Acquiring and Harnessing Verb Knowledge for Multilingual Natural Language Processing

Olga Anna Majewska

Theoretical and Applied Linguistics
University of Cambridge

This thesis is submitted for the degree of
Doctor of Philosophy

Corpus Christi College

February 2021

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the preface and specified in the text. I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee.

Olga Anna Majewska
February 2021

Acquiring and Harnessing Verb Knowledge for Multilingual Natural Language Processing

Olga Anna Majewska

Abstract

Advances in representation learning have enabled natural language processing models to derive non-negligible linguistic information directly from text corpora in an unsupervised fashion. However, this signal is underused in downstream tasks, where they tend to fall back on superficial cues and heuristics to solve the problem at hand. Further progress relies on *identifying* and *filling* the gaps in linguistic knowledge captured in their parameters. The objective of this thesis is to address these challenges focusing on the issues of resource scarcity, interpretability, and lexical knowledge injection, with an emphasis on the category of verbs.

To this end, I propose a novel paradigm for efficient acquisition of lexical knowledge leveraging native speakers' intuitions about verb meaning to support development and downstream performance of NLP models across languages. First, I investigate the potential of acquiring semantic verb classes from non-experts through manual clustering. This subsequently informs the development of a two-phase semantic dataset creation methodology, which combines semantic clustering with fine-grained semantic similarity judgments collected through spatial arrangements of lexical stimuli. The method is tested on English and then applied to a typologically diverse sample of languages to produce the first large-scale multilingual verb dataset of this kind. I demonstrate its utility as a diagnostic tool by carrying out a comprehensive evaluation of state-of-the-art NLP models, probing representation quality across languages and domains of verb meaning, and shedding light on their deficiencies. Subsequently, I directly address these shortcomings by injecting lexical knowledge into large pretrained language models. I demonstrate that external manually curated information about verbs' lexical properties can support data-driven models in tasks where accurate verb processing is key. Moreover, I examine the potential of extending these benefits from resource-rich to resource-poor languages through translation-based transfer. The results emphasise the usefulness of human-generated lexical knowledge in supporting NLP models and suggest that time-efficient construction of lexicons similar to those developed in this work, especially in under-resourced languages, can play an important role in boosting their linguistic capacity.

To my parents, who made all of this possible

Acknowledgements

I would like to thank my supervisor, professor Anna Korhonen, for giving me the one-of-a-kind opportunity to pursue my PhD in this fascinating field of research. Her guidance and support have been critical to the completion of this project. My sincere thanks go to Diana McCarthy, whose generosity and thought-provoking suggestions have motivated me to always look deeper and further. Moreover, I am utterly indebted to Ivan Vulić for sharing his expertise with me throughout the course of this PhD. His technical and strategic advice have been invaluable to my development as a researcher.

I thank Edoardo, my rock and inspiration. His affection, brilliance, and insight have spurred me to be a better scientist and a better person. Thank you for being by my side whenever things got rough and for your precious company on the many adventures we have undertaken together, the true highlights of these four years.

I am forever indebted to my parents, Barbara and Tomasz, without whose unconditional support, sacrifice, and constant encouragement since my first day of school, I would have never found myself in the position to undertake this PhD. I thank my brother Piotr, whose friendship and wit have kept me grounded throughout this journey.

I sincerely thank my collaborators Jasper van den Bosch and Goran Glavaš for their help and Thierry Poibeau for the encouraging invitation to present the early stages of this research. I am grateful to all fellow Language Technology Lab members for their inspiring company and discussions, especially Daniela, for taking me under her wing in my first months as a PhD student, and Billy, Gamal and Milan for making me feel so welcome. I also thank the colleagues I had the pleasure to work with and learn from during my internships at Amazon and Wluper, with special thanks to Nikolai Rozanov for the stimulating collaboration, uplifting humour, and faith in my ability.

Finally, I thank the special friends who helped me stay afloat through it all. Ulrike, for her boundless empathy and surprise care packages in moments of need. My fantastic PhD support group, Anna, Catherine, and Sansan, for all the camaraderie, antics, and laughs. Sara, for the perspective-changing conversations and Ditte, for encouraging me whenever I needed a confidence boost. And Laura, my other bucket half, who put fun back into my PhD existence and made all the cold early mornings on the river worth it.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Human communication involves manipulating nuanced linguistic knowledge, readily available in native speakers' mental lexicon. State-of-the-art natural language processing systems strive to match humans' intuitive mastery of linguistic complexities by leveraging lexical-semantic and structural information found in large volumes of text. However, they are highly reliant on abundant data and our understanding of the nature of the encoded knowledge is still limited. Moreover, when faced with complex tasks requiring sensitivity to fine-grained meaning distinctions and ability to make nuanced inferences about the world, current deep neural architectures have been shown to find task-specific shortcuts, rather than making use of the linguistic signal gleaned from data. The resultant high performance scores on natural language understanding benchmarks suggest an impressive linguistic capacity, but reveal very little about what really contributed to achieving them and may conceal the model's deficiencies. Development of resources and tools for focused evaluation and probing is therefore essential to tackle the problem of limited interpretability of current neural architectures. What is more, given that they learn purely from the distributional signal in text corpora, external knowledge bases can play an important role in supplying the missing linguistic or real-world information. The difficulty lies in the limited availability of such resources in the majority of the world's languages and the time and expense entailed in constructing them.

One challenging area of linguistic knowledge concerns reasoning about verbs. Functioning as pivots within sentence structure, verbs carry information about the event taking place and its extension and position in time, as well as the roles assumed by the participating actors and the relations between them. Moreover, their meaning is

strongly interlinked with their syntactic behaviour, and each level of analysis provides important cues about the other that are crucial both in language acquisition and understanding. Mastering and accurately manipulating the rich lexical information encoded in verbs therefore constitutes one of the keys to successful processing of speech and text by machines. However, their complex properties are difficult to model and automatically acquire from unlabelled data. This is why structured resources organising fine-grained verb knowledge in computer-readable form are critical in facilitating the development of systems with a deeper awareness of linguistic phenomena.

In the following Section 1.1.1, I discuss the importance of focusing resource creation efforts on verbs as bearers of information crucial to sentence understanding and introduce the notion of verb class as an organisational unit of verb knowledge. Then, in Section 1.1.2, I argue that current intrinsic evaluation protocols can be enriched by the inclusion of lexical-semantic benchmarks capturing fine-grained meaning distinctions and the complexity of verbal lexical relations in continuous semantic space. I then outline the main premise and rationale behind the novel two-phase semantic data collection paradigm introduced in this work, as well as the potential of the produced multilingual resource to support cross-lingual comparisons and in-depth analyses of typological variation. Finally, in Section 1.1.3, I argue that the linguistic information encoded in lexical resources can play an important role in filling the knowledge gaps of large data-driven models. I propose to leverage lexical-semantic information pertaining to verbs by means of a computationally efficient, modular, and cross-lingually portable approach, for the benefit of downstream tasks where accurate verb processing is essential.

1.1.1 Potential of Verb Classes

Viewed as clause governors across many syntactic frameworks (Chomsky, 2014; Kaplan and Bresnan, 1982; Pollard and Sag, 1994; Tesnière, 1959), verbs occupy a prominent role of organisational nuclei in sentence structure. As predicates, they assign properties to entities, their arguments, and relate them to one another, thus determining the sentence's propositional content (Davidson, 1967; Parsons, 1990; Vogel, 2016). Their lexical substance carries information about the type of event taking place, while their morphological features anchor the occurrence they describe in time and can determine the nature of the statement being made (e.g., the indicative mood signalling an affirmation or optative mood expressing a wish) (Lenci, 1998).

Due to their complex linguistic properties, verbs pose a particular challenge to machine interpretation of sentence meaning, where a lot of weight is assigned to them

as carriers of information. This is why accurate, nuanced analysis and representation of verbs' semantic and syntactic behaviour is especially important for NLP systems to get closer to human levels of language understanding (Altmann and Kamide, 1999; Ferretti et al., 2001; Jackendoff, 1972; Levin, 1993; McRae et al., 1997; Resnik and Diab, 2000; Sauppe, 2016). In light of the interplay between semantic and syntactic properties in verbs (Jackendoff, 1990; Levin, 1993), lexical classes organising entries based on shared meaning components and structural features have gained prominence as a powerful model of verb behaviour at the syntax-semantics interface. Their usefulness and value for NLP systems lies in their predictive power. By linking an individual verb to its class, we can abstract away from a single word's occurrence in text and access a bundle of rich semantic-syntactic information pertaining to it. What is more, class membership communicates a verb's affinity in terms of structural behaviour and meaning to other class members, which can remedy the shortage of textual data exemplifying the verb's properties (Kipper et al., 2006) and thus help make accurate predictions about less frequent words and improve generalisability of language models.

The largest such classification currently available for English is VerbNet (Kipper et al., 2006; Kipper Schuler, 2005), which extends and refines the English verb classes of Levin (1993). Including rich semantic and syntactic descriptions for each class in a multi-level hierarchy, VerbNet records variations in the surface realisation of the verb's arguments (i.e., *diathesis alternations*) and maps them to predicate-argument structure, complete with a list of thematic roles assumed by the verbs' arguments and selectional preferences imposed on them. Since its inception, VerbNet has been used to support various NLP tasks, from semantic role labelling and word sense disambiguation to information extraction and question answering, as well as machine translation, automated story generation, and language grounding in robotics (Ammanabrolu et al., 2020; Brown and Palmer, 2012; Clark et al., 2018; Crouch and King, 2005; Kawahara and Palmer, 2014; Lignos et al., 2015; Lippincott et al., 2013; Martin et al., 2018; Raman et al., 2013; Rios et al., 2011; Schmitz et al., 2012; Shi and Mihalcea, 2005; Swier and Stevenson, 2004; Windisch Brown et al., 2011, *inter alia*).

Crucially, the relevance of the VerbNet framework extends beyond English, as similar semantic-syntactic interrelatedness of verb behaviour which underlies Levin classes has been attested in other languages (Dixon, 1991; Fillmore, 1967; Jackendoff, 1990; Levin, 1993; Viberg, 1984). Indeed, its translatability and cross-lingual potential have been studied and empirically verified in a number of works (Hautli and Butt, 2011; Jones et al., 1996; Liu et al., 2008; Majewska et al., 2018b; Mousser, 2010; Pradet et al., 2014; Snider and Diab, 2006; Sun et al., 2010; Vulić et al., 2017b). However, although there

is consensus regarding the importance of building VerbNet-style databases in languages other than English, especially those where high-quality resources are lacking, the task poses significant challenges and databases of similar scale or granularity are extremely rare. While there is compelling evidence that the notion of a semantically-syntactically defined verb class has wide cross-lingual applicability, there is a considerable degree of language specificity in the characteristic syntactic frames and class membership at higher granularity levels, which hinders direct transfer from English. Due to the high expertise and effort involved in manual construction of VerbNet-style lexicons, the possibility of automating the process has been explored (Joanis et al., 2008; Kawahara et al., 2014; Peterson et al., 2016; Scarton et al., 2014; Sun et al., 2010, 2013; Vlachos et al., 2009; Vulić et al., 2017b, *inter alia*), at the cost of accuracy and, in the case of semi-automatic translation-based approaches, faithfulness to the properties of the target language. In light of these challenges and the consequent limited availability of such resources, the potential benefits of verb classes in supporting NLP tasks in other languages and domains are still under-explored.

1.1.2 Acquiring Verb Knowledge from Non-experts

Given the centrality of verbs in sentence structure and the challenge of deriving accurate representations of their meaning and behaviour directly from raw text corpora, time-efficient collection and compilation of verb-focused lexical information is key to enable and accelerate developments in multilingual natural language processing. Rich verb lexicons created by experts provide such information for a small group of the world's languages and require years of expensive lexicographic work. Faster and cheaper crowdsourcing with non-expert native speakers therefore offers an attractive alternative (Snow et al., 2008), if the data collection endeavour can be divided into smaller tasks adequately eliciting human judgments about the properties of linguistic expressions and the relationships between them. While creating detailed, systematically organised lexicons is no doubt beyond the reach of a layperson, the intuitive grasp of commonalities in verb semantic-syntactic behaviour that native speakers possess makes them a promising source of information.

In this work, I explore the possibilities offered by non-expert annotation and the potential and cross-lingual applicability of class-based representations of verb knowledge from two perspectives: *resource creation* and *usefulness for downstream applications*. First, starting from the hypothesis that commonalities in verb meaning are a strong predictor of overlap in syntactic behaviour (Fillmore, 1968; Fisher et al., 1991; Gruber, 1965; Hartshorne et al., 2014; Levin, 2015; Vendler, 1967, 1972), I examine

the possibility of building verb classes starting from semantics alone. Drawing on the observations in previous research that language users are capable of making fine-grained judgments about the combinatorial properties of verbs based on their meaning (Hale and Keyser, 1987; Levin, 1993; White et al., 2014), I investigate the potential of acquiring classifications of verbs from non-expert native speakers. In particular, is it possible to obtain lexical verb classes consistently with minimal guidance, where the only prerequisite for class membership is closeness of meaning? The preliminary experiments on a small sample of languages reveal a promising degree of inter-annotator and cross-lingual alignment, while exposing challenges posed by semantic ambiguity. Naturally, such broad classes, characterised by varying degrees of heterogeneity, are only the first step towards partitioning the verb lexicon: within each there hide further fine-grained distinctions between its members along several dimensions of difference.

I subsequently build on these insights to explore the possibility of leveraging native speakers' introspection to accomplish the task of organising each such broad semantic space to reflect relative similarity and dissimilarity of related words located therein. To this end, I develop a novel two-phase semantic dataset creation paradigm, which uses manual semantic clustering to create broad relatedness-based classes, within which fine-grained semantic similarity judgments are made. The methodology uses a spatial multi-arrangement (MA) approach which was first proposed in the field of cognitive neuroscience for capturing multi-way similarity judgments of visual stimuli. Task participants are presented with a set of items on a computer screen and asked to arrange them in a two-dimensional space through drag-and-drop operations. The challenge in adapting it for the purposes of a linguistic problem lay in the non-trivial issue of polysemy of lexical stimuli, as well as the much larger scale of the undertaking, involving over seven times more items than the stimuli samples used hitherto in MA-based research.

Framing the problem of organising a large lexical sample as two sequential subtasks, clustering and spatial multi-arrangement, has a number of advantages, from the methodological standpoint as well as considering the usefulness of the end product. First, given that one of the main challenges in scaling up any annotation task is the cognitive load on the participants, eliciting fine-grained, relative judgments on hundreds of stimuli simultaneously is unfeasible. The rough clustering phase serves the role of partitioning the starting sample into manageable sets without overburdening users' working memory. However, it is also theoretically motivated, as it ensures 'comparability' of stimuli, a precondition for performing meaningful similarity judgments (Turner et al., 1987). Moreover, presenting each verb in the context of related verbs

helps disambiguate it. During the clustering phase, different senses of a single word end up in different clusters based on their meaning. Thus, in any given cluster, the relevant sense of the word is implied by the surrounding related words, which helps avoid large discrepancies in judgments on ambiguous items.

Secondly, the two-phase protocol holds promise for time-efficient creation of large semantic datasets for use in evaluation of NLP models, with potential to enrich the suite of evaluation resources used to date thanks to its distinctive features. Currently, work on development of representation learning models of lexical semantics usually relies on some form of intrinsic evaluation to ensure that the learned representations reflect human semantic judgments. Being faster and easier to implement, intrinsic evaluation tasks serve as a proxy assessment of representation quality, ahead of the model's deployment in the ultimate downstream application. Lexical semantic similarity estimation is a widely used intrinsic evaluation method, where rankings of similarity scores computed between word embeddings yielded by the model are compared against human ratings of similarity of word pairs (Finkelstein et al. 2002; Agirre et al. 2009; Bruni, Tran, and Baroni 2014; Hill, Reichart, and Korhonen 2015). Typically, human judgments have been elicited on lists of word pairs, whose similarity is expressed as a numerical rating on a discrete scale. A seemingly simple task, similarity estimation is in fact a cognitively complex operation requiring a wealth of conceptual knowledge (Batet et al., 2013; Hill et al., 2015; Mur et al., 2013, *inter alia*).

While the pairwise rating-based annotation design has the undoubted advantage of being easy to implement, which allows for quick ad hoc evaluation data collection, it is known to have a number of limitations (Batchkarov et al., 2016; Faruqui et al., 2016; Gladkova and Drozd, 2016, *inter alia*). Judgments elicited on isolated word pairs one by one are prone to being biased by prototypicality, speed of association, word frequency effects, as well as order of presentation, rather than reflecting semantic considerations. What is more, in practice, subtle differences in word meaning are often very difficult to quantify and transpose onto a discrete numerical scale, especially when no context or reference points are provided. This may result in a certain arbitrariness of assigned similarity scores, and often leads to intra- and inter-rater inconsistencies.

In turn, the potential of the spatial multi-arrangement method to remedy the above limitations lies in its focus on relative, multi-way, continuous similarity judgments. Rather than collecting numerical ratings, users express the relative similarity of words by means of spatial placements, allowing a continuous range. The task builds on the intuitive metaphor of distance in a geometric space as a measure of closeness in meaning, leveraging the spatial nature of the representation of concept similarity in the mental

lexicon (Casasanto, 2008; Gärdenfors, 2004; Lakoff and Johnson, 1999). Crucially, judgments are collected in the context of all other words in the sample, over multiple trials, and fine-grained meaning distinctions can be captured by varying pairwise distances to reflect a word's similarity to all other words present. This has important implications for the method's efficiency and scalability. While in the pairwise rating method the number of possible unique pairings of words grows quadratically as the sample size increases, in the spatial multi-arrangement each placement simultaneously signals the semantic distance of the item to all other items present.

Beyond these practical advantages, the two-phase design yields semantic data which open up interesting analytical possibilities. While the spatial similarity data can be readily used for model evaluation in the traditional pairwise ranking setup, the semantic classes from Phase 1 and the complete distance matrices from Phase 2 allow for evaluating models on the semantic clustering task at different granularity levels. This may consist in either partitioning a large word sample into broad classes, or retrieving narrow clusters of semantically proximate terms from within a set of related words, which can aid development of automatic approaches to building lexical taxonomies. What is more, the method's portability to other languages, which I demonstrate in subsequent chapters, as well as to other parts of speech and types of stimuli, holds promise for efficient collection of evaluation data to support NLP systems regardless of target application or domain. While the word similarity benchmarks available thus far have paid more attention to nouns, the multilingual verb dataset presented in this work addresses the shortage of verb-oriented evaluation data, allowing for assessing the capacity of models to capture the complex linguistic properties of verbs. Comprising two types of semantic data, the resource also offers wide possibilities for in-depth lexical-typological analyses of cross-lingual commonalities and variation in the organisation of lexical fields and investigation of the most salient dimensions underlying human similarity judgments.

