still remains a lot of work to do. That's why the President was clear that the best way to fix to address this problem is going to be through comprehensive immigration reform. The President was also clear that that's unlikely to happen in this Congress, and that's why he expects a debate amongst the American people on the best pathway forward.The other, I think, answer to your question is, there does remain additional work that the President has done on this issue that remains in place, and that, as Justin referenced, includes the DACA program announced in the summer of 2012. At the time, the President announced that our lowest priorities for enforcement were these diligent, patriotic young DREAMers who grew up pledging allegiance to the flag, and that they should be able to apply to work, study and pay their taxes here. Since then, more than 730,000 people have come out of the shadows and their lives have been changed as a result. Fortunately, today's decision doesn't affect those people. I'd also say, it's one of the reason why comprehensive immigration reform has such wide support. You've got support from the evangelical community. You've got support from business leaders, from law enforcement, obviously from the Hispanic community, from unions. So you see wide and diverse constituency groups supporting this, and yet you see Republicans on the Hill throwing up their hands and doing nothing about it.

Bibliography

Ahrens, K. (2009). *Politics, Gender and Conceptual Metaphors*. Palgrave Macmillan, London.

Bangalore, S. and Joshi, A. K. (1999). Supertagging: An Approach to Almost Parsing. *Computational Linguistics*, 25(2):237–265.

Barbaresi, A. (2012). German Political Speeches, Corpus and Visualization. Technical report, ENS Lyon. 2nd Version.

Barnes, J. (2014). *The Complete Works of Aristotle: The Revised Oxford Translation, One-Volume Digital Edition*. Bollingen Series (General). Princeton University Press.

Baroni, M., Dinu, G., and Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 238–247, Baltimore, Maryland. Association for Computational Linguistics.

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, 59(1):617–645.

Basile, P., Caputo, A., and Semeraro, G. (2015). Temporal Random Indexing: A System for Analysing Word Meaning over Time. *Italian Journal of Computational Linguistics*, 1(1):55–68.

Beer, F. A. and De Landtsheer, C. (2004). Metaphors, Politics and World Politics. In *Metaphorical World Politics: Rhetorics of Democracy, War and Globalization)*, pages 5–52. Michigan State University Press, East Lansing, MI.

Beigman Klebanov, B. (2007). Experimental and computational investigation of lexical cohesion in english texts. *Unpublished Doctoral Dissertation*.

Beigman Klebanov, B. and Beigman, E. (2010). A Game-Theoretic Model of Metaphorical Bargaining. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 698–709, Uppsala, Sweden. Association for Computational Linguistics.

Beigman Klebanov, B., Beigman, E., and Diermeier, D. (2009). Discourse Topics and Metaphors. In *Proceedings of the Workshop on Computational Approaches to Linguistic Creativity*, pages 1–8, Boulder, Colorado. Association for Computational Linguistics.

Beigman Klebanov, B., Diermeier, D., and Beigman, E. (2008). Automatic Annotation of Semantic Fields for Political Science Research. *Journal of Information Technology & Politics*, 5(1):95–120.

Beigman Klebanov, B., Leong, B., Heilman, M., and Flor, M. (2014). Different Texts, Same Metaphors: Unigrams and Beyond. In *Proceedings of the Second Workshop on Metaphor in NLP*, pages 11–17, Baltimore, MD. Association for Computational Linguistics.

Beigman Klebanov, B. and Shamir, E. (2007). Reader-based exploration of lexical cohesion. *Language Resources and Evaluation*, 41(1):27–44.

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A Neural Probabilistic Language Model. *The Journal of Machine Learning Research*, 3:1137–1155.

Bethard, S., Tzuyin Lai, V., and Martin, J. H. (2009). Topic Model Analysis of Metaphor Frequency for Psycholinguistic Stimuli. In *Proceedings of the*

Workshop on Computational Approaches to Linguistic Creativity, pages 9–16, Boulder, Colorado. Association for Computational Linguistics.

Bevitori, C. (2007). Engendering Conflict? A Corpus-Assisted Analysis of Women MPs' Positioning on the War in Iraq. *Textus*, 20(1):137 – 158.

Bhatia, A. (2006). Critical discourse analysis of political press conferences. *Discourse & Society*, 17(2):173–203.

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media.

Birke, J. and Sarkar, A. (2006). A Clustering Approach for Nearly Unsupervised Recognition of Nonliteral Language. In *11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 680–690.

Black, M. (1954). Metaphor. In *Proceedings of the Aristotelian Society*, volume 55, pages 273–294. Blackwell Publishing.

Black, M. (1962). *Models and Metaphors: Studies in Language and Philosophy*. Studies in Language and Philosophy. Cornell University Press.

Blei, D. M. (2012). Probabilistic Topic Models. *Communications of the ACM*, 55(4):77–84.

Blei, D. M. and Lafferty, J. D. (2009). Visualizing Topics with Multi-Word Expressions. *arXiv:0907.1013v1 [stat.ML]*.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

Bowdle, B. F. and Gentner, D. (2005). The Career of Metaphor. *Psychological Review*, 112(1):193–216.

Boyd-Graber, J. L. and Blei, D. M. (2009). Syntactic Topic Models. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems 21*, pages 185–192. Curran Associates, Inc.

Brigadir, I., Greene, D., and Cunningham, P. (2015). Analyzing Discourse Communities with Distributional Semantic Models. In *Proceedings of the ACM Web Science 2015 Conference*. ACM. Article 27.

Brown, R. and Gilman, A. (1960). The Pronouns of Power and Solidarity. In Sebeok, T. A., editor, *Style in Language*, pages 253–276. MIT Press.

Bruni, E., Tran, N.-K., and Baroni, M. (2014). Multimodal Distributional Semantics. *Journal of Artificial Intelligence Research*, 49(1):1–47.

Burnard, L. (2007). Reference Guide for the British National Corpus (XML Edition).

Burnard, L. and Sperberg-McQueen, C. (2006). TEI Lite: Encoding for interchange: an introduction to the TEI – Revised for TEI P5 release.

Cameron, L. (1999). Identifying and describing metaphor in spoken discourse data. In Cameron, L. and Low, G., editors, *Researching and Applying Metaphor*, Cambridge Applied Linguistics, pages 105–132. Cambridge University Press.

Cameron, L. (2003). *Metaphor in Educational Discourse*. Advances in Applied Linguistics. Bloomsbury Academic.

Cap, P. (2005). Language and legitimization: Developments in the proximization model of political discourse analysis. *Lodz Papers in Pragmatics*, 1(1):7–36.

Cap, P. (2006). *Legitimisation in Political Discourse: A Cross-Disciplinary Perspective on the Modern US War Rhetoric (Second Edition)*. Cambridge Scholars Publishing, Newcastle, UK.

Cap, P. (2013). *Proximization: The pragmatics of symbolic distance crossing*, volume 232 of *Pragmatics & Beyond New Series*. John Benjamins Publishing, Amsterdam.

Cap, P. (2014). Applying cognitive pragmatics to Critical Discourse Studies: A proximization analysis of three public space discourses. *Journal of Pragmatics*, 70:16–30.

Chilton, P. (2004). *Analysing Political Discourse: Theory and Practice*. Routledge.

Chilton, P. (2005). Discourse Space Theory: Geometry, Brain and Shifting Viewpoints. *Annual Review of Cognitive Linguistics*, 3:78–116.

Choi, J. D., Tetreault, J., and Stent, A. (2015). It Depends: Dependency Parser Comparison Using A Web-based Evaluation Tool. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 387–396, Beijing, China. Association for Computational Linguistics.

Chovanec, J. (2010). Legitimation through differentiation. Discursive construction of Jacques Le Worm Chirac as an opponent to military action. *Perspectives in Politics and Discourse*, 36:119–141.

Church, K. W. and Hanks, P. (1990). Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, 16(1):22–29.

Cienki, A., Kaal, B., and Maks, I. (2010). Mapping world view in political texts using Discourse Space Theory: Metaphor as an analytical tool. In *Eight meeting of the Association for Researching and Applying Metaphor (RaAM 8)*, Amsterdam.

Cimmino, T., Esposito, F., and Cutugno, F. (2016). CompWHoB: un corpus Open-Data di testi politici inglesi della Casa Bianca. Unpublished Bachelor of Science Thesis.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

Collins, C. (1999). *Margaret Thatcher, Complete Public Statements 1945-1990 on CD-ROM: A unique Record of One of the Twentieth Century's Most Prominent Political Leaders*. Oxford University Press.

Collobert, R. and Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*, pages 160–167. ACM.

Coombs, W. T. (2007). Protecting Organization Reputations During a Crisis: The Development and Application of Situational Crisis Communication Theory. *Corporate Reputation Review*, 10(3):163–176.