1.1.3 Model Augmentation with Structured Lexical Verb Information

Recent years have witnessed major advances in a wide range of NLP tasks owing to the advent of a new generation of self-supervised pretrained contextual encoders (Devlin et al., 2019; Liu et al., 2019d; Radford et al., 2018, 2019; Raffel et al., 2020). In contrast to the earlier, widely used *static* distributional models which yield the same representation for the same word regardless of context (Bojanowski et al., 2017;

Mikolov et al., 2013b; Pennington et al., 2014), the new architectures are sensitive to the contextual dependence of meaning, producing different representations across different occurrences of the word in text. Notably, they have been used with impressive results in cross-task knowledge transfer, first pretrained on large volumes of unlabelled data through language modeling, and subsequently fine-tuned on labelled data in a supervised fashion to tackle a target task of choice. While their performance across various natural language understanding benchmarks established the new state of the art and garnered an immense interest in the research community, a growing body of work has focused on investigating the exact reasons behind their success. Although a number of analyses have revealed that pretraining on large text corpora allows them to capture a range of linguistic information, for instance, pertaining to syntactic structure and semantic roles (Ettinger, 2020; Tenney et al., 2019b), they are not free from the limitations of their static predecessors. In particular, their insights into the properties of language and the wider extralinguistic reality expressed by its means are purely distributional in nature. In consequence, they are still prone to fall into the trap of mistaking topic relatedness for lexical semantic similarity (Lauscher et al., 2020b) and lack basic factual knowledge about the world (Peters et al., 2019; Zhang et al., 2019b).

To remedy these deficiencies, a number of recent works investigated the possibility to supplement self-supervised encoders with external information. While static representation models lend themselves well to different forms of *specialisation*, where a particular type of information, e.g., a lexical relation, is accentuated in the static embedding space by means of lexical constraint injection (Faruqui et al., 2015; Mrkšić et al., 2017; Ponti et al., 2018, 2019; Wieting et al., 2015), the same methods cannot be directly applied to the contextual encoders. Instead, work on knowledge augmentation of pretrained architectures explored the possibility of adding additional pretraining objectives based on external knowledge bases (Lauscher et al., 2020b; Nguyen et al., 2016a), or alternatively, using such objectives to fine-tune the parameters of the pretained model (or a small set of additional parameters injected for this purpose) post hoc (Lauscher et al., 2020a; Liu et al., 2019b; Peters et al., 2019; Wang et al., 2020a; Zhang et al., 2019b), where the latter dispenses with the computationally expensive retraining of the entire network from scratch.

The final chapter of the present work picks up this research thread and investigates an area yet unexplored in the context of the state-of-the-art pretrained models, i.e., augmentation with verbal lexical knowledge. Its first guiding question is: Can the large self-supervised encoders still benefit from structured lexical information about verbs' semantic-syntactic behaviour? Specifically, I investigate whether fine-grained knowledge

of commonalities in verbs' meaning and argument-taking properties can boost the ability of pretrained encoders to reason about events and their participants, a skill crucial in applications involving processing of narratives and dialogue (Carlson et al., 2002; Miltsakaki, 2009), as well as grounding texts in real-world facts (Doddington et al., 2004). Across languages, events are prominently expressed by means of verbs and their arguments, the former encoding the nature of the occurrence and the latter its participating entities. Structured sources of information about the syntactic patterns exhibited by verbs and the semantic frames which they evoke – such as the above discussed VerbNet, or FrameNet (Baker et al., 1998), grounded in the theory of frame semantics (Fillmore, 1976, 1977, 1982) – may therefore provide a useful supplementary signal to help data-driven models tackle event-focused problems. To put this hypothesis to the test, I investigate approaches to verb knowledge injection based on an auxiliary objective consisting in identifying verbs which share membership in the same class or evoke the same semantic frame, comparing the effectiveness of storing this added knowledge in a dedicated self-contained set of parameters or updating the parameters of the entire model. Evaluation on two event annotation frameworks and the tasks of event trigger and argument identification and classification show that exposing the model to a verb-specific signal indeed helps downstream performance, which holds promise for aiding other NLP applications where accurate verb processing is paramount.

Further, in light of the above discussed shortage of high-quality, high-coverage verb lexicons in the majority of the world's languages, I explore the potential to extend the benefits of structured verb-focused lexical information to under-resourced languages by means of cross-lingual knowledge transfer. To this end, I examine two approaches: *direct model transfer* and *annotation transfer*. In the former, I leverage a massively multilingual encoder (Devlin et al., 2019) pretrained on over a hundred languages and perform zero-shot transfer, where the model is trained on source language labelled data and boosted with source language verb knowledge before making predictions in the target language, without observing any annotated examples in that language. In the latter, I continue the exploration of a theme central to this thesis, the cross-lingual potential of verb classes and frames (Jackendoff, 1990; Levin, 1993). I investigate their direct portability across typologically diverse languages through automatic lexical constraint transfer, which employs a relation prediction model trained on the verb class and verb frame data available in the source language to clean the automatically translated constraints, without any manual target-language adjustments. Results on the chosen event processing tasks show that both approaches succeed in supporting the underlying model in solving the problem at hand. These findings have

interesting implications for one of the important theoretical linguistic debates, that is, to what extent can we consider semantically-syntactically defined verb classes and predicate-argument structures universal, rather than cross-lingually variable (Brown and Bowerman, 2008; Hartmann et al., 2013). Despite the noise introduced in the automatic constraint transfer, there is a non-negligible amount of knowledge about commonalities in verbs' semantic-syntactic patterns of behaviour which transcends language boundaries and benefits target language processing.

Finally, having validated the utility of injecting expert-curated knowledge about verbs' properties into model parameters, I investigate whether the native-speaker expertise of untrained language users can be harnessed for the same purpose. Considering the language-specificity of verb class membership, notwithstanding high-level cross-lingual parallels, I evaluate whether the semantic classes and semantic similarity data collected in this work can supplement the models' distributional knowledge with useful information and benefit their event processing ability. This analysis sheds light on another important question: Do non-expert judgments about verb meaning encode valuable information beyond what can be automatically derived from large text corpora? The final experiments of this thesis show that there is indeed potential to leverage non-expert data as a source of non-distributional verb knowledge at different levels of granularity. Most importantly, their contribution compares favourably to direct cross-lingual transfer of expert knowledge from English to the target language. This reveals that native-speaker judgments capture important language-specific information unavailable in the automatically translated and refined source language lexical constraints. This finding, further corroborated in an evaluation focused on the injection of high-quality but low-coverage information from a small target language database, reaffirms the value of dedicating efforts to the construction of language-specific resources and their potential to enrich state-of-the-art NLP models.

1.2 Main Contributions

The thesis aims to address the continuously growing need for language resources in multilingual NLP to evaluate and support distributional models in tasks requiring sophisticated linguistic knowledge, and empirically investigates the potential to infuse said knowledge into the parameters of state-of-the-art architectures in resource-rich and resource-lean scenarios. Focusing on the category of verbs where shortage of high-coverage lexicons has been particularly pronounced, this work makes the following original contributions:

- **Semantically driven classification.** Drawing on the interrelatedness of verbs' semantic and syntactic behaviour, I investigate whether verb classes can be consistently obtained from non-expert native speakers in a bottom-up fashion by 'bootstrapping' from semantic information. The resultant verb clusterings show a promising degree of inter-annotator alignment, suggesting grouping verbs based on their meaning can be used as a starting point to build verb classifications. Due to the strong semantic-syntactic mapping in verbs, a degree of overlap in syntactic patterns is expected, but it does not explicitly determine class membership. Thanks to its semantic focus, the approach holds promise for facilitating the resource creation process, given the greater ease of the task for untrained subjects compared to syntactic analysis, and generating cross-lingual mappings, in light of the universal nature of many components of verb meaning.

- **Novel dataset creation paradigm.** To overcome the bottleneck of slow and expensive resource creation, I design a novel two-phase semantic dataset construction paradigm. To this end, I adapt a spatial arrangement approach previously used exclusively in visual perception studies with concrete real-world objects to ambiguous lexical stimuli and a sample over seven times larger than those hitherto used in cognitive research. In the first phase, broad relatedness-based classes are created in a rough semantic clustering task, and polysemy is handled through label copying. These classes are subsequently fed into the second phase, where fine-grained semantic similarity judgments are made by means of spatial arrangements in a two-dimensional space. The approach elicits nuanced and continuous similarity judgments on related words, which ensures comparability of the concepts and provides disambiguating contexts in the form of other words appearing in the space.

- **Multilingual verb resource.** To address the issue of shortage of large-scale verb-focused evaluation resources, I employ the above described methodology to produce the first large-coverage dataset targeting verb semantics in a typologically diverse selection of languages, English, Mandarin Chinese, Japanese, Italian, Finnish, and Polish. The resource includes semantic classes and over 20k fine-grained pairwise similarity scores for each language. The scale and broad coverage of the dataset enables, for the first time, evaluation of representation learning models across a number of domains of verb meaning. Further, the resource's typological diversity allows for in-depth examination of commonalities

and patterns of variation in verb behaviour and the organisation of semantic fields across different language families.

- **Verb-focused intrinsic evaluation.** I carry out extensive evaluation of representation learning models on the multilingual semantic data, comparing the capacity of two types of architectures – static word embeddings and Transformer-based pretraining models – to capture word-level semantics in two tasks, semantic clustering and word similarity. This evaluation reveals the primacy of static word embeddings over the lexical representations extracted from their contextualised counterparts, but the improvements gained by averaging the latter representations over multiple contexts show there is scope for extracting more powerful word-level vectors by aggregating multiple token-level embeddings.

- **Verb knowledge injection for event processing.** I investigate whether discrete lexical-level information concerning verbs' semantic-syntactic behaviour can help mitigate the limitations of state-of-the-art representation learning architectures and improve their performance in downstream NLP tasks. To this end, I propose an approach to inject verb knowledge into contextualised self-supervised pretraining models that avoids computationally expensive retraining from scratch and enables easy integration with other types of information. Improvements in event processing tasks demonstrate that complementing the model's distributional knowledge with external verb-specific information indeed boosts its capacity to understand events and their structure. Further, I examine the potential of leveraging rich verb-related knowledge available in well-resourced languages to support NLP models in languages lacking such resources. I achieve this by means of two alternative approaches, either using the discrete verb knowledge available in the source language to support zero-shot transfer, or automatically transferring this knowledge to support monolingual fine-tuning on target language task data. Performance gains yielded by the two techniques suggest that both hold promise for supporting event processing in under-resourced languages, while providing empirical evidence for the existence of a strong cross-lingual component in verb classes and semantic frames. Finally, I demonstrate the potential of harnessing non-expert knowledge about verb meaning by employing the semantic data generated as part of this thesis for verb-oriented fine-tuning. The results reveal that untrained native-speaker judgments encode valuable language-specific knowledge unexploited in direct cross-lingual transfer and can provide useful guidance in the absence of expert-curated resources.

1.3 Thesis Outline

The remainder of the thesis is organised in the following chapters, briefly summarised:

- Chapter 2 provides theoretical background from the area of verb lexical semantics and surveys most relevant related work on representation learning model evaluation, resource creation, and model augmentation with external knowledge.

- Chapter 3 examines the potential of creating verb classes through soft manual verb clustering based on shared semantics. The same starting sample, translated into the target language, is clustered in English, Polish, and Croatian, which enables comparisons of emerging classifications in languages within the same language family (Slavic) and from different language families. The analysis of inter-annotator agreement shows an encouraging degree of overlap in the classifications produced for each language individually, as well as across all three languages. This suggests that verbs can be reliably classified by native speakers without linguistics training, and that there is potential to create verb classifications starting from a simple, purely semantic task. Moreover, the cross-lingual overlap demonstrates that there are cross-linguistic commonalities and shared meaning components governing the semantic organisation of verbs.

- Chapter 4 presents and motivates the novel two-phase semantic dataset creation paradigm, providing a detailed description and analysis of each phase and the unique benefits offered by the produced data. I scrutinise the clusters emerging from Phase 1 by means of a network analysis approach and analyse patterns of agreement and ambiguity in the spatial similarity judgments collected in Phase 2. Next, I carry out an in-depth analysis of the properties of the newly created dataset, SpA-Verb, by means of a quantitative and qualitative comparison with several existing lexicons and datasets. This analysis shows an encouraging degree of overlap between the semantic data from Phase 2 and VerbNet, suggesting potential for creation of similar datasets bottom-up in languages lacking such resources. Moreover, the comparative analysis of the spatial and pairwise-rating paradigms highlights their key differences and the evaluation potential of SpA-Verb beyond what is offered by traditional datasets. The subsequent extensive evaluation of a diverse selection of representation learning architectures on the new dataset reveals it to be a challenging benchmark, while demonstrating that systems drawing on external linguistic knowledge are especially capable of capturing fine-grained meaning distinctions across semantic domains.

- Chapter 5 examines the cross-lingual applicability of the two-phase paradigm to typologically diverse languages, Mandarin Chinese, Japanese, Italian, Finnish, and Polish. I discuss the language-specific factors which need to be taken into account when porting the method to languages distant from English and describe the design choices made to accommodate them. Subsequently, I analyse the semantic verb classes created in Phase 1 and the considerations impacting classification decisions in each language. Aligning the resultant clusters reveals substantial cross-lingual overlap in the semantic areas identified by the participants, which suggests that they relied on similar classification criteria across languages. Subsequently, I carry out comparative analyses of the semantic spaces encoded in the representational dissimilarity matrices in the six languages studied. I then use the multilingual resource to examine the capacity of static and contextualised word embeddings to capture the subtle semantic distinctions encoded in the spatial similarity data across languages and domains of verb meaning, as well as their ability to construct semantic verb classes from scratch.

- In chapter 6, I empirically examine the potential of structured lexical verb resources to support state-of-the-art NLP models in tasks where accurate verb processing is key. I introduce an approach to incorporating discrete knowledge about verb behaviour into large pretrained encoders, offering the benefits of modularity and computational efficiency. Subsequently, I evaluate the method on the tasks of event trigger and argument identification and classification in English, drawing on the information stored in two lexicons, VerbNet and FrameNet. The results demonstrate that verb knowledge injection has a positive impact on performance. I then explore the potential to extend the benefits of the approach to other languages by means of two cross-lingual transfer methods, either leveraging the information available in English directly in a zero-shot transfer setup, or automatically transposing it into the target language to support monolingual downstream fine-tuning. The experiments show that verb knowledge transfer can indeed boost event processing in resource-lean languages. Finally, I demonstrate the potential of employing the semantic data generated as part of this work as an alternative source of verb-related information, revealing the advantages of non-expert in-target data over automatic transfer of expert knowledge.

- Finally, in chapter 7 I discuss the implications of the results of experiments and analyses included in this thesis, and reflect on the avenues for future research which could extend the investigations pursued in this work.

1.4 Publications

All experiments and analyses, as well as the conceptual design of the presented two-phase data collection methodology, were carried out by the author. The web implementation of the spatial multi-arrangement method (Phase 2) used for online data collection is by Nikolaus Kriegeskorte, Jasper van den Bosch, and Ian Charest of Meadows Research (www.meadows-research.com). Jasper van den Bosch integrated new custom features into the online software at the request of the author to adapt the method to text stimuli and accommodate soft clustering (Phase 1). Ivan Vulić assisted in deriving word-level representations from Transformer-based models evaluated by the author in Chapters 4 and 5, Goran Glavaš produced Figure 6.1. I am grateful to all co-authors for their suggestions and help. The thesis includes material from the following papers:

- **Acquiring Verb Classes Through Bottom-Up Semantic Verb Clustering.** [Ch. 3]
  Olga Majewska, Diana McCarthy, Ivan Vulić, Anna Korhonen. 2018. *Proceedings of LREC*, pages 952–958.

- **Spatial Multi-Arrangement for Clustering and Multi-way Similarity Dataset Construction.** [Ch. 4]
  Olga Majewska, Diana McCarthy, Jasper van den Bosch, Nikolaus Kriegeskorte, Ivan Vulić, Anna Korhonen. 2020. *Proceedings of LREC*, pages 5749–5758.

- **Semantic Dataset Construction from Human Clustering and Spatial Arrangement.** [Ch. 4]
  Olga Majewska, Diana McCarthy, Jasper van den Bosch, Nikolaus Kriegeskorte, Ivan Vulić, Anna Korhonen. 2021. *Computational Linguistics* 47.1, pages 69–116.

- **Manual Clustering and Spatial Arrangement of Verbs for Multilingual Evaluation and Typology Analysis.** [Ch. 5]
  Olga Majewska, Ivan Vulić, Diana McCarthy, Anna Korhonen. 2020. *Proceedings of COLING*, pages 4810–4824.

- **Verb Knowledge Injection for Multilingual Event Processing.** [Ch. 6]
  Olga Majewska, Ivan Vulić, Goran Glavaš, Edoardo M. Ponti, Anna Korhonen. 2021. *Proceedings of ACL*, pages 6952–6969.

Chapter 2

Background

The aim of this chapter is to provide background information relevant to the investigations carried out in this work, as well as place it within a larger context of previous research. The discussion is divided into three parts. First, I introduce the framework of lexical semantics and discuss approaches to organising verb knowledge in structured form. Next, I give a brief overview of the developments in representation learning and review methods for creating intrinsic evaluation datasets. Finally, I discuss the need for augmenting distributional models with external lexical knowledge and outline some of the recent techniques used for this purpose.

2.1 Lexical Semantics and NLP

Very little is predetermined about the names that languages assign to concepts. Lexicons arise spontaneously and heterogeneously (Ralph, 1980), and aside from small areas of form-meaning systematicity (e.g., onomatopoeic expressions) (Monaghan et al., 2011), there is no direct, natural association between word forms and the concepts they stand for (Greenberg, 1957; Saussure, 1916). Acquiring word meaning is crucial to communicate in a language, and indeed, children learn to label the salient elements of the surrounding reality with words before developing any awareness of how they can be combined to form sentences (Clark, 2009; Wolf and Stoodley, 2008).

In linguistic theory, the lexicon of a language has been viewed as the record of all the idiosyncratic information about its elements, which can be arranged according to the finite set of rules of grammar to express an unlimited number of meanings (Beaugrande, 1991; Bloomfield, 1933; Chomsky, 1965). However, views on how much information is contained in the lexicon and how much of what we know about concepts makes up the meaning of the words which denote them (Geeraerts, 2010) vary across

linguistic theories. For structuralist and formal approaches to lexical semantics, lexical meaning is viewed in terms of the relations of similarity and contrast between the meanings of words present in the lexicon (Bosch, 1988; Lyons, 1968; Saussure, 1916). Knowing a word means knowing its place in the larger system and its relations to other elements of that system. For more recent cognitive approaches, focusing on language use rather than on the language system, word meaning is highly contextual and flexible (Langacker, 2008; Wray, 2015). It emerges through interaction with a wider, extra-linguistic reality, with words acting as entry points to the speaker's conceptual knowledge (Dancygier, 2017).

The investigations carried out by lexical semanticists concern the problems of representing, decomposing and classifying word meaning and accounting for its contextual variability. Moreover, they involve the examination of relations between words and their senses and the relation between lexical meaning and its realisation in the syntactic structure. This latter inquiry, in particular, has provided evidence in support of a wider view of the lexicon than merely a record of the minimum idiosyncratic information specific to each word. In fact, speakers' knowledge of lexical items transcends word boundaries and allows them to make nuanced judgments about their combinatorial properties even in completely novel syntactic contexts and expressions (Levin, 1985). This interrelatedness of lexical meaning and syntax is particularly conspicuous in the case of verbs, whose meaning has been shown to provide essential cues as to the realisations of their arguments and their possible alternative arrangements.