Cousins, K. and McIntosh, W. (2005). More than Typewriters, More than Adding Machines: Integrating Information Technology into Political Research. *Quality & Quantity*, 39(5):581–614.

Dagan, I., Marcus, S., and Markovitch, S. (1993). Contextual word similarity and estimation from sparse data. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, pages 164–171. Association for Computational Linguistics.

Davidson, D. (1978). What Metaphors Mean. *Critical Inquiry*, 5(1):31–47.

Davies, M. (2009). The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, 14(2):159–190.

Do Dinh, E.-L. and Gurevych, I. (2016). Token-Level Metaphor Detection using Neural Networks. In *Proceedings of the Fourth Workshop on Metaphor in NLP*, pages 28–33, San Diego, California. Association for Computational Linguistics.

Drulák, P. (2005). Identifying and assessing metaphors: the discourse on European future. In *European Consortium for Political Research (ECPR) Joint Sessions of Workshops*, Granada. Draft Paper. Available at http://eis.bris.ac.uk/ potfc/Granada/Papers/Drulak.pdf.

Dunmire, P. (2011). *Projecting the Future through Political Discourse: The case of the Bush doctrine*. Discourse Approaches to Politics, Society and Culture. John Benjamins Publishing Company.

Dunn, J. (2010). Gradient Semantic Intuitions of Metaphoric Expressions. *Metaphor and Symbol*, 26(1):53–67.

Dunn, J. (2013a). Evaluating the Premises and Results of Four Metaphor Identification Systems. In *CICLing'13 Proceedings of the 14th International Conference on Computational Linguistics and Intelligent Text Processing*, pages 471–486. Springer Heidelberg.

Dunn, J. (2013b). How linguistic structure influences and helps to predict metaphoric meaning. *Cognitive Linguistics*, 24(1):33–66.

Dunn, J. (2013c). What metaphor identification systems can tell us about metaphor-in-language. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 1–10, Atlanta, Georgia. Association for Computational Linguistics.

Dunn, J. (2014). Measuring metaphoricity. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 745–751, Baltimore, Maryland. Association for Computational Linguistics.

Engelbart, D. (1962). *Augmenting Human Intellect: a Conceptual Framework*. AFOSR. Air Force Office of Scientific Research, United States Air Force.

Esposito, F., Basile, P., Cutugno, F., and Venuti, M. (2015). The CompWHoB Corpus: Computational Construction, Annotation and Linguistic Analysis of the White House Press Briefings Corpus. In Bosco, C., Tonelli, S., and Zanzotto, F. M., editors, *Proceedings of the Third Italian Conference on Compuational Linguistics CLiC-it 2015*, pages 120–124.

Esposito, F., Corazza, A., and Cutugno, F. (2016). Topic Modelling with Word Embeddings. In Corazza, A., Montemagni, S., and Semeraro, G., editors, *Proceedings of the Third Italian Conference on Compuational Linguistics CLiC-it 2015*, pages 129–134.

Esposito, F., Esposito, E., and Basile, P. (2017). White House Under Attack. Introducing Distributional Semantic Models for the Analysis of U.S. Crisis Communication Strategies. In Grassia, M. and Marino, M., editors, *Data Science and Social Research*. Springer. To appear.

Evert, S. (2005). *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. PhD thesis, University of Stuttgart.

Evert, S. (2008). Corpora and collocations. *Corpus linguistics. An International Handbook*, 2:1212–1248.

Evert, S. (2010a). Distributional Semantic Models. In *NAACL HLT 2010 Tutorial Abstracts*, pages 15–18, Los Angeles, California. Association for Computational Linguistics.

Evert, S. (2010b). Google Web 1T 5-Grams Made Easy (but not for the computer). In *Proceedings of the NAACL HLT 2010 Sixth Web as Corpus Workshop*, pages 32–40, NAACL-HLT, Los Angeles. Association for Computational Linguistics.

Faruqui, M., Dodge, J., Jauhar, S. K., Dyer, C., Hovy, E. H., and Smith, N. A. (2014). Retrofitting Word Vectors to Semantic Lexicons. *CoRR*, abs/1411.4166.

Fass, D. (1991). met*: A method for discriminating metonymy and metaphor by computer. *Computational Linguistics*, 17(1):49–90.

Firth, J. (1957). *Papers in linguistics, 1934-1951*. Oxford University Press.