The attention given to the issues pertaining to the lexicon and the inquiry into lexical meaning has grown together with the shift in linguistics towards more lexically-driven theories of grammar, and its importance has long been recognised in the field of natural language processing. Indeed, in light of the centrality of lexical semantic knowledge for the majority of NLP tasks, much seminal work on lexical semantics has been carried out within the computational linguistics framework, including the development of large-scale digital lexicons and influential models of meaning such as WordNet (Fellbaum, 1998). Computational lexical databases have since supplied NLP systems with detailed information about words' semantic and syntactic behaviour, as well as morphological and phonetic properties, at different levels of granularity. For instance, information about semantic roles taken by verbs' arguments has been leveraged in semantic parsing (Giuglea and Moschitti, 2006) and coreference resolution (Bejan and Harabagiu, 2010; Rahman and Ng, 2011). Whereas knowledge of lexical-semantic relations between words has been shown to help word sense disambiguation (Vial et al., 2018) and lexical entailment (Vulić and Mrkšić, 2018; Vulić et al., 2019).

In parallel to the expert knowledge-driven approach to analysing and organising word meaning, lexical semantic information has been acquired automatically from text corpora and encoded in human- or computer-readable form. Data-driven automatic lexical acquisition methods have offered the advantages of speed and large coverage. However, they are prone to error and fall short of capturing fine-grained meaning distinctions encoded in expert-curated knowledge bases. As the advances in language modeling unlock overwhelming potential to learn about meaning in a completely unsupervised fashion, the precision and high granularity of structured lexicons make them highly valuable as complementary sources of information and reference benchmarks (Shwartz et al., 2015).

In what follows, I will discuss different approaches to the organisation of lexical knowledge in structured form, starting with two prominent expert-built semantic resources employed in the analyses included in this work (2.1.1). Next, I will shift the focus to the lexical semantics of verbs specifically, discussing the theoretical perspectives on the interrelatedness of syntactic behaviour and meaning of verbs (2.1.2). I will subsequently introduce two key resources built on these foundations, Levin's English verb classes (Levin, 1993) and VerbNet (Kipper Schuler, 2005). Finally, I will discuss approaches to creating verb classes automatically and deriving them directly from non-expert language users.

2.1.1 Semantic Databases

WordNet

Among the most prominent structured lexical resources is the English WordNet (Fellbaum, 1998; Miller, 1995). It is a large hierarchical semantic network that organises concepts into over 117 thousand unordered synonym sets (the so-called *synsets*),[1] drawing on psycholinguistic and computational models of human lexical memory. The primary relation in WordNet, (cognitive) synonymy, links words which denote the same concept and can be used interchangeably across different contexts, e.g., *let, allow, permit* or *accurate, exact, precise*. Synsets, in turn, are interconnected by several other conceptual-semantic and lexical relations. Hypernymy/hyponymy links specific concepts to more general sets and superordinate categories across several taxonomy levels (e.g., *vehicle – wheeled vehicle – bicycle – tandem*). Meronymy/holonymy holds between concepts denoting component parts of other entities (*mother board – central processing unit – computer*). Antonymy links individual contrasting word senses (e.g.,

[1] Version 3.1 of WordNet contains 155,287 words organised in 117,659 synsets.

light – heavy, open – close), whereas troponymy connects verb synsets across different degrees of specificity in the manner characterising a given action or event denoted by the verbs in the set (e.g., *cook – fry – sauté*). Importantly, WordNet operates at the level of specific senses, rather than word forms. Thus, it provides a comprehensive inventory of meanings a word can take, ordered by their frequency, a feature which is especially useful in the task of word sense disambiguation.

The wide usefulness of WordNet across many NLP tasks and applications (Mandala et al., 1998; Mihalcea and Moldovan, 2000; Moldovan and Mihalcea, 2000; Ngo et al., 2018; Vijayarajan et al., 2016; Wei et al., 2015; Wiebe et al., 1998) has inspired analogous project in other languages. These include Japanese (Isahara et al., 2008), Italian (Toral et al., 2010), Chinese (Huang et al., 2010), Polish (Vetulani et al., 2010), Arabic (Black et al., 2006), and German (Hamp and Feldweg, 1997), as well as initiatives aimed at unifying or extending language-specific resources into large multilingual knowledge bases (Atserias et al., 2004; Jansen, 2004; Navigli and Ponzetto, 2010; Stamou et al., 2002; Vossen, 1998). English WordNet has also served as the source of lexical relation knowledge in much recent research on semantic specialisation of distributional representation models (e.g., Faruqui et al., 2015; Glavaš and Vulić, 2018b; Lauscher et al., 2020b; Ponti et al., 2019; Vulić et al., 2018; Vulić and Korhonen, 2018; Vulić and Mrkšić, 2018). In this thesis, it is employed for qualitative and quantitative analysis of the fine-grained semantic relationships encoded in the output of the spatial multi-arrangement method presented in Chapter 4.

FrameNet

In contrast to WordNet's paradigmatic approach, where word senses are grouped into synsets based on their mutual substitutability, another prominent English lexical resource, FrameNet (Baker et al., 1998), has a syntagmatic focus. Rooted in the theory of Frame Semantics (Fillmore, 1976, 1977, 1982), it adopts a model where the meaning of a word is captured within the context of a prototypical situation,[2] a frame, in relation to the other entities participating in it (i.e., frame elements). For example, the situation of telling usually involves a person (Speaker) addressing another person (Addressee) with a message (Message) about something (Topic). While the concept of losing typically involves someone (Owner) losing something (Possession) that belongs to them. The word senses which may evoke each of these situations, e.g., *notify* and *inform* in the first and *lose* and *misplace* in the second scenario, are the *lexical units* of the 'Telling' and 'Losing' frame, respectively. Each frame provides

[2]Aside from events, frames can also represent relations, states or entities.

examples illustrating the surface realisation of the frame elements (in other words, semantic roles characteristic of the frame) in relation to its evoking predicate (TOLD in the example below), based on corpus evidence, e.g.:

[SPEAKER Leroy] TOLD [ADDRESSEE his mother] [TOPIC about his arrest].

Although the sets of lexical units listed for each frame comprise different parts of speech, verbs are the prototypical frame-evoking words, naturally lending themselves to a frame-based representation consisting of a predicate and its arguments. While the frames are semantically, rather than syntactically, motivated, the annotated example sentences drawn from the British National Corpus (BNC Consortium, 2007) supply rich information about the combinatorial properties of the lexical units associated with them. Beyond the relations between individual frame elements, FrameNet also records relationships between frames themselves. For example, the relation of inheritance applies where a more specific child frame elaborates a more general parent frame (e.g., the 'Telling' frame inherits from the 'Statement' frame), whereas the relation of inchoativity holds between a stative frame and the inchoative frame that refers to it and denotes a change of state (e.g., the 'Becoming_dry' frame is inchoative of 'Being_dry').

The semantic orientation of the FrameNet database makes it amenable to cross-lingual extensions, as many of the 1224 frames[3] it currently includes have their analogs in languages other than English. For instance, the notion of cooking captured in the 'Apply_heat' frame is likely to involve the same core elements in different languages, that is, a person doing the cooking, the food being prepared, a source of heat or the container holding the food. Indeed, projects to construct counterparts of the English FrameNet or frame-annotated corpora have been undertaken in Japanese (Ohara, 2012), Swedish (Heppin and Gronostaj, 2012), Spanish (Subirats and Sato, 2004), Danish (Bick, 2011), Chinese (Liping You and Kaiying Liu, 2005), French (Candito et al., 2014), Korean (Kim et al., 2016a), and German (Boas, 2002), also including annotations for selected domains (e.g., tourism and football in Brazilian Portuguese (Torrent et al., 2014)) and an ongoing Multilingual FrameNet initiative (Torrent et al., 2020). The experiments reported in Chapter 6 further corroborate the wide cross-lingual applicability of the FrameNet paradigm, demonstrating that there is scope to directly port knowledge of verbs' frame-evoking properties from one language to another to boost both cross-lingual transfer and monolingual event processing in the target language.

[3]The frames in turn include the total of 13,685 lexical units in the version 1.7 of the resource.

2.1.2 Verb Classes and the Syntactic-Semantic Interface

The inquiry into the interplay of words' semantic and syntactic properties spans several fields of study, including psycholinguistics, cognitive science, and lexical semantics. Research in child language acquisition has explored the semantic-syntactic link in the context of the theory of bootstrapping, which aims to explain children's extraordinary ability to learn a language by means of innate mental strategies that help initiate the learning process, leveraging systematic correspondences between different levels of linguistic structure. According to the syntactic bootstrapping hypothesis, the syntactic frames in which verbs appear help children constrain their possible meanings. Their ability to implicitly associate certain words with certain syntactic categories (possibly aided by acoustic cues in the speech signal (Gleitman and Wanner, 1982)), is thought to assist a first language learner in making inferences about words' semantics, despite the limited extralinguistic cues available (Gleitman, 1990; Gleitman et al., 2005). Faced with recurring structures like *The cat drank the milk. Dad is feeding the dog*, the child classifies *the cat* and *dad* as 'the doers', and *drink* and *feed* as something being done to *milk* and *dog*. The likely meaning of words is thus partly derived from the structural slots which they fill, as observed in early experiments eliciting children's interpretations of nonsense words (Brown, 1957).

Of course, syntactic frames alone do not eliminate ambiguity. As argued by the advocates of semantic bootstrapping, according to which learning about words' semantics from extra-linguistic context *first* helps children acquire syntax, syntactic frames may help narrow down that a verb is referring to some action, but can hardly provide enough information about its specific meaning (Pinker, 1994). Both perspectives have faced challenging evidence and indeed, it has been argued that rather than being mutually exclusive, they are two techniques available to a child learner at different stages of language acquisition (Pinker, 1994). Nonetheless, there is an emerging consensus about the existence of a natural link between meaning and function, stable enough to support the learning process, and the importance of their interplay in verb acquisition (Fisher et al., 1991; Gleitman, 1990; Lederer et al., 1995). The conclusions from this line of research have powerful implications beyond child language acquisition: even if their relationship is complex, we can view verbs' syntax as a regular projection from their semantics (Fisher et al., 1991). We should therefore expect the degree of verbs' semantic similarity to be mirrored in the overlap in the syntactic structures which they permit, and we should see verbs representing a given type of action or event select similar kinds of arguments.

Levin's Verb Classes

Semantic-syntactic mappings have been the subject of study in a body of linguistic research (Fillmore, 1965, 1968; Goldberg, 1995; Green, 1974; Gropen et al., 1989; Gruber, 1965; Hartshorne et al., 2014; Jackendoff, 1972, 1983; Levin, 1993, 2015; Pinker, 1989; Talmy, 1985; Vendler, 1967, 1972; White et al., 2014, *inter alia*). Crucially, these correspondences, leveraged by the child learner, are also what allows adult native speakers to make subtle judgments about grammaticality of sentences and arguments licensed by any given verb, whether already known, novel, or made up (Hale and Keyser, 1987; Levin, 1985; Zwicky, 1971).[4] Based on the assumption that native speakers' predictions about verb combinatorial behaviour and sensitivity to alternations in verbs' valency and permitted arguments (i.e., *diathesis alternations*) are possible thanks to the existence of reliable links between particular syntactic properties and verbs representing a certain semantic type, Levin (1993) stipulated that verbs which cluster together based on their behaviour will also exhibit shared semantics. For example, verbs such as *break, split, tear* and *rip* all take part in the causative/inchoative (1) and middle alternation (2), but do not permit the simple reciprocal alternation (3):

(1) a. I broke the twig off (of) the branch.
The twig broke off (of) the branch.
b. I broke the twig and the branch apart.
The twig and the branch broke apart.

(2) a. I broke the twigs off (of) those branches.
Twigs break off of those branches easily.
b. I broke those twigs and branches apart.
Those twigs and branches break apart easily.

(3) a. The twig broke off (of) the branch.
b. * The twig and the branch broke.
with the intended meaning of (a)

All the verbs participating in these alternations share an extended meaning which involves 'separating by V-ing', where 'V' denotes the basic meaning of the verb (Levin,

[4]Zwicky (1971) illustrates this with an invented verb *greem*: knowing that it means 'to communicate verbally with a particular voice quality (hoarse and loud)' tells an English native speaker that it is possible 'to greem' (i.e., to speak loudly), 'to greem for someone to get you a glass of water', or 'to greem about the price of doughnuts'. In short, a native speaker has an intuitive sense of the acceptable usages of the novel verb based on its semantics.

1993), giving rise to the class of *Split* verbs. Notably, many of the members of this class simultaneously belong to other classes (e.g., *Break* verbs), manifesting a different predominant sense in a different set of alternations, and may not have any inherent meaning component of 'separating' outside these particular syntactic contexts (Dang et al., 1998). Using diathesis alternations as the main class membership criteria (along with morphological and subcategorisation properties), Levin manually constructed a semantic-syntactic taxonomy of 3,024 English verbs. Her verb lexicon includes 48 broad and 192 fine-grained classes, each characterised by a set of alternations in which member verbs participate, and remains one of the most widely used English resources of this kind in NLP.

While Levin's work concerns English verbs only, the rationale behind it is thought to have cross-lingual applicability. The evidence in support of the existence of a similar semantic-syntactic interplay exhibited by verbal alternations has been found in languages both typologically close and distant from English. For example, the conative alternation (where an action described by the verb is attempted but not necessarily completed, e.g., *The woodchopper hacked the tree – The woodchopper hacked at the tree*) displays analogous patterns in English and Warlpiri (Guerssel et al., 1985; Laughren, 1988), a language belonging to the Pama-Nyungan family spoken in Australia. Verbs which can participate in it belong to the *cut* and *hit* type (e.g., *saw, slash, bash, kick*), while those belonging to the *break* and *touch* type do not permit it. According to a number of investigations of this phenomenon (Fillmore, 1967; Guerssel et al., 1985; Hale and Keyser, 1986, 1987), these two sets of verbs share meaning components which determine whether or not they license the conative construction, i.e., the components of motion and contact (Guerssel et al., 1985) (both are absent in *break*, while *touch* only involves the latter). Further examples of parallels traversing language families and morphological alignment types are found in English and an ergative language spoken in northeastern Siberia, Chukchi, with regards to the causative-inchoative alternation (Nedjalkov, 1976), which appears in both languages with *fill*-type verbs (Baker and Bobaljik, 2017):

(4) a. ətləg-e jərʔen-nin əʔtvʔet miml-e
father-ERG fill-3SG>3.SG boat.ABS water-INSTR
'Father filled the boat with water.'

b. əʔtvʔet jərʔet-gʔi miml-e
boat.ABS fill-3SG water-INSTR
'The boat filled with water.'

Additional similarities are found in the conative and locative alternations, where the syntactic expression of a given argument varies between a prepositional phrase and a noun phrase (e.g., *Mum smeared butter on the bread – Mum smeared the bread with butter*) (Baker and Bobaljik, 2017; Nedjalkov, 1976).

Despite the differences in verb and alternation inventories across languages, the same meaning components are thought to serve as criteria determining verbs' behaviour and the expression of their arguments. In other words, the patterns of verb behaviour tend to be sensitive to the same set of aspects of verb meaning. In this thesis, the analyses of the organisation of concepts within broad semantic categories based on spatial similarity judgments of verb meaning in Chapter 5 show cross-lingual commonalities concerning the most salient meaning components underlying the arrangements, suggesting speakers of different languages implicitly rely on similar meaning dimensions when organising related verbal concepts.

VerbNet

Levin's model of verb classification gave rise to the largest computational English verb lexicon currently available, VerbNet (Kipper et al., 2006; Kipper Schuler, 2005). Using the correspondence between verbs' syntactic and semantic properties as a classification criterion, VerbNet extends and refines Levin's original classes yielding a fine-grained taxonomy comprising nearly 3,800 verb lemmas across several granularity levels. Each of the 274 first-level classes groups together verbs which share certain meaning components and patterns of syntactic realisation of their arguments. For instance, the class 'performance-26.7' includes verbs *chant, rap, vocalize* and *improvise* which participate in three syntactic frames with the following types of arguments:

- NP V NP
  Sandy sang a song.
  Agent V Theme

- NP V
  Sandy sang.
  Agent V

- NP V NP PP.BENEFICIARY
  Sandy sang a song for me.
  Agent V Theme {for} Beneficiary

Further, a subset of members (e.g., *sing, whistle, hum*) license the benefactive alternation, that is, can appear with a beneficiary argument realised as part of a prepositional phrase headed by *for* (above) as well as with one realised as an immediately postverbal noun phrase:

- NP V NP.BENEFICIARY NP
  Sandy sang me a song.
  Agent V Beneficiary Theme

Verbs which display this alternating behaviour form a subclass ('performance-26.7-1'), which inherits the basic set of frames from the parent class, thus producing a hierarchical structure. Beyond the syntactic frames and the associated semantic predicates, each class lists the thematic roles and selectional restrictions (e.g., *animate, organization*) for the verbs' arguments. Further, VerbNet provides mappings to other lexical databases including WordNet, FrameNet, and Xtag (XTAG Research Group, 2001).

VerbNet classes act as generalised representations of groupings of verbs characterised by shared semantic-syntactic properties and enable inferring rich lexical information about them based on class membership. Their predictive power has benefited a number of NLP applications, including semantic role labelling (Giuglea and Moschitti, 2006; Hartmann et al., 2016; Swier and Stevenson, 2004), word sense disambiguation (Dang, 2004; Kawahara and Palmer, 2014; Popov et al., 2019; Windisch Brown et al., 2011), semantic parsing (Shi and Mihalcea, 2005), information extraction (Schmitz et al., 2012), question answering (Clark et al., 2018) and text mining (Lippincott et al., 2013; Rimell et al., 2013), as well as automated story generation (Ammanabrolu et al., 2020; Martin et al., 2018) and detection (Eisenberg and Finlayson, 2017). However, its utility in supporting NLP systems has been constrained by the very limited availability of similar resources in other languages and domains, due to the extended lexicographic effort and financial expense entailed by fine-grained manual classification work at such large scale. At present, manually built VerbNet-style resources can be found in Arabic (Mousser, 2010), Spanish and Catalan (Aparicio et al., 2008), and Czech (Pala and Horák, 2008), while the resource available in Mandarin Chinese (Liu et al., 2008) includes a class hierarchy based on frame semantics.

Automatic and Semi-automatic Verb Classification

Given that manual construction of verb classifications is time-consuming and costly, the possibility of automating the process has attracted a lot of interest in the research community. The problem has been framed either as a supervised classification task,

where new verbs are assigned to pre-defined classes, or as an unsupervised clustering task, where groupings of verbs are discovered based on shared properties and behaviour. The latter approaches have often employed subcategorisation information (i.e., *subcategorisation frames*, which represent the syntactic realisation of the arguments a verb can take) as clustering features (Brew and Schulte im Walde, 2002; Im Walde, 2000; Kamp, 2019; Lapata, 1999; Sedoc et al., 2017; Sun et al., 2008b; Vlachos et al., 2009), sometimes supplemented with additional information about selectional preferences (Im Walde, 2006; Sun and Korhonen, 2009). These techniques were applied to acquire VerbNet-style classes in French (Sun et al., 2010) and Brazilian Portuguese (Scarton et al., 2014). While these approaches have focused on clustering verb *types*, a number of works explored the potential of clustering verb *instances*, taking into account polysemy. For example, Materna (2012), Kawahara et al. (2014) and Kawahara and Palmer (2014) performed clustering based on semantic frames, where resultant groupings of verbs share predicate-argument structures.