Fleiss, J. L. (1971). Measuring Nominal Scale Agreement Among Many Raters. *Psychological Bulletin*, 76(5):378 – 382.

Gedigian, M., Bryant, J., Narayanan, S., and Ciric, B. (2006). Catching Metaphors. In *Proceedings of the Third Workshop on Scalable Natural Language Understanding*, pages 41–48. Association for Computational Linguistics.

Gelo, O. C. G. (2008). *Metaphors and emotional-cognitive regulation in psychotherapy: a single-case analysis*. Ulmer Textbank.

Gelo, O. C. G. and Mergenthaler, E. (2012). Unconventional metaphors and emotional-cognitive regulation in a metacognitive interpersonal therapy. *Psychotherapy Research*, 22(2):159–175.

Gibbs, R. W. (1984). Literal Meaning and Psychological Theory. *Cognitive Science*, 8(3):275–304.

Goldberg, Y. and Levy, O. (2014). word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. *CoRR*, abs/1402.3722.

Graff, D. and Cieri, C. (2003). English Gigaword. Linguistic Data Consortium (LDC) 2003 T05. Phiadelphia: Linguistic Data Consortium.

Guerini, M., Giampiccolo, D., Moretti, G., Sprugnoli, R., and Strapparava, C. (2013). The New Release of CORPS: A Corpus of Political Speeches Annotated with Audience Reactions. In *Multimodal Communication in Political Speech. Shaping Minds and Social Action*, pages 86–98. Springer.

Guerini, M., Strapparava, C., and Stock, O. (2008). Corps: A corpus of Tagged Political Speeches for Persuasive Communication Processing. *Journal of Information Technology & Politics*, 5(1):19–32.

Guiraud, P. (1954). *Les Charactères Statistiques du Vocabulaire. Essai de méthodologie*. Presses Universitaires de France, Paris.

Haagsma, H. and Bjerva, J. (2016). Detecting novel metaphor using selectional preference information. In *Proceedings of the Fourth Workshop on Metaphor in NLP*, pages 10–17, San Diego, California. Association for Computational Linguistics.

Hall, D., Jurafsky, D., and Manning, C. D. (2008). Studying the History of Ideas Using Topic Models. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 363–371, Honolulu, Hawaii. Association for Computational Linguistics.

Halliday, M. and Hasan, R. (1976). *Cohesion in English*. English language series. Longman, London.

Hardie, A. (2012). CQPweb – combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, 17(3):380–409.

Harris, Z. (1970). Distributional structure. In *Papers in Structural and Transformational Linguistics*, pages 775–794. Springer.

Harris, Z. S. (1954). Distributional structure. *Word*, 10(2-3):146–162.

Hart, C. (2010). *Critical Discourse Analysis and Cognitive Science: New Perspectives on Immigration Discourse*. Palgrave Macmillan.

Heintz, I., Gabbard, R., Srivastava, M., Barner, D., Black, D., Friedman, M., and Weischedel, R. (2013). Automatic Extraction of Linguistic Metaphors with LDA Topic Modeling. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 58–66, Atlanta, Georgia. Association for Computational Linguistics.

Hirst, G., St-Onge, D., et al. (1998). Lexical chains as representations of context for the detection and correction of malapropisms. *WordNet: An electronic lexical database*, 305:305–332.

Hoffman, M., Bach, F. R., and Blei, D. M. (2010). Online Learning for Latent Dirichlet Allocation. In Lafferty, J. D., Williams, C. K. I., Shawe-Taylor, J., Zemel, R. S., and Culotta, A., editors, *Advances in Neural Information Processing Systems 23*, pages 856–864. Curran Associates, Inc.

Hofmann, T. (2001). Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Machine Learning*, 42(1):177–196.

Honnibal, M. and Johnson, M. (2015). An Improved Non-monotonic Transition System for Dependency Parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal. Association for Computational Linguistics.

Hovy, D., Shrivastava, S., Jauhar, S. K., Sachan, M., Goyal, K., Li, H., Sanders, W., and Hovy, E. (2013). Identifying Metaphorical Word Use with Tree Kernels. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 52–57, Atlanta, Georgia. Association for Computational Linguistics.

Iacobacci, I., Pilehvar, M. T., and Navigli, R. (2016). Embeddings for Word Sense Disambiguation: An Evaluation Study. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 897–907, Berlin, Germany. Association for Computational Linguistics.