Although the unsupervised approaches offer the advantage of applicability in low-resource scenarios, the output partitions are inevitably noisy. Therefore, when gold standard training data are available, supervised techniques have been shown to yield higher quality classifications (e.g., Falk et al., 2012; Joanis et al., 2008; Li and Brew, 2008; Merlo and Stevenson, 2001; Ó Séaghdha and Copestake, 2008; Sun, 2013; Sun et al., 2008a). Further, a number of works have demonstrated that even a small amount of supervision can significantly improve the accuracy of the resultant groupings. For instance, Vlachos et al. (2009) boosted their unsupervised clustering algorithm with a small set of pairwise constraints representing if a given pair of verbs belong together, while Peterson et al. (2016) achieved improved clustering performance by additionally predicting a VerbNet class for each verb sense, relying on annotated VerbNet data. More recently, Peterson et al. (2020) reported further gains obtained by partially supervising the joint sense induction and clustering model of Peterson and Palmer (2018) with some directly observed sentences annotated with VerbNet class labels.

Alternatively, in scenarios where some lexical resources were already available, verb classes have been obtained in a semi-automatic fashion, thus reducing the time required and improving on fully automatic methods in terms of accuracy. Scarton and Alusio (2012) used alignments between English VerbNet and Wordnet, and the Brazilian Portuguese WordNet, to infer VerbNet-style classes in the latter. Mikelić Preradović and Boras (2013) used a manually constructed Czech verb classification (based on 49 English Levin classes), VerbaLex, and a Croatian valency lexicon (Preradovic et al., 2009) to assign Czech valency frames to the verbs in the closely related Croatian.

Further, Uresova et al. (2018) proposed a Czech-English bilingual verb synonym lexicon, where groupings of semantically equivalent verb senses are based on valency and predicate-argument structure information, drawing on several language-specific lexicons (e.g., English VerbNet and FrameNet, English, Czech, and Czech-English valency lexicons (Cinková et al., 2014; Urešová et al., 2016, 2014)) as well as a richly annotated bilingual corpus (Hajič et al., 2012). Moreover, Pradet et al. (2014) built a French VerbNet based on translation of top-level VerbNet class members from English, followed by a de-noising step using a semantic (Dubois et al., 1997) and syntactic verb lexicon (Boons et al., 1976; Gross, 1975) already available in the target language.

Whether supervised or unsupervised, automatic or semi-automatic, these approaches often require substantial expertise and effort for feature engineering, as well as access to NLP tools and resources, such as accurate parses, subcategorisation frames, or pre-existing lexical databases, unavailable in most languages and focused domains. More recently, neural classification approaches based on automatically learned features have been shown to achieve results superior to those produced with previous methods reliant on sophisticated linguistic features and language resources across diverse languages (Vulić et al., 2017b). Further, recent work has shown that neural methods can be particularly useful for extending the coverage of existing resources with high accuracy (Chiu et al., 2019). Still, utility of such methods relies on the availability of human-generated data, for evaluation and model supervision. Approaches to their time-efficient acquisition, like the one presented in this thesis (Chapter 4), can therefore play an important role in facilitating development and application of automatic classification techniques, thus extending their benefits to under-represented languages and domains.

Deriving Verb Clusters from Non-experts

While much verb classification work in NLP relied on acquisition of patterns of verb behaviour from text corpora, some linguistically motivated research explored acquiring such information from language users. For example, Öztürk et al. (2011) studied human clustering decisions in a sorting task with 41 Norwegian motion verbs. Based on the assumption that humans perform some form of analysis of relative similarities between words in order to sort them, they examined what features underlie those judgments and their relative impact on the sorting performance. The collected sorting data and user feedback on the salient features taken into consideration were converted into a verb co-occurrence matrix and a feature-verb matrix. These were then used as input in a stepwise method to obtain a verb similarity matrix based on feature

weights (computed using linear regression), and subsequently fed into an agglomerative hierarchical clustering algorithm yielding clusters across different granularity levels.

Further, in a study extending the experimental findings about the systematic relationship between components of verb meaning and syntactic behaviour in language acquisition research, White et al. (2014) investigated what kind of semantic distinctions are retrievable from syntax, focusing specifically on propositional attitude verbs (e.g., *think, want*). For a set of 30 attitude verbs, they collected human judgments of syntactic compatibility (i.e., whether a given verb can appear in a particular syntactic construction) and semantic similarity (elicited on triads of verbs, where the task involved choosing the verb with a meaning least like the others). They subsequently applied hierarchical clustering to both types of data to retrieve semantically and syntactically driven groupings of verbs. What is especially relevant for this thesis, they found that non-experts are sensitive to fine-grained semantic features of verbs. Crucially, they observed significant correlation between the classes obtained based on semantic similarity judgments and those derived from syntactic compatibility judgments. This finding provides additional support for the potential of acquiring verb classes based on semantic judgments alone, investigated in this thesis (Chapters 3 and 4).

Beyond Levin Classes

Levin's work aimed at exploring the limits of the hypothesis that verbs' syntactic behaviour is semantically determined and the potential to organise the English verb inventory using it as the guiding principle – with the caveat that the resultant taxonomy is just one possible solution to the classification problem. Since a number of meaning components are characteristic of several different classes (as identified by Levin), class boundaries can be drawn in more than one way, each of them equally valid. For instance, the partition structure proposed in the above discussed FrameNet organises English verbs in terms of their potential to evoke certain semantic frames, which capture the interaction between a verb's meaning and the syntactic realisation of its arguments. Given the different frameworks guiding the creation of both lexicons (diathesis alternations vs. frame semantics), the alignment between FrameNet frames and VerbNet classes is not perfect. The mappings between these two resources created as part of the SemLink project (Palmer, 2009) are many-to-many, with a number of VerbNet members mapping to more than a single FrameNet frame. Accordingly, the experimental results in Chapter 6 show that these two types of knowledge representations can have a different impact on downstream task performance, when used to enhance the underlying model's grasp of verb meaning and behaviour. Still, both frames and

classes capture powerful generalisations about verbs' properties, which makes them a useful predictive tool.

In this thesis, the emphasis is put on the ultimate usefulness of verb classes as organisational units of verb knowledge for helping NLP systems abstract away from individual words and infer information about their meaning and shared patterns of behaviour (Chapter 6). To extend those benefits across different languages, I explore the potential for efficient acquisition of fine-grained verb knowledge based on semantic judgments alone (Chapters 4 and 5). Crucially, while Levin used the insights from studies probing native speakers' lexical knowledge to inform the rationale behind her lexicographic effort, the aim of the work in this thesis is to tap into that knowledge to generate verb classes in a bottom-up fashion, with minimal expert supervision, and use them to organise verb meaning (Chapters 3 – 5).

2.2 Evaluation and Data Collection Paradigms

The quest to learn and represent the meaning of linguistic expressions based on their patterns of occurrences in text corpora continued in the field of distributional semantics has resulted in a number of important breakthroughs in natural language processing. Fuelled by the ever-growing amount of easily accessible digital textual data, representation learning has become a central endeavour in the pursuit of more powerful NLP systems across a wide range of applications. Since accurately capturing the meaning of text is a prerequisite in a multitude of tasks, evaluation of representation quality is a key step before model deployment. The availability of high-quality evaluation resources therefore plays a crucial role in spurring on advances in this area. Estimation of semantic likeness between natural language expressions, which humans perform intuitively, is a complex task which serves as a useful yardstick for measuring the model's linguistic ability. However, different views on what meaning relationships semantic models should capture, and how that ability should be tested, have informed work on semantic similarity-based evaluation, producing evaluation benchmarks with different characteristics and underlying motivations. In what follows, I discuss current approaches to intrinsic evaluation of representation models and dataset construction, focusing on lexical meaning. Prior to that, I briefly introduce the rationale behind some of the currently most widely used representation learning architectures in the NLP community, which are also employed in the experimental work presented in this thesis. I conclude this section with a discussion of the characteristics of the spatial

arrangement method employed in this work and highlight its benefits for multilingual semantic data collection.

2.2.1 Representation Learning

The distributional semantic framework rests on a view of lexical meaning as a function of word usage (Wittgenstein, 1953), as reflected in word distributions in a corpus of text. The hypothesis that words with similar distributions are also similar in meaning (Harris, 1954) is central to the research on representation learning, aimed at automatic discovery and extraction of features for classification and prediction directly from data (Bengio et al., 2003). Particularly influential models of distributional semantics are those based on vector representations, where each dimension corresponds to a feature. For each word w in the vocabulary V, there is an associated vector of real numbers $\mathbf{w} \in \mathbb{R}^d$, which maps its meaning to a point in a d-dimensional space. The underlying intuition in *Vector Space Models* is that the closer the meaning of two words is, the closer they will be located in that space. Vector representations can be constructed based on counts of occurrences of a given word in different contexts (e.g., Baroni and Lenci, 2011; Lin, 1998; Padó and Lapata, 2007), or alternatively learned as part of a machine learning model based on some objective, e.g., predicting the next word in a sentence given the preceding ones (Bengio et al., 2003; Collobert and Weston, 2008). While the frequency-based approach requires large vector dimensions to represent co-occurrence patterns of all words in a corpus, prediction-based techniques which encode distributional information in dense lower dimensional vectors by means of a neural network have become widely popular (Mikolov et al., 2013b; Pennington et al., 2014), and are also the focus of the model evaluations in the upcoming chapters.

Static Representations

Among the most impactful recent word embedding models are the two representation learning algorithms proposed by Mikolov et al. (2013a,b), Skip-gram and Continuous Bag-of-Words (CBOW). While both focus on predicting words in a given window of text, they approach the task from two opposing perspectives. The former predicts the context words given a word in the centre of the window, while the latter predicts the centre word given the surrounding words. For example, given a sequence of text *Leopold Bloom ate with relish*, where the verb *ate* is the centre word and we assume a context window of size 2, the Skip-gram maximises the conditional probability of the context words given the centre word, considering words within a distance of no more

than 2 words away from the centre and predicting each independently from the rest, that is:

$$P(\text{“Leopold”}|\text{“ate”}) \cdot P(\text{“Bloom”}|\text{“ate”}) \cdot P(\text{“with”}|\text{“ate”}) \cdot P(\text{“relish”}|\text{“ate”})$$

Assuming that w_i is an ith word in the vocabulary and that $\mathbf{v}_i \in \mathbb{R}^d$ is its d-dimensional vector when it is the target word (i.e., the input vector of w_i) and $\mathbf{u}_i \in \mathbb{R}^d$ is its d-dimensional vector when it is the context word (i.e., the ouput vector of w_i), let w_t be the centre word and w_o be the outside context word, and $V = \{0, 1, ..., |V| - 1\}$ be the set of vocabulary indices. The conditional probability of predicting the context word given the centre word is obtained using a softmax operation (where **exp** denotes the exponential function):

$$P(w_o|w_t) = \frac{\exp(\mathbf{u}_o^\top \mathbf{v}_t)}{\sum_{i \in V} \exp(\mathbf{u}_i^\top \mathbf{v}_t)} \tag{2.1}$$

In CBOW, on the other hand, the situation is reversed. Given a context, we are concerned with the conditional probability of the centre word given its context:

$$P(\text{“ate”}|\text{“Leopold”},\text{“Bloom”},\text{“with”},\text{“relish”})$$

CBOW uses the same method to obtain the conditional probability as Skip-gram, however, it takes the average of context word vectors $\mathbf{v}_i$ (now the input) within a given window of size m, calculated as:

$$\bar{\mathbf{v}} = \frac{\mathbf{v}_{o-m} + \mathbf{v}_{o-m+1} + ... + \mathbf{v}_{o+m}}{2m} \tag{2.2}$$

Then, assuming that $\mathbf{u}_i$ is now the centre word vector and $\mathbf{v}_i$ is the context word vector, given the centre target word w_t and context words $W_o = \{w_{o-m}, w_{o-m+1}, ..., w_{o+m}\}$, the conditional probability of the centre word given the context is calculated as:

$$P(w_t|W_o) = \frac{\exp(\mathbf{u}_t^\top \bar{\mathbf{v}})}{\sum_{i \in V} \exp(\mathbf{u}_i^\top \bar{\mathbf{v}})} \tag{2.3}$$

The more computationally expensive Skip-gram, which trains each word-context pair individually, predicting for every word one word from its context, has been found to be more robust than CBOW when dealing with rare words. However, the latter, which uses averaged contexts for predictions, offers the advantage of greater efficiency for processing large corpora. Word vectors pretrained using these algorithms, released as part of the `word2vec` resource, have been shown to provide powerful input features in

a variety of NLP tasks (Al-Rfou et al., 2015, 2013; Chen and Manning, 2014; Tsvetkov et al., 2015). They also inspired a number of extensions and modifications, including the generalisation of the Skip-gram with negative sampling to arbitrary, non-linear contexts by Levy and Goldberg (2014). Notably, Levy and Goldberg (2014) showed that deriving syntactic contexts from dependency parse trees produces word embeddings which capture different information from the original model. Consequently, the semantic relations encoded in the embedding space are also of a different nature: Mikolov et al. (2013b)'s bag-of-word contexts yield broader topical relatedness (e.g., *snow* and *sleigh*), while Levy and Goldberg (2014)'s syntactic contexts capture functional similarities between words which share similar behaviour (Turney, 2012) (e.g., *sleigh* and *snowboard*).

A feature of these approaches, however, is that word tokens are treated as opaque entities, with no attention being paid to their internal structure. For example, they are oblivious to the relationship between inflectional variants of the same verb, *gives* and *given*. They assign a distinct vector to each word seen during training, and as a consequence, fail to yield vectors for words which are not part of their vocabulary, even if they are decomposable into already known words (e.g., compounds like *applesauce*). This is a serious limitation, particularly in the case of languages with very rich and productive morphology, and hence a large number of infrequent word types.

This shortcoming has been addressed by an extension of the `word2vec` model, `fastText` (Bojanowski et al., 2017; Mikolov et al., 2018), which enriches the embeddings by representing words as bags of character n-grams, rather than only learning their vectors directly. For instance, for $n = 3$, the following n-grams would be generated for the word *relish*: <re, rel, eli, lis, ish, sh> (< and > denoting the beginning and end of a word), and its vector would be computed as the sum of the word-level and averaged subword-level representations (i.e., of the constituent n-grams). This makes the method much more robust to rare words, and more suited for modeling morphologically rich languages.

Contextualised Representations

The discussed models yield powerful semantic representations, capable of improving model generalisability in tasks where limited training data is available (Collobert et al., 2011). However, their limitations lie in their static view of meaning, disregarding its variability according to a changing context. The challenge of adequately representing different usages of the same word type in text was taken up by the family of *contextualised word embeddings* (McCann et al., 2017; Peters et al., 2017, 2018). The model

of Peters et al. (2018) ELMo (*Embeddings from Language Models*) yields a different vector for each instance of a given word, which is a function of the whole input sentence (e.g., compare *Leopold Bloom ate with relish* vs. *Use salsa as a relish with grilled fish*, where `word2vec` would produce the same vector each time). ELMo word vectors are derived from a bidirectional Long short-term memory (LSTM) neural network, which captures contextual information in two directions, both following and preceding the input word. It is trained on a large unlabelled text corpus with a coupled language model (LM) objective (i.e., to predict the next word in a word sequence). All the hidden states of the biLM are linearly combined for each input word to yield rich context-dependent word representations. The authors show that, indeed, a different kind of signal is captured in the higher-level states, which encode the contextual aspects of word meaning, and the lower-level states, which capture syntactic properties.

The addition of ELMo embeddings has been shown to significantly improve performance on a variety of NLP tasks, including question answering, textual entailment and sentiment analysis (Peters et al., 2018). In the 'plug-in' scenario, pretrained ELMo embeddings are incorporated into task-specific models as additional features for the purposes of target downstream applications – the so-called *feature-based* approach. Another useful way of using pretrained language representations in downstream tasks is the *fine-tuning* approach. This involves adding a limited number of task-specific parameters, and then training the model on the target task by updating all the parameters pretrained using the language modeling objective, instead of randomly initialising and training them from scratch (Howard and Ruder, 2018; Radford et al., 2018). This second type of transfer approach enables efficiently utilising the information acquired during pretraining for the benefit of any downstream NLP application. However, in approaches such as that proposed by Radford et al. (2018), the general language representations are learned via a unidirectional forward language model, without capturing context in both directions, which can be detrimental in tasks such as question answering where this information is key.

Devlin et al. (2019) proposed to address this limitation by introducing Bidirectional Encoder Representations from Transformers (BERT). Based on the Transformer architecture (Vaswani et al., 2017), the model pretrains deep bidirectional representations on unlabelled text using a *masked language model* pretraining objective (based on the Cloze task (Taylor, 1953)), jointly conditioning on both the left and right context in all Transformer layers. In masked language modelling, a proportion of input tokens is masked at random (e.g., `Leopold Bloom ate with [MASK]`), and the objective is to predict the vocabulary ID of the masked word based on its context, which allows the

learned representation to fuse the left and right context. Note that this is different from the above discussed ELMo, which employs a shallow concatenation of independently trained left-to-right and right-to-left language models. The second pretraining task is next sentence prediction, which additionally allows the model to capture the relationship between two sentences (not directly captured by language modeling), an important ability for sentence-level tasks involving inference or question answering. The task is framed as binary classification consisting in predicting whether the second sentence is the true next sentence of the first sentence, given a pair of sentences in the following format (tokenised into wordpieces (Wu et al., 2016)):

```
[CLS] two letters and a card lay on the [MASK] [SEP]
he stopped and [MASK] them [SEP]
```

A special classification token is added to the start of every input sequence (`[CLS]`, used as an aggregate of the whole sequence representation) and a separator token (`[SEP]`) is added between the two sentences, and at the end of the sequence. BERT obtains a fixed-length vector used as input to the Transformer by summing three embedding layers: the token layer, the segment layer (capturing which sentence a given token belongs to), and position layer (capturing the token's position in the sequence). The resultant model can be subsequently fine-tuned on labelled data from a downstream task of choice with only one additional output layer, with minimal task-specific adjustments of the model architecture, achieving state-of-the-art performance in numerous NLP tasks (Devlin et al., 2019). At the same time, it is also possible to extract fixed pretrained representations from BERT as input features in a feature-based approach, which is useful for tasks which do not lend themselves to a Transformer encoder architecture. In the following chapters, I employ the model in both scenarios. First, I extract token-level representations for the purposes of intrinsic evaluation, where I compare its performance against its predecessor models discussed in this section (Chapter 4 and 5). Next, I fine-tune the pretrained BERT on downstream task data (Chapter 6).

2.2.2 Word Similarity Estimation

The characteristic of the above discussed approaches is that they yield numerical word representations that lend themselves to vector arithmetic, which provides a useful tool for evaluating their quality. Given their underlying assumption that words similar in meaning tend to share similar contextual distributions (Miller and Charles, 1991), we can mathematically evaluate whether the learned representations reflect this property by simply computing the distance between the words' vectors in a multi-dimensional

semantic space. Importantly, the dimensions which constitute this space are intrinsically linked to and a product of the specific contextual information used to build a given distributional model (Section 2.2.1).