Jang, H., Moon, S., Jo, Y., and Rose, C. (2015). Metaphor Detection in Discourse. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 384–392, Prague, Czech Republic. Association for Computational Linguistics.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., and Darrell, T. (2014). Caffe: Convolutional Architecture for Fast Feature Embedding. In *Proceedings of the 22nd ACM International Conference on Multimedia*, pages 675–678. ACM.

Jones, N. (1996). *Soundbites and Spin Doctors: How Politicians Manipulate the Media and Vice Versa*. Indigo.

Karin Kipper, Anna Korhonen, N. R. M. P. (2006). A Large-Scale Extension of VerbNet with Novel Verb Classes. In Elisa Corino, Carla Marello, C. O., editor, *Proceedings of the 12th EURALEX International Congress*, pages 173–184, Torino, Italy. Edizioni dell'Orso.

Karlgren, J. and Sahlgren, M. (2001). From Words to Understanding. In Uesaka, Y., Kanerva, P., and Asoh, H., editors, *Foundations of Real World Intelligence*, pages 294–308. Stanford: CSLI Publications.

Karov, Y. and Edelman, S. (1998). Similarity-based Word Sense Disambiguation. *Computational linguistics*, 24(1):41–59.

Kiela, D. and Bottou, L. (2014). Learning Image Embeddings using Convolutional Neural Networks for Improved Multi-Modal Semantics. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 36–45, Doha, Qatar. Association for Computational Linguistics.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlỳ, P., and Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, 1(1):7–36.

Kittay, E. F. (1990). *Metaphor: Its Cognitive Force and Linguistic Structure*. Oxford University Press.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *MT summit*, volume 5, pages 79–86.

Kopytowska, M. (2010). Television News and the 'Meta-War'. In Witczak-Plisiecka, I., editor, *Pragmatic Perspectives on Language and Linguistics*, pages 301–326. Cambridge Scholars Publishing, Newcastle.

Kovecses, Z. (2010). *Metaphor: A Practical Introduction*. Oxford University Press.

Krishnakumaran, S. and Zhu, X. (2007). Hunting Elusive Metaphors Using Lexical Resources. In *Proceedings of the Workshop on Computational Approaches to Figurative Language*, pages 13–20, Rochester, New York. Association for Computational Linguistics.

Kumar, M. J. (2010). *Managing the President's Message: The White House Communications Operation*. JHU Press.

Lakoff, G. (1993). The Contemporary Theory of Metaphor. In Ortony, A., editor, *Metaphor and Thought*, pages 202–251. Cambridge University Press.

Lakoff, G. (2008). *The Political Mind: A Cognitive Scientist's Guide to Your Brain and Its Politics*. Penguin Publishing Group.

Lakoff, G., Espenson, J., and Schwartz, A. (1991). Master Metaphor List: 2nd draft copy. Cognitive Linguistics Group. University of Berkeley.

Lakoff, G. and Johnson, M. (1980). *Metaphors We Live By*. University of Chicago Press, Chicago.

Lakoff, G. and Turner, M. (1989). *More Than Cool Reason: A Field Guide to Poetic Metaphor*. University of Chicago Press.

Lakoff, G. and Wehling, E. (2012). *The Little Blue Book: The Essential Guide to Thinking and Talking Democratic*. Simon and Schuster.

Landauer, T. K. and Dumais, S. T. (1997). A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological Review*, 104(2):211.

Landis, J. R. and Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., Van Kleef, P., Auer, S., et al. (2015).

DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web*, 6(2):167–195.

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Italian Journal of Linguistics*, 20(1):1–31.

Lenci, A. (2010). Modelli Distribuzionali del Lessico. *Informatica Umanistica*, 3:57–69.

Levin, B. (1993). *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press.

Levin, B. and Hovav, M. R. (1992). The Lexical Semantics of Verbs of Motion: The Perspective from Unaccusativity. In Roca, I., editor, *Thematic Structure: Its Role in Grammar*, Linguistic Models, pages 247–269. Walter de Gruyter, Berlin.

Levy, O. and Goldberg, Y. (2014). Dependency-Based Word Embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 302–308, Baltimore, Maryland. Association for Computational Linguistics.

Li, L., Roth, B., and Sporleder, C. (2010). Topic Models for Word Sense Disambiguation and Token-Based Idiom Detection. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1138–1147, Uppsala, Sweden. Association for Computational Linguistics.

Lichtman, M. (2013). *Qualitative Research for the Social Sciences*. Sage Publications.