To compare word vectors, a simple and widely used measure is cosine similarity, i.e., the cosine of the angle between two vectors, where the similarity is inversely proportional to the angle (i.e., the smaller the angle, the greater the cosine similarity). Given two words, w_1 and w_2, and their vector representations $\mathbf{u}$ and $\mathbf{v}$, a vector dimensionality d, with u_i being a component of vector $\mathbf{u}$ in dimension i, cosine similarity of words w_1 and w_2 is calculated as:

$$sim_{cos}(w_1, w_2) = \cos(\theta) = \frac{\mathbf{u} \cdot \mathbf{v}}{||\mathbf{u}||||\mathbf{v}||} = \frac{\sum_{i=1}^{d} u_i v_i}{\sqrt{\sum_{i=1}^{d} u_i^2}\sqrt{\sum_{i=1}^{d} v_i^2}} \tag{2.4}$$

This geometric view of semantic closeness, where we compare the relative positions of words in a multi-dimensional space representing observable uses of language, aligns with the influential spatial model of similarity proposed in cognitive science research (Shepard, 1962).[5] In spatial models, similarity between concepts – or rather, their mental *representations* – is an exponentially decreasing function of the distance between them in a metric space (Ashby and Perrin, 1988; Shepard, 1987). Meaning is thus continuous, continuously varying along the space's multiple dimensions (Osgood, 1952).

Indeed, distributed representations of meaning have been found to be well suited for modelling how humans reason about meaning and similarity, allowing for solving word analogies (e.g., *mason:stone – carpenter:wood* (Turney et al., 2003)), as well as finding related words (e.g., *dancing → singing, rapping, breakdancing* (Levy and Goldberg, 2014)). How well these representations align with human appreciation of similarity between words can be assessed by comparing the similarity scores computed between vectors (Eq. 2.4) with numerically represented human similarity judgments for the same word pairs. Given the subjective nature of such judgments (i.e., there is no *true* value of a semantic similarity meausre (Harispe et al., 2015)), human consensus is established by averaging judgments across individuals, and then treated as a benchmark for models. The comparison of model output to averaged human judgments can be made by computing the correlation between the two, which avoids the problem of comparing absolute scores associated with different scales and only considers how the two signals behave. Frequently, Spearman's rank correlation coefficient is used for

[5]It is worth noting that despite its popularity, the geometric model has also been a subject of criticism due to the inconsistencies found between humans' perception of similarity and the constraints imposed by the axiomatic properties of distance (Tversky, 1977; Tversky and Gati, 1978, 1982).

this purpose, which takes into account the relative ordering (i.e., ranks) of pairwise similarities based on their values, computed as:

$$\rho(a, b) = 1 - \frac{6 \sum_{i=1}^{n} (a_i - b_i)^2}{n(n^2 - 1)} \tag{2.5}$$

where a and b represent the two arrays of similarity scores (from humans and those yielded by the model) for n word pairs. The result is bounded in the $[-1, 1]$ interval, where correlation of 1 signifies identity of ranks. The closer the correlation is to 1, the closer the model representations mirror human judgments, and hence, the higher their quality.

Datasets comprising such pairwise judgments are essential for evaluating representation quality. However, evaluation benchmarks differ in the underlying assumptions about what kind of similarity these judgments should reflect, and which meaning relationship the models should capture to meet the needs of their ultimate application. What is more, it has been recognised that different parts of speech impose different demands on the representation models due to their unique linguistic properties (Bansal et al., 2014; Melamud et al., 2016; Schwartz et al., 2015; Vulić et al., 2017c). As discussed in Section 2.1, cognitive models of the lexicon generally assume that representations of verbs contain interdependent properties which noun representations lack, such as subcategorisation information, selectional preferences, and event structure (Fisher et al., 1991; Gleitman, 1990; Resnik and Diab, 2000). In representation learning literature, it has been shown that modelling verbs requires paying attention to different contextual information than in the case of nouns. For example, Schwartz et al. (2015) demonstrated that different contexts prove informative for learning representations of verbs and adjectives (coordination structures x *and* y), and different ones are optimal for nouns (the traditional bag of words). Hence, focused evaluation and fine-tuning of representation learning algorithms on individual lexical categories, such as verbs, can aid the models in better capturing their characteristics and increase their utility for category-specific applications (e.g., automatic verb classification). However, most evaluation resources currently available focus on noun similarity, paying less attention to verb semantics or excluding verbs completely. What is more, the issue is further exacerbated in languages other than English, where large-coverage evaluation resources are still rare. In what follows, I briefly introduce the notions of *relatedness* and *similarity*, and subsequently discuss the approaches to collecting human judgments of these two meaning relationships on the example of datasets commonly used in intrinsic evaluation of NLP models.

Similarity vs. Relatedness

Comparison is an essential part of human cognition. The ability to judge the similarity between entities is a key component of learning, involved in a multitude of processes such as decision making, problem solving, memory retrieval, categorisation, and pattern recognition (Gentner and Markman, 1997; Goldstone and Son, 2012; Hahn et al., 2003; Holyoak and Koh, 1987; Markman and Gentner, 1993; Novick, 1988; Ross, 1987, 1989; Vosniadou and Ortony, 1989). As a key feature of an intelligent system and a manifestation of the ability to reason, similarity estimation is considered as a good proxy task to assess the capacity of general-purpose representation learning models to learn and represent meaning encoded in text (Baroni et al., 2010; Collobert and Weston, 2008). However, when the notion of similarity is left under-specified, it can refer to different types of relation: a broader *relatedness* or a narrower *semantic similarity*, and a model's capacity to estimate the former does not necessarily entail the same ability with regard to the latter. Different applications may also favour sensibility to one or the other type of relation. For example, accurately estimating relatedness is important in information retrieval (Akmal et al., 2014; Chen et al., 2017a; Gurevych et al., 2007; Lopez-Gazpio et al., 2017; Srihari et al., 2000; Uddin et al., 2013) and text summarisation (Kozima, 1994), while recognising similarity as distinct from association is vital in generation of thesauri, language correction tools, machine translation, and question-answering systems (Biemann, 2005; Cimiano et al., 2005; Freitas et al., 2011; He et al., 2008; Iordanskaja et al., 1991; Li et al., 2006; Marton et al., 2009; Wang et al., 2012a).

The term *semantic relatedness* has been used to describe words linked by different kinds of semantic relations (Budanitsky and Hirst, 2001, 2006; Gentner, 1983; Turney and Pantel, 2010), including synonyms (*puzzle – bemuse*), meronyms and holonyms (*peel – fruit*), as well as antonyms (*light – dark*). Similarity defined as association (*associative relatedness*), that is, the mental activation of a term when another is presented (Chiarello et al., 1990; Lemaire and Denhiere, 2006) (e.g., *butter – knife, hammer – nail*), has been estimated in terms of how frequently the two words co-occur in the same contexts in language (and the physical world) (Bruni et al., 2012; McRae et al., 2012; Turney, 2001; Turney and Pantel, 2010). When asking human raters about relatedness of words, researchers are typically interested in quantifying the strength of the perceived connection between the concepts they represent, without making any assumptions about the factors underlying this link (Taieb et al., 2020).

Related words may be strongly associated, but they are not substitutable. For example, *bikini* is associated with *beach*, but one cannot substitute them for one another

in the following sentences without changing the underlying semantics: *I tried on a new bikini* and *The beach was crowded.* The word *trunks*, on the other hand, could be considered both related and semantically similar to *bikini*, as both refer to garments often worn on the beach. Their similarity can thus be quantified in terms of the impact of substituting one for the other on the meaning of the sentence (Budanitsky and Hirst, 2001; Resnik, 1995; Weeds, 2003). This second relation, semantic similarity, has been defined in terms of shared superordinate category (Lupker, 1984; Resnik, 1995) (also referred to as *taxonomical similarity* (Turney and Pantel, 2010)) or common semantic features (Frenck-Mestre and Bueno, 1999; Turney, 2006; Tversky, 1977). Viewed this way, similarity is quantified in terms of degree of overlap in semantic properties, such as shared function or physical features, with synonyms occupying the top region of the similarity scale (e.g., *fiddle – violin* (Cruse, 1986)). Similarity is often considered as a kind of relatedness, where the compared concepts belong to the same ontological class (Jurafsky, 2000).

In the context of distributed word representations, the distinction between semantic similarity and relatedness can also be made in terms of *syntagmatic* and *paradigmatic* relationships (Saussure, 1916). The former refers to the relations between linguistic entities co-occurring in a sequence of text, e.g., in the sentence *I parked the car on the street*, *car* and *street* are syntagmatically related. Whereas paradigmatic relationships hold between entities which belong to the same class and can assume the same function in the sentence. For instance, in the above sentence, the slot occupied by *car* (*I parked the ___ on the street*) can be taken up by other members of the category *vehicle*, e.g., *motorbike, van, Mercedes.* Paradigmatic relationships therefore involve substitutability, which can be understood as non-violation of grammatical coherence (e.g., members of the lexical category 'noun') or in narrower semantic terms as interchangeability without affecting the meaning conveyed by the sentence (e.g., synonyms). As discussed in 2.2.1, in representation learning bag-of-words contexts are useful for capturing syntagmatic relatedness, while syntactic dependency-based contexts help the model learn paradigmatic relationships between *functionally* similar words.

In this thesis, I reserve the term (semantic) similarity for this latter definition of closeness of meaning, as distinct from the more general relatedness, which also includes association, following previous work (Agirre et al., 2009; Gerz et al., 2016; Hill et al., 2015; Resnik, 1995; Resnik and Diab, 2000). In Chapter 4, I explore how this distinction is captured by native-speaker judgments in the two tasks constituting the introduced annotation design, rough semantic clustering and spatial arrangements of

Dataset	Task	Scale	N pairs	N verb pairs	References
RG65	sim	0.0-4.0	65	0	Rubenstein and Goodenough (1965)
Resnik & Diab	sim	0-5	27	27	Resnik and Diab (2000)
WordSim-353	rel	0-10	351*	0	Finkelstein et al. (2002)
YP-130	rel	0-4	130	130	Yang and Powers (2006)
MEN	rel	N/A	3000	390**	Bruni et al. (2014)
Verb-143	sim	0-4	143	143	Baker et al. (2014)
SimLex-999	sim	0-6	999	222	Hill et al. (2015)
SimVerb-3500	sim	0-6	3500	3500	Gerz et al. (2016)

Table 2.1 Examples of English language datasets based on human judgments of the similarity (**sim**) or relatedness (**rel**) of word pairs (rating-based or comparative), including N verb pairs. *Scale* refers to the rating scale used for data collection (i.e., the raw judgments), rather than the score interval after post-processing or normalisation. *The final pair number after excluding 1 duplicate and 1 identity pair. **The number of pairs including at least one verb.

words, through qualitative and quantitative analysis with reference to existing lexical resources.

Datasets and Data Collection Paradigms

Pairwise Word Similarity. One of the most widespread methods for collecting human similarity judgments is by eliciting ratings of similarity for lists of word pairs on a discrete scale. The method imposes minimal technological demands, is fast and easy to implement, and potentially requires very little annotator training. In this approach, raters are presented with sets of word pairs and are asked to judge the degree of similarity/relatedness (see discussion above) between the words in each. This method is particularly amenable to crowdsourcing, where the task can be split into small batches, requiring very short time to complete, and assigned to multiple *turkers* who participate on an online platform (e.g., Mechanical Turk[6], Appen[7]). The design choices involve the construction of pairs (e.g., focusing on pairing frequently co-occurring words (Radinsky et al., 2011) or capturing explicit semantic relations (Luong et al., 2013)), as well as the type of words (e.g., common or rare, concrete or abstract) and part of speech, up to the choice of rating scale. Table 2.1 presents examples of popular English language datasets, highlighting the type of semantic relation being measured (similarity or relatedness), the rating scale used for judgment collection, and their coverage of verb similarity.

Datasets constructed using this approach have been amongst the most widely used in the research community for evaluation of semantic models. In the pioneering

[6]https://www.mturk.com
[7]https://www.appen.com

work of Rubenstein and Goodenough (1965), the method was first adopted to test the distributional hypothesis (Harris, 1954) using human judgments of word semantic similarity, collected on 65 pairs of nouns. The subjects in the experiment were presented with a deck of 65 paper slips, each containing one word pair. The task required them to first order the pairs from the most similar to the least similar, and subsequently assign a numerical similarity score to each on the scale between 0.0 and 4.0 (e.g., *coast – shore* 3.60, *bird – woodland*, 1.24)[8]. The experiment's setup took into account the importance of eliciting similarity judgments in the context of other pairs, rather than pairs in isolation. Indeed, the assignment of ratings followed the comparative appreciation of relative differences in similarity of the pairs in the set. However, although the study's scale was relatively small, the task of considering 65 pairs simultaneously entails a high cognitive load. Approaches which elicit relative judgments in a more efficient way by simultaneously capturing multiple pairwise similarities, such as the spatial arrangement method discussed later in this section (and in Chapter 4), therefore offer a considerable advantage in terms of scalability and the cognitive demands on the annotators.

Given the unique focus on noun similarity in the dataset of Rubenstein and Goodenough (1965), Resnik and Diab (2000) directed their attention to the meaning of verbs. Taking into account the additional factors potentially impacting verb similarity judgments, such as subcategorisation behaviour, thematic properties (i.e., argument structure), and lexical aspect, the authors constructed a set of 48 verb pairs[9] in such a way as to limit the dimensions along which any two words could differ. Specifically, in each pair, the two verbs had to both require a Theme argument, belong to the same aspectual class, and have matching subcategorisation frames, so as to disentangle semantic considerations from effects of syntactic likeness. They also experimented with presenting verb pairs with and without context, and found the two sets of judgments to be highly correlated (Pearson's $r = .89$). The participants were asked to rate each pair on the scale from 0 to 5, focusing on similarity rather than relatedness. Additionally, a "don't know" response was allowed in case the meaning of a low-frequency verb was unclear. Importantly, like in the work of Rubenstein and Goodenough (1965), each

[8]Although the task guidelines explicitly stated that the word pairs should be judged based on the "amount of *similarity of meaning*", word association clearly impacted the judgments. For instance, the averaged similarity score for *bird – woodland* (1.24) is higher than for the pair *monk – slave* (0.57). While the first pair is strongly associated (i.e., birds inhabit woodlands), there is hardly anything similar between the two words – one is an animal and the other refers to land covered in trees. Whereas the words in the second pair, although not associated, both describe human beings.

[9]Although judgments were collected on 48 pairs, the authors later excluded the total of 21 pairs upon discovery of violations of the semantic-syntactic qualification criteria set out at the beginning and due to some words being unknown to some participants. The final released set thus includes only 27 non-excluded pairs.

subject was presented with all 48 pairs and asked to compare the similarities across pairs. Notably, they found inter-rater agreement to be significantly lower than in an analogous study on nouns (.79 on verbs in context and .76 without context, compared to .90 achieved on nouns), which they interpret as indicative of the greater difficulty of verb similarity estimation. Interestingly, they also found that metaphorical extensions of verb meanings play a role, where raters are inclined to assign non-zero scores to dissimilar pairs such as *unfold – divorce* and *initiate – enter*. In Chapter 3, I discuss similar effects in the semantic clustering task, where participants sometimes followed a "storyline" approach (e.g., clustering verbs such as *approach, conquer, marry* and *move* together).

RG65 and the dataset of Resnik and Diab (2000) have been followed by several larger datasets focusing either on noun or verb semantics. Finkelstein et al. (2002) collected relatedness judgments on 353 noun pairs, without however making an explicit distinction between semantic similarity and association. The guidelines instructed raters to assign the score of 0 to unrelated and 10 to very closely related words, and explicitly asked to consider antonyms similar ("belonging to the same domain or representing features of the same concept"). To allow for distinguishing between similarity and relatedness, the dataset was later divided into two separate sets focusing on each of the semantic relations, WordSim-203 and WordRel-252 (Agirre et al., 2009). Yang and Powers (2006) later proposed a relatedness-based dataset focused entirely on verbs, drawing verb pairs from the TOEFL (Test of English as a Foreign Language) (Landauer and Dumais, 1997) and ESL (English as a Second Language) (Tatsuki, 1998) synonym sets. They subsequently asked six participants (4 native speakers and 2 near-native) to rate 144 pairs according to their relatedness on a discrete 0-4 scale (a number of pairs were eliminated in post-processing resulting in a 130-pair dataset). Baker et al. (2014) collected human similarity judgments on 143 pairs of verbs following WordSim-353 guidelines. The pairs included different forms of a single verb (e.g., *affect*, *affecting*, *affected*) and were examined by an independent group of human judges prior to the experiment so that only those with some degree of similarity were used.

Given the ambiguous interpretation of similarity adopted in some of these works, often confounded with association, Hill et al. (2015) proposed a dataset where the distinction between the two is made explicit, SimLex-999. The annotation guidelines included examples of pairs illustrating both relations and instructed the raters to beware of related but dissimilar pairings, which should be assigned low scores (e.g., *movie – theatre*). In sampling the pairs, Hill et al. (2015) ensured a wide coverage of different levels of concreteness and inclusion of different parts of speech (adjective, noun,

and verb), as well as associated and unassociated pairs, which allows for evaluating the capacity of models to distinguish between associated and similar, and associated but dissimilar words. Specifically, they sampled 900 associated pairs from over 70 thousand pairs included in the University of South Florida Free Association Database (USF) (Nelson et al., 2004), and subsequently complemented the sample with 99 unassociated pairs by randomly combining words from the associated set. Importantly, in light of the different cognitive processes at play when judging similarity of different parts of speech (Gentner, 2006; Markman and Wisniewski, 1997), they only permitted same-POS pairings, resulting in 666 nouns pairs, 222 verb pairs, and 111 adjective pairs. Five hundred native-speaker raters were recruited on a crowdsourcing platform (Amazon Mechanical Turk) to rate the similarity of pairs presented in batches of 6-7 on a 0-6 integer scale by moving a slider (non-integral scores were not permitted).

Adopting the same human judgment collection method as SimLex-999, Gerz et al. (2016) constructed a large-scale dataset focused exclusively on verb similarity, SimVerb-3500. To ensure coverage of verbs with a wide range of semantic-syntactic properties, they used VerbNet class membership to inform their sampling procedure. Specifically, they started by sampling all possible verb pairs from USF, subsequently removing pairs with multi-word verbs (e.g., *take on*), auxiliaries, or non-infinitive forms, yielding 3,072 pairs. They subsequently sampled verbs directly from VerbNet to ensure coverage of all VerbNet classes by at least 3 verbs,[10] which formed 81 additional pairs. Finally, they followed the approach of Hill et al. (2015) to supplement the set with 347 unassociated pairs by randomly combining associated verbs (and excluding those which coincided with existing USF pairs). Over 800 raters recruited on a crowdsourcing platform[11] rated the similarity of 79 verb pairs each, presented in batches of 7-8, on a discrete 0-6 scale using a slider. The study resulted in averaged scores for 3500 verb pairs, linearly scaled to the 0-10 interval. Given its scale and exclusive focus on verbs, SimVerb-3500 is used as the main reference resource for evaluating the methodology and dataset presented in this thesis (*SpA-Verb*). In Chapter 4, I carry out an extensive comparative analysis of the two resources, discussing their relative strengths and weaknesses, and highlight the unique properties of the SpA-Verb dataset constructed through spatial arrangements of words, compared to the data gathered by means of the pairwise rating method.