Lin, D. and Wu, X. (2009). Phrase Clustering for Discriminative Learning. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1030–1038, Suntec, Singapore. Association for Computational Linguistics.

Liu, L., Tang, L., Dong, W., Yao, S., and Zhou, W. (2016). An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus*, 5(1):1608.

Louwerse, M. M. (2011). Symbol Interdependency in Symbolic and Embodied Cognition. *Topics in Cognitive Science*, 3(2):273–302.

Lund, K. and Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28(2):203–208.

Maltese, J. (1994). *Spin Control: The White House Office of Communications and the Management of Presidential News*. University of North Carolina Press.

Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Martin, J. H. (1990). *A Computational Model of Metaphor Interpretation*. Academic Press Professional, Inc.

Mason, Z. J. (2004). CorMet: A Computational, Corpus-Based Conventional Metaphor Extraction System. *Computational linguistics*, 30(1):23–44.

Merz, N., Regel, S., and Lewandowski, J. (2016). The Manifesto Corpus: A new resource for research on political parties and quantitative text analysis. *Research & Politics*, 3(2):1–8.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient Estimation of Word Representations in Vector Space. *CoRR*, abs/1301.3781.

Mikolov, T., Le, Q. V., and Sutskever, I. (2013b). Exploiting Similarities among Languages for Machine Translation. *CoRR*, abs/1309.4168.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013c). Distributed Representations of Words and Phrases and their Compositionality. *CoRR*, abs/1310.4546.

Miller, G. A. (1995). WordNet: a Lexical Database for English. *Communications of the ACM*, 38(11):39–41.

Miller, G. A. and Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

Mnih, A. and Hinton, G. (2007). Three New Graphical Models for Statistical Language Modelling. In *Proceedings of the 24th International Conference on Machine learning*, pages 641–648. ACM.

Mohammad, S., Shutova, E., and Turney, P. (2016). Metaphor as a Medium for Emotion: An Empirical Study. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 23–33, Berlin, Germany. Association for Computational Linguistics.

Morris, J. and Hirst, G. (1991). Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text. *Computational linguistics*, 17(1):21–48.

Musolf, A. (2000). *Mirror Images of Europe: Metaphors in the Public Debate about Europe in Britain and Germany*. Iudicium Verlag.

Navarro-Colorado, B. and Tomas, D. (2015). A fully unsupervised Topic Modeling approach to metaphor identification. *Actas del XXXI Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural*, 55:91–98.

Nelson, W. (2000). *Who Speaks for the President?: The White House Press Secretary from Cleveland to Clinton*. Syracuse University Press.

Nivre, J., Hall, J., and Nilsson, J. (2006). Maltparser: A Data-Driven Parser-Generator for Dependency Parsing. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, volume 6, pages 2216–2219.

Niwa, Y. and Nitta, Y. (1994). Co-occurrence vectors from corpora vs. distance vectors from dictionaries. In *Proceedings of the 15th Conference on Computational linguistics*, volume 1, pages 304–309. Association for Computational Linguistics.

Nunberg, G. (1987). Poetic and Prosaic Metaphors. In *Proceedings of the 1987 workshop on Theoretical issues in Natural Language Processing*, pages 198–201. Association for Computational Linguistics.

Ó Séaghdha, D. (2010). Latent Variable Models of Selectional Preference. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 435–444, Uppsala, Sweden. Association for Computational Linguistics.

Osenova, P. and Simov, K. (2012). The Political Speech Corpus of Bulgarian. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Padó, S. and Lapata, M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

Partington, A. (2003). *The Linguistics of Political Argument: The Spin-Doctor and the Wolf-Pack at the White House*. Routledge Advances in Corpus Linguistics. Taylor & Francis.

Passos, A., Kumar, V., and McCallum, A. (2014). Lexicon Infused Phrase Embeddings for Named Entity Resolution. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 78–86, Ann Arbor, Michigan. Association for Computational Linguistics.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Perloff, R. (1998). *Political Communication: Politics, Press, and Public in America*. Lawrence Erlbaum Associates (LEA)' Communication Series. Lawrence Erlbaum Associates.

Prabhakaran, V., Arora, A., and Rambow, O. (2014). Staying on Topic: An Indicator of Power in Political Debates. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1481–1486, Doha, Qatar. Association for Computational Linguistics.