[10]Exceptions include VerbNet class 56 (which includes denominal verbs such as *honeymoon, winter, weekend*) and small classes 91 and 93.

[11]https://www.prolific.co

Modifications and Alternative Approaches. While being conceptually simple and practical, the pairwise rating approach suffers from certain limitations. In particular, assigning discrete scores to capture subtle differences in meaning is often a challenge in the absence of context or points of reference, which results in inter-rater inconsistencies, as well as varying judgments from the same rater. Further, raters are likely to be biased by the order of presentation, speed of association, and word frequency effects, as well as their own rating history. For instance, raters have been found to be influenced by how frequently they assigned a given score before: to avoid repeated judgments, they may be inclined to switch to a different one (Helson, 1964; Helson et al., 1954; Parducci, 1965; Wedell, 1995).

To address some of these challenges, modifications to the rating method have been proposed. For example, Bruni et al. (2014) circumvented the issue of arbitrariness of numerical scores by asking participants to express a binary preference, i.e., to decide which of two word pairs presented to them is more related (e.g., the words in the pair *car – wheels* are more related to each other than *wallet – moon*). The scores for each pair were later transformed through evaluation against 50 randomly selected comparison pairs (the ultimate score being equal to the number of times out of 50 that the pair was picked as the most related of the two), and then normalised to the $[0-1]$ interval to yield the 3,000-pair MEN dataset (Table 2.1). In another method based on relative judgments, *paired comparisons* (Dalitz and Bednarek, 2016), the task is to determine which of the two items at hand has more of a given property (e.g., is more positive). To deal with word frequency effects when constructing a rare word dataset, Luong et al. (2013) permitted annotators to give a "don't know" answer for pairs including unknown words, like Resnik and Diab (2000) before. Whereas the issue of judging word pairs in isolation has been addressed by eliciting judgments of words presented in sentential contexts, which takes into account context-dependence of meaning and triggers a specific sense of each word (Armendariz et al., 2020; Huang et al., 2012; Pilehvar and Camacho-Collados, 2019).

Further, some approaches avoid the pairwise setup altogether. For instance, in best-worst scaling (Asaadi et al., 2019; Avraham and Goldberg, 2016; Kiritchenko and Mohammad, 2017, 2016; Louviere et al., 2015; Louviere and Woodworth, 1991), participants are presented with sets of n items (usually four) and perform relative judgments to decide which displays a given property to the highest (*best*) and which to the lowest degree (*worst*). Each item is assessed a number of times in different sets, and the ultimate real-valued scores are obtained through the counting procedure (Orme, 2009) (i.e., the item's score is calculated as the proportion of times it was chosen as best

minus the percentage of times it was chosen as worst). Kiritchenko and Mohammad (2017) showed that for the same number of collected judgments, best-worst scaling yields more reliable results than the pairwise rating method. Another example is the task of outlier detection from clusters of semantically similar words (Blair et al., 2017).

Beyond pairwise word similarity or relatedness, representation models have also been evaluated on synonym detection datasets using English as foreign language test data (Landauer and Dumais, 1997; Turney, 2001) and word games (Jarmasz and Szpakowicz, 2003), where the aim is to identify one correct synonym of the target word among 4 candidates. Other types of evaluation benchmarks include word analogy (Gladkova et al., 2016; Mikolov et al., 2013a) and semantic relation datasets (Baroni and Lenci, 2011). Further, evaluation of word embeddings has also involved concept categorisation, or word clustering, where the task is to divide a set of words into subsets representing categories of similar words (Baroni et al., 2014) (e.g., *bicycle* and *car* would be clustered together as vehicles, *pear* and *apple* would form a cluster of fruits). Word vectors produced by the model are clustered using an unsupervised clustering algorithm (e.g., k-means) into n groups, based on the gold standard classes. In Chapters 4 and 5, I use the classes produced in the two-phase data collection protocol presented in this thesis to evaluate a selection of current representation learning models on this task.

Spatial Arrangement Method. A characteristic of the above discussed pairwise approaches is that the final set of word pairs does not cover all the possible pairings of the unique verb types in the sample. In other words, for any given word, only a subset of its similarities to the rest of the words is known. In psychology, similarity judgments have often been collected for the purpose of analyses using multidimensional scaling (MDS) (Borg and Groenen, 2005; Kruskal and Wish, 1978; Rabinowitz, 1975; Shepard, 1980), a set of exploratory data analysis techniques which allow visualising similarities spatially. Specifically, MDS computes a configuration of points in a k-dimensional space that best reflects the pairwise similarities in the data. The algorithm minimises a stress function which quantifies the fit between the distances in the k-dimensional space and those in the input data by iteratively reorganising the points in space, until an optimal configuration is attained. The product of MDS is a visualisation of the similarity relationship in the dataset and its underlying dimensions (Giguère, 2006; Nosofsky, 1992), which can shed light on the salient features impacting similarity judgments (e.g., size, intensity, polarity).

Given that MDS requires complete matrices of item-to-item similarity scores as input, many studies have relied on pairwise rating-based similarity data collection.

However, one of the main shortcomings of this approach lies in its poor scalability, as the number of possible pairwise combinations of items in the sample (n) grows as a quadratic function of its size ($(n^2 - n)/2$), and with it the time required for completing the study. To overcome this problem, Goldstone (1994) proposed the *spatial arrangement method* (SpAM), which elicits simultaneous judgments on multiple stimuli in a two-dimensional space. Rather than assigning a numerical score to each pair, participants arrange items geometrically to reflect their relative similarities, thus avoiding the problems associated with discrete rating scales (Section 2.2.2). Not only is the method fast, but it is also intuitive, tapping into the spatial aspects of human appreciation of similarity as closeness and dissimilarity as distance (Casasanto, 2008; Lakoff and Johnson, 1980). Upon finishing the arrangement, the matrix of pairwise dissimilarities is obtained from item-to-item Euclidean distances.

Among the method's strengths is the fact that it allows simultaneous consideration of all items in the set. In the pairwise method, this would require keeping in memory an unfeasibly long list of pairs (e.g., for 50 items presented in the spatial method there would be 1225 unique pairs to judge). As discussed earlier (Section 2.2.2), it is an important factor enabling consistent, comparative judgments, which also avoids changes in rating strategy and shifts in scores due to considerations external to the task (e.g., avoiding score repetitions). These properties make the spatial arrangement method an attractive alternative to pairwise ratings, however, its classic single-arrangement implementation has certain limitations. For one, there is the practical limitation of the size of the computer screen where the items are arranged, which imposes constraints on the sample size. Further, a single arrangement only allows appreciation of the relative similarities along two dimensions. This latter shortcoming has been addressed by Kriegeskorte and Mur (2012), who proposed a spatial *multi*-arrangement approach where subsets of stimuli are iteratively sampled from the whole set from trial to trial. This means that items clustered together in one trial are subsequently subsampled and arranged in a less crowded space, which allows for consideration of fine-grained differences between items previously judged as similar (e.g., *carrot, apple* and *pear* may be placed close together when surrounded by *car, bike, book, pen*, but separated into two groupings when judged on their own).

Given the potential of the spatial method for efficient acquisition of similarity judgments, in Chapter 4 I present a novel two-phase data collection paradigm which adapts the technique to lexical stimuli. Crucially, to address the restrictions imposed on the sample by the size of the computer screen, I introduce a precursor method based on manual semantic clustering which divides a large sample into manageable sets within

which meaningful similarity judgments can be made. I investigate the potential of the method to facilitate efficiently creating large-scale verb-focused resources composed of fine-grained spatial similarity judgments that allow for capturing the complexity of verb meaning in a multi-dimensional space. Although recent efforts contributed to opening up new evaluation possibilities in the domain of verb meaning (Gerz et al., 2016), the demand for challenging, wide-coverage lexical resources targeting verb semantics has not yet been fully met. As I demonstrate in Chapter 5, the two-phase method holds promise for alleviating the problem of resource scarcity thanks to its direct portability to languages other than English. To place this endeavour in context, the next section reviews the current approaches to multilingual dataset construction, highlighting some of the challenges involved in translation-based extension of existing English-language resources.

Multilingual Pairwise Datasets. Although English has dominated resource creation endeavours, the need to extend their coverage to other languages has been long recognised. This has been attempted either through monolingual data collection in the language of interest, or leveraging an existing English language resource through a translation-based method to speed up the construction process. In both cases, the go-to annotation paradigm is the pairwise rating method, where numerical scores are assigned to word pairs one by one.

Examples of monolingual resources created from scratch include pairwise datasets of semantic relatedness in German, created manually (Gurevych (2006), 350 pairs, including 103 verbs) and semi-automatically based on corpora (Zesch and Gurevych (2006), 222 pairs), as well as in Chinese (Wang et al. (2011b), 240 pairs), Turkish (Sopaoglu and Ercan (2016), 101 pairs) and Arabic (Saif et al. (2014), 40 pairs). Datasets focusing on semantic similarity include a 70-pair datset in Arabic (Almarsoomi et al., 2013) and two datasets in Chinese (Wu and Li (2016), 500 pairs; Huang et al. (2019), 960 pairs including 240 verb pairs). In Turkish (Ercan and Yıldız, 2018), a 500-pair dataset was created using a sampling procedure aimed at achieving a balanced coverage of different word frequencies, lexical relations, degrees of concreteness, and morphological properties. Further, Akhtar et al. (2017) manually created monolingual word similarity datasets for six Indian languages, Urdu, Telugu, Marathi, Punjabi, Tamil and Gujarati. Whereas Sakaizawa and Komachi (2018) proposed a large word similarity resource in Japanese, including 1,464 verb pairs.

In the translation-based approach, word pairs have been typically translated from one of the established English datasets, either by experts or through crowdsourcing, and

often re-annotated by native-speaker raters in the target language, to ensure that the scores faithfully capture the relatedness of the target words. Three English resources, the RG-65 dataset, WordSim-353 and SimLex-999, have been used for this purpose by a large number of works (Barzegar et al., 2018; Camacho-Collados et al., 2015a; Freitas et al., 2016; Granada et al., 2014; Gurevych, 2005; Hassan and Mihalcea, 2009; Konopík et al., 2017; Leviant and Reichart, 2015a; Netisopakul et al., 2019; Panchenko et al., 2016; Tóth, 2013; Van Tan et al., 2017). For example, Barzegar et al. (2018) employed professional translators to produce translations of these datasets into 11 languages (German, French, Russian, Italian, Dutch, French, Portuguese, Swedish, Arabic, Farsi, and Chinese), however, without re-annotating the translated pairs. Earlier, SimLex-999 was also translated by means of crowdsourcing into German, Italian, Russian (Leviant and Reichart, 2015a), Hebrew and Croatian (Mrkšić et al., 2017), and Polish (Mykowiecka et al., 2018). Vulić et al. (2020a) extended and translated SimLex-999 to cover a larger spectrum of word frequency, lexical fields, and concreteness levels, following a systematic translation and annotation procedure to guarantee cross-lingual consistency of the data. A different approach was adopted in the work of Camacho-Collados et al. (2017), who manually curated a set of 500 English word pairs, including named entities and multi-word expressions, ensuring a balanced distribution across the similarity scale and a wide coverage of different domains (e.g., biology, computing, history, health and medicine). The pairs were subsequently translated and re-annotated in Farsi, German, Italian, and Spanish. Both Vulić et al. (2020a) and Camacho-Collados et al. (2017) additionally automatically generated cross-lingual datasets by intersecting pairs of monolingual resources following the method of Camacho-Collados et al. (2015a).

Although dataset translation provides a convenient shortcut facilitating extensions to other languages, translation of pairs from one language to another entails certain challenges. First of all, languages vary in how they lexicalise different areas of meaning, differing in the number of lexical items used to cover the meanings associated with each semantic field. For instance, Germanic languages tend to have large inventories of verbs expressing the manner in which an action is performed (e.g., *peer*, *goggle*, *slither*, which express a manner of looking or moving). In contrast, languages such as Italian and Japanese (the so-called verb-framed languages (Talmy, 1985)) tend to express subtle meaning distinctions regarding manner outside the verb (Talmy, 2000), for example in adverbials:

(5) Il ragazzo beveva il suo tè rumorosamente.
the boy drink.IMP.3.SING his tea noisily
'The boy was slurping his tea.'

In the above sentence, the meaning component of the verb 'slurp' relative to drinking is expressed by *beveva* (Italian *bere*, 'to drink'), while the manner of drinking is captured by the adverb *rumorosamente* ('noisily'). In light of such mismatches, it is likely that multiple words in one language may translate into a single word in the target language, and vice versa, which creates problems for translating on a pair-by-pair basis. For instance, the English pair *drink – slurp* may translate to a single Italian word yielding an identity pair *bere – bere*, while the words *difficult* and *hard* (= difficult) have a single translation in Polish (*trudny*). A similar example can be found in Estonian, where there is no gender-marked word for *husband* or *wife*, only *abikaasa*, 'spouse', rendering direct translation of the pair *husband – wife* (e.g., found in SimLex-999) infeasible. In Chapter 5, I argue that the two-phase data collection paradigm proposed in this thesis circumvents these problems by starting from translating a verb sample, rather than verb pairs. This facilitates the translation process and avoids imposing the meaning distinctions present in the source language upon the target language, only to preserve a given word pairing (e.g., through periphrasis). Further, by permitting one-to-many and many-to-one translations between the source and the target language it allows adequate reflection of different patterns of polysemy and distinct lexicalisation patterns.

2.3 Lexical Knowledge Injection

Recent breakthroughs in representation learning coupled with an increased capacity of computational resources have enabled deep learning models to extract unprecedented amounts of information directly from large text corpora. Starting without any prior knowledge, these architectures develop an awareness of a range of linguistic phenomena through exposure to vast amounts of raw data. However, just like their predecessors, they are only able to acquire distributional knowledge available in text, and have been shown to fall back on superficial signals and task-specific shortcuts to excel in a given natural language problem. Therefore, there has been a growing interest in developing methods for supplying them with external knowledge, irretrievable from unlabelled text, to boost their language capacity. In this section, I give a brief overview of approaches to injecting lexical information through *specialisation* of static word embeddings and proceed to discuss more recent techniques aimed at infusing large pretrained models with discrete knowledge.

2.3.1 Semantic Specialisation

The popularity of pretrained word vectors (Mikolov et al., 2013a,b) generated a lot of interest in improving them using discrete knowledge from external databases. This has generally been approached from two different perspectives: (i) either by augmenting the training procedure with an additional training objective based on an external resource and jointly optimising it with the distributional objective, or (ii) by fine-tuning a pretrained word embedding space using the knowledge from such a resource. The shared goal is to draw the representations of words in a desirable relation close together (e.g., synonyms) while pushing words in undesirable relations further apart (e.g., antonyms). Several works explored the potential of *joint specialisation* (Bian et al., 2014; Kiela et al., 2015; Liu et al., 2015; Ono et al., 2015; Osborne et al., 2016; Xu et al., 2014). For example, the first method was employed by Yu and Dredze (2014) who proposed a combined training objective which maximised both the probability of the raw text and the probability of synonymy relations derived from WordNet and the Paraphrase Database (Ganitkevitch et al., 2013; Pavlick et al., 2015). They demonstrated that this added knowledge resulted in improved performance of the CBOW model of Mikolov et al. (2013a) on the tasks of language modelling and semantic similarity and relatedness estimation. Similarly, Nguyen et al. (2017) employed WordNet as a source of hypernym-hyponym pairs (e.g., *animal – cat*, *cat – Maltese*) in an extension of the SGNS model (Mikolov et al., 2013b) with two additional objective functions to learn hierarchical embeddings capturing hypernymy.

In the second approach, the semantic signal encoded by distributed word vectors is enriched by means of *retrofitting*: the input embeddings undergo fine-tuning to satisfy lexical constraints extracted from an external database (Faruqui et al., 2015; Jo and Choi, 2018; Lengerich et al., 2018; Mrkšić et al., 2016, 2017; Wieting et al., 2015). The constraints typically take the form of word pairs holding a lexical relation of interest, e.g., synonymy, hyper-hyponymy. This method has the advantage of portability, computational efficiency and flexibility, without being tied to a specific model, as it is in the case of joint specialisation approaches above. Specialising for semantic relations has been shown to benefit diverse downstream tasks, including lexical text simplification (Glavaš and Vulić, 2018b; Ponti et al., 2018), intent detection for spoken language understanding (Kim et al., 2016b), or detection of abusive language (Koufakou and Scott, 2020). For instance, Mrkšić et al. (2017) demonstrated that by *attracting* synonymous words and *repelling* antonyms in the embedding space, the produced representations better capture the distinction between similarity and relatedness. Indeed, evaluation on semantic clustering and word similarity using the

spatially induced data reported in Chapter 4 shows that these linguistically informed vectors are especially capable of accurately capturing fine-grained differences in word similarity, proving most robust across tasks and even outperforming the more recent large pretrained Transformer-based model. What is important, Mrkšić et al. (2017) also demonstrated that their representations boost downstream performance in dialogue state tracking and hold promise for cross-lingual knowledge transfer. By deriving cross-lingual semantic constraints (e.g., translation pairs like EN *dark* – IT *scuro*) from a multilingual semantic network, BabelNet (Navigli and Ponzetto, 2012), the embedding spaces of different languages can be merged into one shared cross-lingual space, for the benefit of task performance in the target language with limited monolingual resources available.

Further, Vulić et al. (2017b) demonstrated that this approach can be applied to other types of lexical information, specifically, shared class membership in an external lexical resource (e.g., VerbNet). Using cross-lingual translation links derived from BabelNet and English VerbNet class membership information, they tie source-target language pairs together in a bilingual embedding space and jointly specialise it to capture the structured information available in the source language. They then use the VerbNet-specialised representations as input to an unsupervised clustering algorithm and achieve state-of-the-art performance on automatic verb class induction in six target languages.

The line of research focused on extending the benefits of discrete knowledge available in resource-rich languages to resource-leaner languages has been continued in the work on *explicit retrofitting* and *post-specialisation*. These types of methods recognised the limitation of the classic retrofitting approaches, which only specialised the words seen in the particular external resources used. In the former (Glavaš and Vulić, 2018b, 2019) a global specialisation function (i.e., a deep non-linear feed-forward network) is learned using the lexical constraints as training examples to transform the entire embedding space. In the latter (Biesialska et al., 2020; Kamath et al., 2019; Ponti et al., 2018; Vulić and Mrkšić, 2018), a general mapping function is learned based on the transformations undergone by the words in the initial specialisation (i.e., by predicting the specialised vectors from the original ones) so as to propagate the external lexical-semantic signal to the entire vocabulary. What is crucial, the method can be used to port structured knowledge from one language to another, even completely lacking lexical resources. To this end, Ponti et al. (2019) proposed a cross-lingual specialisation transfer method based on Lexical Relation Induction, where pairwise constraints are automatically translated into the target language and subsequently

cleaned using a neural model for lexical-semantic relation prediction (Glavaš and Vulić, 2018a), before being employed as training examples for learning a global specialisation function. In Chapter 6, I demonstrate how the method can be adopted for the purposes of transferring and injecting verbal lexical knowledge into pretrained encoders for the benefit of downstream task performance across typologically diverse languages.