Prabhakaran, V., John, A., and Seligmann, D. D. (2013). Power Dynamics in Spoken Interactions: A Case Study on 2012 Republican Primary Debates. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 99–100. ACM.

Pragglejaz Group, M. (2007). MIP: A Method for Identifying Metaphorically Used Words in Discourse. *Metaphor and Symbol*, 22(1):1–39.

Purpura, S. and Hillard, D. (2006). Automated Classification of Congressional Legislation. In *Proceedings of the 2006 International Conference on Digital Government Research*, pages 219–225. Digital Government Society of North America.

Purpura, S., Hillard, D., and Howard, P. (2006). A Comparative Study of Human Coding and Context Analysis against Support Vector Machines (SVM) to Differentiate Campaign Emails by Party and Issues. *Draft Paper*. Available at www.purpuras.net.

Ratnaparkhi, A. et al. (1996). A Maximum Entropy Model for Part-Of-Speech Tagging. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, volume 1, pages 133–142. Philadelphia, PA.

Rayson, P. (2003). *Matrix: A statistical method and software tool for linguistic analysis through corpus comparison*. PhD thesis, Lancaster University.

Reddy, M. (1979). The Conduit Metaphor – A Case of Frame Conflict in Our Language about Language. In Ortony, A., editor, *Metaphor and Thought*, volume 2, pages 284–310. Cambridge University Press, Cambridge.

Rehurek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. European Language Resources Association (ELRA).

Reimers, N., Eckle-Kohler, J., Schnober, C., Kim, J., and Gurevych, I. (2014). Germeval-2014: Nested Named Entity Recognition with Neural Networks. In *Workshop Proceedings of the 12th Edition of the KONVENS Conference*, pages 117–120.

Roberts, W. and Egg, M. (2014). A Comparison of Selectional Preference Models for Automatic Verb Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 511–522, Doha, Qatar. Association for Computational Linguistics.

Rothe, S. and Schütze, H. (2015). AutoExtend: Extending Word Embeddings to Embeddings for Synsets and Lexemes. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1793–1803, Beijing, China. Association for Computational Linguistics.

Rubenstein, H. and Goodenough, J. B. (1965). Contextual Correlates of Synonymy. *Communications of the ACM*, 8(10):627–633.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252.

Sahlgren, M. (2006). *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. PhD thesis, Stockholm University.

Sahlgren, M. (2008). The distributional hypothesis. *Italian Journal of Disability Studies*, 20:33–53.

Schütze, H. (1993). Word Space. *Advances in Neural Information Processing Systems*, 5:895–902.

Schütze, H. and Pedersen, J. O. (1997). A Cooccurrence-Based Thesaurus and Two Applications to Information Retrieval. *Information Processing & Management*, 33(3):307–318.

Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. *ACM Computing Surveys (CSUR)*, 34(1):1–47.

Shutova, E. (2010a). Automatic Metaphor Interpretation as a Paraphrasing Task. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1029–1037, Los Angeles, California. Association for Computational Linguistics.

Shutova, E. (2010b). Models of Metaphor in NLP. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 688–697, Uppsala, Sweden. Association for Computational Linguistics.

Shutova, E. (2015). Design and Evaluation of Metaphor Processing Systems. *Computational Linguistics*, 41(4):579–623.

Shutova, E., Kiela, D., and Maillard, J. (2016). Black Holes and White Rabbits: Metaphor Identification with Visual Features. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 160–170, San Diego, California. Association for Computational Linguistics.

Shutova, E. and Sun, L. (2013). Unsupervised Metaphor Identification Using Hierarchical Graph Factorization Clustering. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 978–988, Atlanta, Georgia. Association for Computational Linguistics.

Shutova, E., Sun, L., and Korhonen, A. (2010). Metaphor Identification Using Verb and Noun Clustering. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1002–1010. Association for Computational Linguistics.

Shutova, E. and Teufel, S. (2010). Metaphor Corpus Annotated for Source - Target Domain Mappings. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, pages 3255–3261, Valletta, Malta. European Language Resources Association (ELRA).

Shutova, E., Teufel, S., and Korhonen, A. (2013). Statistical Metaphor Processing. *Computational Linguistics*, 39(2):301–353.

Sim, Y., Acree, B. D. L., Gross, J. H., and Smith, N. A. (2013). Measuring Ideological Proportions in Political Speeches. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 91–101, Seattle, Washington, USA. Association for Computational Linguistics.