Alternatively, the work of Yuan et al. (2019) transferred semantic specialisation to low-resource languages by exploiting annotations of task-specific keywords from bilingual speakers to refine cross-lingual word representations for a given classification problem. Whereas Zhang et al. (2020) retrofit cross-lingual word embeddings to their training dictionary, pulling translation pairs closer in the embedding space. This allows them to fully exploit the information encoded in the training data, resulting in improved performance in document classification and dependency parsing.

2.3.2 Knowledge Injection into Pretrained Language Models

The success of large encoders pretrained on language modelling objectives (Devlin et al., 2019; Radford et al., 2018, 2019) has inspired investigations into the potential of supplementing their distributional knowledge with external structured information. Since they only learn from the co-occurrence signal in text corpora, they lack commonsense and factual world knowledge, as well as struggle to distinguish between fine-grained lexical relations. However, the contextualised representations yielded by these architectures impose new demands on knowledge injection methods, impeding direct application of specialisation approaches compatible with static word embeddings. Experimentation into the possibility of specialising these models has mainly targeted the Transformer-based BERT (Devlin et al., 2019), pretrained on masked language modelling and next sentence prediction (see also Section 2.2.1). The proposed approaches employed one of the following general strategies: masking, graph reshaping, and multi-task objectives, to infuse the model with external factual, commonsense, or linguistic information (He et al., 2019; Lauscher et al., 2020a,b; Levine et al., 2020; Peters et al., 2019; Xiong et al., 2019; Zhang et al., 2019b).

For example, Zhang et al. (2019b) acquire knowledge implicitly from training corpora by means of a two-fold masking strategy consisting of entity-level and phrase-level masking. As discussed in Section 2.2.1, in masked language modeling, BERT masks a proportion of input tokens and then conditions on both left and right context to predict the masked word's vocabulary ID. Here, Zhang et al. (2019b) mask multiple-word named entities (e.g., *Harry Potter*), identified via Named Entity Recognition, and entire phrases constituting single conceptual units, identified through chunking

(e.g., *a series of*). The aim is to induce the model to implicitly learn long distance semantic dependencies and develop an awareness of the interactions and relations linking different entities. The knowledge-aware model proves to achieve superior results on a range of tasks, including natural language inference, named entity recognition, and question answering.

In contrast, Liu et al. (2020b) exploit three Chinese knowledge graphs to boost the model with encyclopedic and linguistic knowledge by hybridising text with graphs. In particular, they transform each input sentence into a knowledge-rich tree structure augmented with triples from the knowledge graph (e.g., *Beijing – China – capital*), linked to entity mentions (e.g. *The national team traveled to Beijing last week*). To prevent external knowledge from distorting sentence meaning and avoid undesirable entity-word interactions, they proposed a special attention mask and soft-position embeddings which limit the impact of the external signal while keeping its structure intact. Their augmented model significantly outperformed the baseline BERT across twelve NLP tasks, both general and domain-specific (e.g., finance-oriented Named Entity Recognition and question answering), achieving a particularly significant advantage in the latter.

The objective of incorporating linguistic knowledge into BERT has been pursued by Lauscher et al. (2020b), who adapt the pairwise-constraint approach for the purpose of steering the pretrained language model to better distinguish similarity from relatedness via joint specialisation. Specifically, they augment BERT's pretraining tasks with an additional one, i.e., binary classification of lexical-semantic relations from the external databases. First, they employ two go-to lexical resources in work on semantic specialisation, WordNet and Roget's Thesaurus (Roget, 1911), to derive positive examples of synonymy and direct hyper/hyponymy relations. Negative examples are then generated by substituting each member of a positive pair with the semantically most similar word in the same batch, i.e., the nearest neighbour in an auxiliary static embedding space. Both positive and negative constraints are transformed into BERT input format and fed into a binary single-layer classifier.

Adapter-Based Fine-Tuning. The above discussed approaches have certain limitations. For example, methods relying on joint specialisation with additional knowledge-based training objectives require pretraining the whole network from scratch whenever the injected information undergoes any modification, which is computationally expensive and inefficient. Whereas post hoc fine-tuning of the pretrained encoder on external

knowledge, especially at a large scale, may lead to catastrophic forgetting[12] of the distributional information acquired in pretraining (Goodfellow et al., 2014; Kirkpatrick et al., 2017).

Recently, training with *adapters* has been proposed as a way of alleviating both these problems (Houlsby et al., 2019; Rebuffi et al., 2018; Wang et al., 2020a). The technique consists in adding a limited number of additional parameters to the underlying model and only tuning their values, i.e., at each training step only the adapter weights are updated and the original transformer parameters remain fixed. This allows for preserving the distributional signal intact, while acquiring new information on top in shorter training cycles. Further, adapters are compact and easily portable to new tasks, enabling integration with new types of structured or application-specific knowledge, without the risk of interference between multiple types of information. In the context of knowledge injection, Wang et al. (2020a) utilised the method to infuse BERT with factual knowledge from Wikidata (Vrandečić and Krötzsch, 2014), whereas Lauscher et al. (2020a) experimented with adapter-based injection of commonsense knowledge from ConceptNet (Liu and Singh, 2004; Speer et al., 2017) and the Open Mind Common Sense (OMCS) corpus (Singh et al., 2002).

In the last chapter of this thesis, I explore the potential of adapter-based injection of verb-specific semantic-syntactic knowledge derived from expert-curated resources. Moreover, I also investigate the impact of incorporating verb class membership information acquired via manual semantic verb clustering and spatial arrangements by non-experts, using the semantic datasets compiled as part of this thesis. I demonstrate that pretrained verb knowledge-aware adapters can boost model performance in event-oriented tasks where accurate verb processing is key. Crucially, the experimental results reveal the potential of plugged-in 'verb adapters' to boost zero-shot cross-lingual transfer from a resource-rich to resource-poorer languages, offering the advantages of computational efficiency and easy reusability.

2.4 Summary

The aim of this chapter was to contextualise the research presented in this thesis from three perspectives mirroring its central themes. First, I established a link between the present work and the inquiry into the organisation of verbal lexical knowledge. Drawing on a body of research into the interactions of verb semantic and syntactic behaviour, I

[12]The term *catastrophic forgetting* refers to a phenomenon whereby a neural network model completely forgets previously learned information as a consequence of exposure to new information.

hypothesised that there is potential in leveraging native speakers' non-expert intuitions about verb meaning *alone* to efficiently acquire class-based verb knowledge. This could be especially beneficial in languages lacking rich lexical resources such as those reviewed in Section 2.1.1. Then, I proceeded to examine approaches to intrinsic evaluation of representation learning models, focusing on word similarity estimation. I emphasised the need for challenging verb-focused benchmarks to enable scrutinising the ability of state-of-the-art models to make fine-grained meaning distinctions. Specifically, I argued that collecting human semantic similarity judgments by means of intuitive spatial arrangements offers the advantage of capturing the continuous nature of meaning, while avoiding some of the shortcomings of the traditional rating-based methods reviewed in Section 2.2.2. Finally, I discussed recent approaches to augmenting representation learning models with non-distributional information, introducing some of the key notions and techniques relevant for the latter part of this work: cross-lingual specialisation transfer and adapter-based fine-tuning. In what follows, I pursue these three research threads to explore (i) efficient bottom-up acquisition of verb knowledge from native speakers, (ii) its utility for focused evaluation and probing of current semantic models, and (iii) its potential to boost the capacity of NLP models to reason about events in different languages.

Chapter 3

Verb Class Induction Through Bottom-up Semantic Clustering

3.1 Introduction

Owing to the pivotal role played by verbs in sentence structure, the problem of creating verb classifications has attracted a lot of attention in natural language processing. Different approaches have been proposed to this end, varying with regard to the guiding criteria by which the class architecture is organised, prioritising semantic (WordNet (Fellbaum, 1998; Miller, 1995), FrameNet (Baker et al., 1998), PropBank (Palmer et al., 2005)) or syntactic information (COMLEX (Grishman et al., 1994), VALEX (Korhonen et al., 2006)), or combining the two (Kipper et al., 2000; Kipper Schuler, 2005; Levin, 1993). VerbNet (Kipper Schuler, 2005), grouping English verbs into classes defined by shared meaning components and syntactic behaviour, is one of the richest lexical verb resources currently available, and its utility in various NLP applications has been repeatedly demonstrated (Bailey et al., 2015; Lippincott et al., 2013; Rios et al., 2011; Schmitz et al., 2012; Windisch Brown et al., 2011).

However, creation of a similar resource from scratch, drawing simultaneously on semantic and syntactic criteria, is a challenging and time-consuming task when attempted by annotators without theoretical linguistics background (Majewska et al., 2018b). A number of automatic verb classification techniques have been developed (Falk et al., 2012; Joanis et al., 2008; Kawahara et al., 2014; Peterson et al., 2016, 2020; Scarton et al., 2014; Sun et al., 2010; Vulić et al., 2017b), allowing to minimise the time required and eliminate the need to employ trained lexicographers. However, evaluation of such systems relies on the availability of gold standard classes, and these are still lacking for a great majority of languages.

In light of these challenges and the high demand for verbal resources, in this chapter I investigate whether semantic verb classes can be reliably acquired from non-expert native speakers based solely on verb semantics and following simple instructions. Drawing on the hypothesis that syntactic and semantic behaviour of verbs are tightly interrelated (Jackendoff, 1990; Levin, 1993; Pinker, 1989), I simplify the classification task by eliminating the need to refer to explicit syntactic knowledge and assess whether intuitive native-speaker perception of closeness of verb meaning provides enough guidance to produce consistent verb classifications. Previous classifications have used syntactic behaviour to guide the construction of verb classification but this necessitates linguistic training.

In order to examine the potential of manual semantic clustering in different languages, I carry out verb clustering experiments with native speakers of English, Polish, and Croatian. I describe the setup of the task in Section 3.2. Subsequently, I analyse the inter-annotator agreement for each language individually and examine the overlap between classes cross-lingually. Section 3.3 includes the results of this evaluation. Finally, in Section 3.4, I discuss observations made with respect to the easily classifiable verbs and those which caused problems in all the languages considered, which shed light on cross-linguistic semantic commonalities and polysemy patterns.[1]

3.2 The Semantic Verb Clustering Task

The task involved a group of 8 native-speaker participants without formal linguistics training, 3 annotators for English and Polish, and 2 in Croatian, who performed soft clustering of a sample of verbs in their native language based on the verbs' semantic similarity. The verb samples were created as follows: first, a sample of 267 English verbs was automatically extracted from the pool of SimVerb-3500 (Gerz et al., 2016) verb types. The verbs were sampled so as to ensure that the top 34 VerbNet classes (according to the number of verbs in the class) from SimVerb-3500 are represented by at least 5 member verbs each, to guarantee 'clusterability' of the verbs presented to the annotators. Next, the English sample was translated by native-speaker translators into Polish and Croatian, and the three samples were manually inspected.

Before the start of the task, the annotators were provided with instructions (Appendix B.1) and a list of 267 verbs in a text file, presented in random order, one word in each line. Since the goal was to keep the task as simple as possible for participants without linguistics training, the annotation guidelines were intentionally minimal: they

[1]The results presented in this chapter have been published in Majewska et al. (2018a).

	English				**Polish**				**Croatian**		
	A1	A2	A3	**Ave**	A1	A2	A3	**Ave**	A1	A2	**Ave**
Number of classes	61	77	58	**65.3**	47	46	35	**42.7**	88	76	**82.0**
Average class size	4.4	3.5	4.6	**4.1**	5.7	5.8	7.6	**6.3**	3.0	3.5	**3.3**
Time spent [hours]	2	1	3	**2.0**	3	3	3	**3.0**	3	2	**2.5**

Table 3.1 Results and statistics of semantic clustering of 267 verbs for English, Polish, and Croatian, for each annotator (A1-A3) and the average scores for each language (Ave).

instructed the annotators to put verbs together using a spreadsheet program (e.g. Microsoft Excel) so as to form groups containing verbs that are used to express similar or related meanings. The groups could vary in size, but annotators were asked to aim for at least 3-5 members. A verb could be put in more than one class (e.g., when it had several different meanings), and any verb which did not seem to fit with any group could be placed in a 'Miscellaneous' class. Annotators were encouraged to make a note of any relationship links between groups where they felt the meanings of member verbs were in some way related, e.g., a bidirectional link between similar groups, or a unidirectional link between a broader class and its subclass(es).

3.3 Results and Inter-Annotator Agreement

The results and statistics of the semantic clustering task for each annotator individually and across annotators, in each of the three languages considered, are reported in Table 3.1. On average, it took 2.5 hours to complete the task across all annotators, ranging from 1 to 3 hours. The average number of classes obtained was 65.3 for English, 42.7 for Polish, and 82 for Croatian, with class size ranging from the average of 3.3 member verbs in Croatian to 6.3 members in Polish.[2]

3.3.1 Percentage IAA

In order to measure the overlap between classifications produced by annotators for each language individually and across languages, I calculate percentage inter-annotator agreement (% IAA) for all pairings of verbs. First, I extract all the pairs of verbs on which there is perfect agreement (i.e., all annotators either grouped them together or not), for each of the languages independently, and compute the ratio of observed

[2]The classes are made available online at https://github.com/om304/semantic-verb-classes.

	English	Polish	Croatian	All
% IAA	88.5%	92.5%	97.8%	79.9%

Table 3.2 The percentage inter-annotator agreement calculated for all possible pairings of verbs, for each language individually and across the three languages.

agreement pairs to all the possible pairings of verbs. Subsequently, I repeat the same procedure for all the English, Polish and Croatian annotators together.

The computations yield a high inter-annotator agreement score for each of the languages, with 88.5% observed for English, 92.5% in Polish, and 97.8% in Croatian (Table 3.2). The percent inter-annotator agreement calculated across all three languages is 79.9%. It must be noted that the very high agreement score obtained for Croatian, compared to the other two languages, is likely to be due to the smaller average class size. Since many Croatian classes included as few as 3 member verbs, there was a large number of pairs of verbs which were not classified together. Whenever the annotators agreed on not putting two verbs together, that pair constituted an 'agreement' pair for the purposes of inter-annotator agreement calculation. The smaller classes gave rise to the somewhat inflated % IAA score for Croatian because of the larger number of true negatives (verbs that are correctly found not to go in the same class). Its inclusion of true negatives gives % IAA rather high scores generally. In order to address this issue, in the following section (3.3.2) I calculate inter-annotator agreement using a different evaluation metric, Fuzzy B-Cubed for overlapping clusters (Amigó et al., 2009; Jurgens and Klapaftis, 2013),[3] which avoids the problem of inflation due to scoring true negatives.

3.3.2 B-Cubed for Overlapping Clusters

In the verb-clustering task, the total number of classes was left unspecified and the annotators were free to put a single verb in as many different classes as they felt was appropriate, whenever they recognised it had more than one distinct sense. In order to adequately evaluate the results, the evaluation measure applied to the data has to be able to accommodate these characteristics of the task. I chose the B-Cubed metric (Bagga and Baldwin, 1998) extended by Amigó et al. (2009) to compare overlapping clusters, and by Jurgens and Klapaftis (2013) to fuzzy clusters, used to evaluate the

[3] I used the Fuzzy B-Cubed implementation of Jurgens and Klapaftis (2013) but did not associate the clusters with weights, and therefore the metric is equivalent to that of Amigó et al. (2009).

	Average B-Cubed
English	0.262
Polish	0.338
Croatian	0.172
All	0.205
1c1inst	0.0
All-instances, One class	0.069

Table 3.3 The average B-Cubed F-score (i.e., harmonic mean of B-Cubed precision and recall) calculated for all possible pairings of annotators, for each language individually and across the three languages, and for two SemEval baselines: 1c1inst and All-instances, One class.

performance of Word Sense Induction systems in SemEval tasks (Jurgens and Klapaftis, 2013).

The B-Cubed metrics (B-Cubed precision and recall) compare two clusterings (say, X and Y) at the item level: for an item i, precision measures how many items sharing a cluster with i in clustering X are placed in its cluster in clustering Y; whereas B-Cubed recall measures how many items sharing a cluster with i in Y are also placed in its cluster in X, with the final B-Cubed score equivalent to the harmonic mean of the two values.

In this task, rather than comparing each clustering against a gold-standard set of classes, I calculate the B-Cubed score for each pair of clusterings produced by the annotators, for each language individually and across all three languages. I report the results of this evaluation in Table 3.3. The highest agreement score is observed for Polish, where the average B-Cubed F-score is 0.338. Less overlap was found between the clusterings produced for English (0.262), with the lowest B-Cubed F-score obtained for the Croatian clusterings (0.172). The low score reported for Croatian is especially noteworthy in light of the inflated percent agreement result reported in Section 3.3.1. With percent agreement computed for every possible pairing of verbs, based on a binary choice between two verbs being either clustered together or kept separate, the two annotators seemed to agree in a vast majority of their clustering decisions. Applying an alternative evaluation metric allows for identifying the bias from scoring true negatives, i.e., all cases in which the annotators agreed that two verbs should not be clustered together. As predicted, this inflation is particularly high in the case of Croatian due to the small average class size compared to the other two languages. Indeed, manual inspection of the classes produced by the Croatian annotators shows that in some

cases the minimum class size of 3-5 members recommended by the guidelines was not adhered to.

The average B-Cubed F-score calculated for all possible pairings of annotators across the three languages (using translational equivalents for cross-lingual comparisons) is 0.205. Notably, the average cross-lingual agreement score is higher than the value obtained for Croatian itself, which suggests a promising degree of overlap between English and Polish classes (the average B-Cubed F-score for these two languages is 0.237).

Keeping in mind the differences in the nature of the present task and a Word Sense Induction task (which can be seen as an example of unsupervised clustering, with usages of a word grouped into clusters, each representing uses of the same meaning (Jurgens and Klapaftis, 2013)), comparing the results against the scores obtained by the SemEval participating systems may help interpret the reported values. Overall, the top-performing system surpasses the highest result for Polish (scoring 0.483), on the other hand, in the multi-sense setting (i.e., on instances labelled with multiple senses), the best performing system achieves the B-Cubed score of 0.134, a result below the lowest agreement score in the present task.

In order to make the comparison more meaningful, I calculate two SemEval baselines for the present task: (1) 1c1inst, where each instance is assigned to its own class, and (2) All-instances, One class, which assigns all instances to a single class. The result for the first baseline, 0.0, is the same as in SemEval, and a natural consequence of B-Cubed since there are no pairs within a class. However, while the overall performance of the All-instances, One sense baseline in SemEval surpasses its best participating system (achieving the score of 0.623), the result for this baseline on the present verb clustering task is much lower (0.069), suggesting the task is significantly more difficult, due to the high number of clusters. And yet, despite the greater difficulty of the task, the agreement between the annotators exceeds the performance of the baselines, which is an encouraging outcome.