Spinzi, C. and Venuti, M. (2013). Tracking the change in an Institutional Genre: A Diachronic Corpus-based study of White House Press Briefings. In *The Three Waves of Globalization: Winds of Change in Professional, Institutional and Academic Genres*, pages 182–197. Cambridge Scholars Publishing, Newcastle upon Tyne.

Spinzi, C. and Venuti, M. (2016). A Corpus-based Discourse Analysis of Security in White House Press Briefings. In *The Languages of Politics Volume 2*, pages 91–114. Cambridge Scholars Publishing, Newcastle upon Tyne.

Sporleder, C. and Li, L. (2009). Unsupervised Recognition of Literal and Non-Literal Use of Idiomatic Expressions. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 754–762, Athens, Greece. Association for Computational Linguistics.

Steen, G. J., Dorst, A. G., Herrmann, J. B., Kaal, A., Krennmayr, T., and Pasma, T. (2010). *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*. Converging evidence in language and communication research. John Benjamins Publishing Company.

Strzalkowski, T., Broadwell, G. A., Taylor, S., Feldman, L., Shaikh, S., Liu, T., Yamrom, B., Cho, K., Boz, U., Cases, I., and Elliot, K. (2013). Robust Extraction of Metaphor from Novel Data. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 67–76, Atlanta, Georgia. Association for Computational Linguistics.

Taghipour, K. and Ng, H. T. (2015). Semi-Supervised Word Sense Disambiguation Using Word Embeddings in General and Specific Domains. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 314–323, Denver, Colorado. Association for Computational Linguistics.

Thomas, M., Pang, B., and Lee, L. (2006). Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 327–335, Sydney, Australia. Association for Computational Linguistics.

Thompson, S. (1996). Politics Without Metaphors is Like a Fish Without Water. *Metaphor: Implications and Applications*, pages 185–201.

Tsvetkov, Y., Boytsov, L., Gershman, A., Nyberg, E., and Dyer, C. (2014). Metaphor Detection with Cross-Lingual Model Transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Lin-*

guistics, pages 248–258, Baltimore, Maryland. Association for Computational Linguistics.

Turney, P., Neuman, Y., Assaf, D., and Cohen, Y. (2011). Literal and Metaphorical Sense Identification through Concrete and Abstract Context. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 680–690, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Turney, P. D. and Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

Veale, T., Shutova, E., and Klebanov, B. (2016). *Metaphor: A Computational Perspective*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Venuti, M., Cava, A., de Candia, S., Riccio, G., and Spinzi, C. (2009). Evolution of institutional genres in time: the case of the White House press briefings. In Mahlberg, M., Gonzaĺez-Diáz, V., and Smith, C., editors, *Proceedings from the Corpus Linguistics Conference Series*. Available at http://ucrel.lancs.ac.uk/publications/cl2009/387_FullPaper.doc.

Venuti, M., Spinzi, C., and De Candia, S. (2012). "I don't know the answer to that question": A corpus-assisted discourse analysis of White House Press Briefings. In *Critical Approaches to Discourse Analysis across Disciplines (CADAAD)*, volume 7 of *1*, pages 66–81.

Volkens, A., Lehmann, P., Matthieß, T., Merz, N., Regel, S., and Werner, A. (2015). The Manifesto Data Collection. Manifesto Project (MRG/CMP/MARPOR). *Berlin: Wissenschaftszentrum Berlin für Sozialforschung (WZB)*. Version 2015a.

Vrandečić, D. and Krötzsch, M. (2014). Wikidata: A Free Collaborative Knowledgebase. *Communications of the ACM*, 57(10):78–85.

Widdows, D. (2004). *Geometry and Meaning*. Center for Study of Language and Information Publications Stanford.

Wilks, Y. (1975). A preferential, pattern-seeking, semantics for natural language inference. *Artificial intelligence*, 6(1):53–74.

Wilks, Y. (1978). Making preferences more active. *Artificial Intelligence*, 11(3):197–223.

Wittgenstein, L. (1973). *Philosophical Investigations*. Philosophical Investigations: The English Text of the Third Edition. Macmillan.

Woolley, J. T. and Peters, G. (2008). The American Presidency Project. Santa Barbara, CA.

Xiao, R. (2006). REVIEW: Xaira – An XML Aware Indexing and Retrieval Architecture. *Corpora*, 1(1):99–103.

Zipf, G. (1949). *Human Behavior and the Principle of Least Effort*. Addison-Wesley Press, Cambridge, MA.