As noted earlier, the present verb clustering task and the SemEval task are different. The SemEval annotation was performed using predefined senses for graded-tagging (on a Likert scale) and the systems' clusters were compared to clusters induced from these graded sense annotations. Since the senses for consideration by the annotators were defined in WordNet this is not comparable with the present task of clustering verbs. The present task allowed for complete flexibility in the number of classes, which resulted in varying levels of granularity (e.g., Croatian classifications had up to 88 clusters, while the smallest Polish clustering had 35), and a higher number of clusters

overall with respect to the SemEval task. The different setups of the two tasks entail different levels of difficulty and hence, different agreement scores, and this is reflected in the results obtained for the same baselines discussed above. What is important, this analysis constitutes the first attempt at measuring agreement on clustering of verbs performed by humans.

The encouraging degree of overlap observed between the classifications produced in the manual clustering task, particularly for Polish and English, suggests that there are consistent patterns in how humans group verbs based on their semantic similarity, not only in each language independently, but also across languages from different language families. Collecting more classification data for Croatian, while controlling for class size (as per the minimum class size stated in the guidelines), will allow us to verify whether the lower B-Cubed score reported for that language has to do with the peculiarities of the collected data or is indicative of a general greater difficulty in classifying verbs in Croatian with respect to the other two languages. Extending the experiments to other diverse languages will allow for investigating even further the degree to which those regularities are observed cross-linguistically; however, these are already promising inter-annotator agreement results for a multilingual semantic task.

3.3.3 Cross-Linguistic Areas of Overlap

Manual inspection of the resultant classes from all annotators allows for observing what class types and semantic domains are shared by the three languages. Five classes emerge which share the core of at least 2 member verbs across annotators in all three languages (with extra members added by some annotators) and can be described with the following labels (denoting 'verbs of_'): 'looking', 'cooking', 'existing', 'movement in water', 'emitting sound'. The total of 30 classes can be identified where the core of at least 2 member verbs is shared by at least two languages (by all annotators), and whose members belong to the same semantic domains across languages (but with more variation in specific member verbs recorded by individual annotators). In Section 3.4, I look more closely at the semantic patterns observable in all three languages and discuss which aspects of verb meaning make the classification task consistently easier or harder, regardless of the language in question. While the interpretation of the cross-lingual patterns discussed below is limited by the small number of participants involved in the present study, this preliminary investigation sheds light on the way non-expert speakers of different languages tackle the clustering task, as well as the potential sources of cross-lingual variation and its impact on the clustering output.

3.4 Analysis and Discussion

Despite the encouraging inter-annotator agreement scores, several issues affecting the agreement and overlap between the resultant classifications could be observed. First of all, as the task did not impose a fixed number of classes, the levels of granularity varied between annotators: the difference between the minimum and maximum number of classes equals 12 for Polish and Croatian, and 19 for English. This discrepancy is even more noticeable across languages: while Polish annotators grouped verbs into 35-47 classes, Croatian classifications comprise between 76-88 verb classes. As the task consisted in grouping verbs into flat classes, the resultant classifications do not capture hierarchical relationships between verb groups (these could, however, be signalled as 'relationship links', as noted above). Therefore, potential inclusion of one class by another (e.g., in the case of 'movement' verbs, which in Croatian are split into two small classes, depending on the medium (water vs. ground), and are grouped together in one broader 'movement' class in Polish (*swim, dive, walk, crawl*)), is interpreted as class disjunction in automatic pairwise evaluation, which results in a lower overlap between the classifications. What is more, in some cases distinct patterns of ambiguity in the languages considered resulted in different clustering decisions. For example, while in English two senses of the verb *shine* (i.e., emit light and polish (a shoe)) were considered, resulting in pairings *shine – glow* and *shine – brush*, only the former sense is available in Polish and Croatian.

3.4.1 Problematic and Easily Classifiable Verbs

In order to investigate whether some verbs are inherently easier or harder to classify, and examine to what extent this is observed across languages, I extracted all the pairs of verbs on which there is perfect agreement and those on which the annotators disagreed for each language individually, and examined the overlap between these groups of verbs across the three languages. This allowed identification of 72 'problematic' and 24 'easy' verbs, shared by the three languages. Manual inspection of these groups allows for making a number of observations regarding the aspects of verb meaning which pose problems or make them easier for humans to classify, regardless of the language considered.

'Problematic' Verbs

Most of the verbs which ended up in the 'problematic' group share the characteristic of having a broad, vague or abstract meaning, sometimes with several related senses

which allow them to appear in a number of slightly different contexts. For example, annotators in all three languages disagreed on how to classify verbs such as *affect, treat, engage* or *spare*. What is more, some display a degree of semantic vacuity, that is, have little semantic content of their own and tend to express a more precise meaning when combined with some other word (e.g., a noun), with which they form a predicate, such as *make* or *have*, examples of the so-called 'light verbs' (Jespersen, 2013). Inspection of the VerbNet classes from which the 'problematic' verbs were sampled revealed that the 'Change of State' class (45) is particularly often represented. Examples of class members include verbs such as *slip, vary, tumble*, whose meanings may not immediately appear as related. Moreover, each has several senses, which is reflected in the fact that each participates in a number of distinct VerbNet classes. Understandably, this results in more variation in clustering decisions, as different annotators are likely to take different verb senses into consideration, and consequently, produce divergent classifications.

'Easy' Verbs

Verbs which lend themselves better to manual semantic classification are those with narrow, concrete meanings, for example, verbs describing sounds (*chirp, buzz, roar*) or those belonging to a clearly defined semantic field, e.g., 'cooking' verbs (*fry, bake, cook*). Synonymous verbs such as *study* and *examine*, or *observe* and *stare*, were also among those on which the same clustering decisions were made across annotators, in all three languages. Interestingly, there was full agreement on antonymous pairs such as *vanish* and *appear*, which were consistently grouped together in all languages. As discussed in lexical semantics literature (Cruse, 1986), antonyms have a paradoxical nature. On the one hand, they constitute the two opposites of a meaning continuum, and therefore could be seen as semantically remote, on the other hand, they are paradigmatically similar, having almost identical distributions, and hence seem closely related. Despite these conflicting properties of antonyms, humans seem to intuitively recognise their relatedness and consistently group them together. The perception of relatedness overrides the sense of 'oppositeness' and being maximally distant along a dimension of meaning, and opposites end up clustered together. This regularity is observed in the case of pairs of relational antonyms, i.e., verbs which describe an event from opposite points of view, for example, *lend* and *borrow*, which differ along only one dimension of meaning, that is, the direction of the action (the object of the verb either travels away from the participant (*A lends something to B*) or towards the participant

(*B borrows something from A*)), and are essentially identical with regard to all other features, which makes them appear semantically close.

3.4.2 Semantic Similarity vs. Relatedness

The importance of distinguishing between the concepts of semantic similarity (e.g., *cup* and *mug*) and relatedness (e.g., *coffee* and *cup*) has been noted in the literature (Hill et al., 2015), and the analysis of the present data provides more evidence illustrating the influence of loose association on how humans conceptualise similarity between words, and the difficulty of keeping similarity and relatedness apart. In all three languages we can observe instances of what can be described as a 'storyline approach' to judging semantic similarity and verb classification. This is particularly noticeable in Croatian classifications, where several classes formed by the annotators group verbs describing quite different actions, linked via loose thematic ties: (1) *marry, conquer, approach, move*, where putting semantically dissimilar verbs *marry* and *move* together seems to suggest an underlying 'storyline' with courtship leading to marriage and moving house; (2) *visit, communicate, treat, operate*, where the association of verbs *visit* with *treat* and *operate* brings to mind a hospital visit, or (3) *finish, frame, announce, submit*, which can be seen as belonging to an 'academic' thematic domain. Relying on association rather than actual consideration of semantic components of verbs' meaning is visible in the cases where a class contains verbs which express a consequence of the action or state described by other verbs in the same class, e.g., *glow, shine, squint* in one of the Polish classifications, with 'squinting' being a reaction to 'glowing' or 'shining', or *ache, hurt, kick, rub, cry* in Croatian. Although verb groupings in which loosely thematically related verbs are classified together are in the minority, their presence in the data suggests that, in order to obtain classes based solely on semantic similarity judgments, unbiased by loose association, the annotation guidelines should explain the similarity-relatedness distinction and instruct the annotators accordingly.

3.4.3 Polysemy

An in-depth investigation of the resultant classes also offers an insight into the patterns of polysemy in the three languages considered. In the present task, the annotators could accommodate a verb's ambiguity by placing it in several different classes, putting each of its distinct senses in a separate cluster. However, since the annotators were provided with just word forms and the senses were not specified a priori, there were some discrepancies in which senses were identified, across annotators and, expectedly,

across the different languages, which led to a lower cross-lingual agreement in the resultant classes. For example, the Croatian translation of the English verb 'to vary', *odstupati*, expresses not only the sense of 'differing', but also 'withdrawing', unavailable in English or Polish, which explains why it was placed in the same class with *move* and *renounce* only by the Croatian annotators. Analogously, the Croatian equivalent of *remark* (*primijetiti*), ambiguous between senses 'to comment' and 'to notice', ended up together with verbs such as *look, stare, observe*, while in Polish and English it was grouped with verbs of 'communicating'. Similarly, the Polish translation equivalent of the verb *weave* (*pleść)* is ambiguous between two senses, 'to interlace' and 'to blabber', and was grouped both with *join* and *combine*, and with *tell* and *communicate*, while no such ambiguity was recorded in English and Croatian. Finally, while two senses of the verb *sway* ('to move rhythmically from side to side' and 'to control or influence') are available in English, only the former is recognised in the Polish and Croatian classifications and its translation equivalents are never grouped together with verbs such as *convince* or *persuade*, as it is the case in English.

In the present task, where guidelines were intentionally restricted, so as to avoid imposing any preconceived semantic categories or classification structure onto the annotators and elicit possibly spontaneous similarity judgments, such discrepancies in detecting ambiguity are inevitable. In order to have more control over which sense of a given verb is taken into consideration in the clustering task, word senses rather than word forms would have to be provided at the start of the task. For instance, short phrases with the target verb accompanied by a direct object could be used to facilitate sense disambiguation (e.g., *ran the track* vs. *ran the shop* (Brown, 2008)). Such a setup would also allow comparison of the elicited classes with the existing multilingual sense inventories, like Open Multilingual WordNet (Bond and Foster, 2013) or BabelNet (Navigli and Ponzetto, 2012). Since the aim of the present study was to elicit judgments on basic word forms, without any guidance as to the different word senses available, such comparisons are beyond its scope.

3.5 Conclusion

In this chapter, I have presented the first cross-lingual analysis and evaluation of semantic clustering of verbs by non-expert human annotators. The inter-annotator agreement scores reported for English, Polish, and Croatian are encouraging and demonstrate that verbs can be reliably classified by humans without a linguistics background. What is important, this suggests that there is potential to create verb

classifications starting from a simple, purely semantic task. Moreover, the degree of overlap in the resultant classifications observed across languages implies that there are cross-linguistic commonalities and shared meaning components governing the semantic organisation of verbs. A cross-lingual scrutiny of low-agreement verbs and those on which annotators made identical clustering decisions allowed for investigating to what extent the same verbs are problematic and whether some verbs are inherently easier to classify. Manual inspection of the thus identified 'easy' and 'problematic' verbs provided interesting insights into the aspects which may affect 'clusterability' of verbs across different languages (e.g., vague or abstract meanings).

The small-scale manual clustering experiments provide several methodological insights that may inform similar projects in the future. First, while the simplicity of the chosen annotation tool (i.e., a spreadsheet) permits quick data collection, it has limited scaling-up potential: Even with the current sample of 267 verbs, the task of keeping track of the created clusters proved arduous for the study participants. Further, the analysis of the discrepancies in clustering granularity across annotators revealed that semantic clusters are naturally hierarchical, with narrow clusters of similar verbs in one clustering solution subsumed by a larger grouping in another clustering. Therefore, using an annotation interface that would explicitly encourage creating cluster hierarchies could yield results more accurately capturing the annotators' perceptions of similarity and relatedness among cluster members. Moreover, given that polysemy is an important source of variation in clustering solutions, incorporating minimal disambiguating contexts into the task could help constrain the set of senses considered for each verb and thus improve inter-annotator alignment.

In the next chapters, I build on these analyses and findings and use the idea underlying the method validated in this preliminary study as a starting point for large-scale, two-phase semantic data collection.

Chapter 4

Semantic Dataset Construction from Clustering and Spatial Arrangement

4.1 Introduction

Recent advances in representation learning have transformed the NLP landscape introducing new powerful architectures that achieve unprecedented results on a plethora of natural language tasks (Devlin et al., 2019; Liu et al., 2019d; Peters et al., 2018; Radford et al., 2019; Yang et al., 2019b, *inter alia*). While high performance in downstream tasks may be the ultimate goal,[1] intrinsic evaluation benchmarks continue to provide a useful intermediary test of representation quality, with the advantages of simplicity and speed of execution. Estimation of lexical semantic similarity has been widely used as an intrinsic evaluation task, where the quality of word embeddings is assessed through comparison of distances between words in the embedding space against human judgments of semantic similarity and/or relatedness (Agirre et al., 2009; Bruni et al., 2014; Finkelstein et al., 2002; Hill et al., 2015). Further progress relies on the availability of high-quality evaluation benchmarks, challenging enough to test the limits of models' capacity to capture word semantics. However, these are still limited to a small number of typically well resourced languages. Moreover, they predominantly focus on nouns, and less attention has been paid to the challenges posed to natural language models by the complex linguistic properties of verbs. Due to the verbs' central role in sentence structure as bearers of information pertaining to both

[1] Another goal may be modelling of human language reflecting cognitive performance for scientific purposes.

structural and semantic relationships between clausal elements, attaining accurate and nuanced representations of their properties is essential to decrease the gap between human and machine language understanding (Altmann and Kamide, 1999; Jackendoff, 1972; Levin, 1993; McRae et al., 1997; Resnik and Diab, 2000; Sauppe, 2016, *inter alia*).

While recent efforts resulted in a large verb similarity dataset for English, SimVerb-3500 (Gerz et al., 2016), the demand for challenging, wide-coverage lexical resources targeting verb semantics has not yet been fully met. Expert-built lexicons encoding rich information about verbs' semantic features and behaviour such as FrameNet (Baker et al., 1998) or VerbNet (Kipper et al., 2006; Kipper Schuler, 2005) are still only available in a handful of languages, and noun-focused benchmark datasets are prevalent (Agirre et al., 2009; Bruni et al., 2012; Finkelstein et al., 2002; Hill et al., 2015). In light of these considerations, in this chapter I introduce a methodology which promises to mitigate the evaluation data scarcity problem and help overcome the bottleneck of slow and expensive manual resource creation.

I present a novel approach to collecting semantic similarity data by means of a two-phase design consisting of (1) *bottom-up semantic clustering* of verbs into theme classes, and (2) *spatial similarity judgments* obtained via a multi-arrangement method so far employed exclusively in psychology and cognitive neuroscience research and with visual stimuli (Charest et al., 2014; Kriegeskorte and Mur, 2012; Mur et al., 2013). I demonstrate how it can be adapted for the purposes of a large-scale linguistic task with *polysemous lexical stimuli* and yield wide-coverage verb similarity data. The method's promise lies in the intuitive nature of the task, where relative similarities between items are signalled by the geometric distances within a two-dimensional arena, as well as a user-friendly drag-and-drop interface. These properties of the annotation design significantly facilitate and speed up the task, as many concurrent similarity judgments are expressed with a single mouse drag. What is more, no classification structure or criteria are pre-imposed on the annotators, and similarities between individual verbs are judged *in the context of other verbs* appearing in the arena, rather than in isolation. Crucially, the method enables word clustering and registers pairwise semantic similarity scores *at the same time*, which can be especially beneficial as a means of rapid creation of evaluation data to support NLP.

The final resource comprises 17 theme classes and *SpA-Verb*, a large intrinsic evaluation dataset including 29,721 unique pairwise verb (dis)similarity scores for 825 target verbs.[2] The SpA-Verb scores are Euclidean distances corresponding to

[2]The resource is available online at https://github.com/om304/SpA-Verb.

dissimilarities between words, assembled in the representational dissimilarity matrix (RDM) (Kriegeskorte et al., 2008).[3] It surpasses the largest verb-specific evaluation resource previously available, SimVerb with 3,500 pairwise similarity scores, by a significant margin. Thanks to its scale and vast coverage, as well as its inclusion of complete matrices of pairwise similarities for all possible pairings of verbs within a given class, SpA-Verb offers a wealth of possibilities for nuanced analyses and evaluation of semantic models' capacity to accurately represent concepts pertaining to different meaning domains and displaying different semantic properties. In Section 4.8, I demonstrate the resource's utility by evaluating a selection of state-of-the-art representation learning architectures on two tasks, corresponding to the two phases of the design: (1) clustering, using Phase 1 classes as gold truth, and (2) word similarity, using pairwise scores from the entire SpA-Verb (29,721 pairs) and the thresholded subset of 10k+ pairs, as well as selected subsets focusing on different semantic characteristics.

The chapter is organised as follows. Section 4.2 briefly discusses related work, existing datasets, and alternative annotation protocols. Section 4.3 presents the structure and motivation of the proposed annotation design and discusses the challenges involved in adapting the spatial arrangement method from visual to lexical stimuli. The two phases of the protocol are analysed in Sections 4.4 and 4.5, respectively. The results of the inter-annotator agreement analysis are discussed in Section 4.6, while Section 4.7 presents an in-depth examination of the semantic information captured in each phase, by means of qualitative and quantitative comparative analyses with existing lexical resources. Section 4.8 presents the results of the evaluation of a diverse selection of representation models on the created dataset.[4]

4.2 Related Work

Recent years have seen word representation learning take centre stage in NLP, with novel architectures pushing performance to new heights. Further advances rely on the availability of high-quality evaluation resources, which are still limited, and few and far between. Rich expert-created resources such as WordNet (Fellbaum, 1998; Miller, 1995), VerbNet (Kipper et al., 2006; Kipper Schuler, 2005), or FrameNet (Baker et al., 1998) encode a wealth of semantic, syntactic and predicate-argument information for English words, but are expensive and time-consuming to create. Crowdsourcing with non-expert

[3]This effectively means that lower scores are assigned to similar verbs, and larger scores to dissimilar verbs.

[4]The research presented in this chapter has been published in Majewska et al. (2020a, 2021a).

annotators has been adopted as a quicker alternative to produce evaluation benchmarks. Semantic models have been predominantly evaluated on datasets consisting of human similarity ratings collected for sets of word pairs (Baroni et al., 2014; Bojanowski et al., 2017; Dhillon et al., 2015; Levy and Goldberg, 2014; Mrkšić et al., 2017; Pennington et al., 2014; Schwartz et al., 2015; Wieting et al., 2016).

Although widely useful, most of the datasets used for intrinsic evaluation are restricted in size and coverage, many conflate similarity and relatedness, and only a few target verbs specifically. Amongst the English language resources used for intrinsic evaluation of semantic models, word pair datasets such as WordSim-353 (Agirre et al., 2009; Finkelstein et al., 2002), comprising 353 noun pairs, and SimLex-999 (Hill et al., 2015), comprising 999 word pairs out of which 222 are verb pairs, have been prominent. Resources focused exclusively on verbs include YP-130 (Yang and Powers, 2006) (130 verb pairs) and the dataset of Baker et al. (2014) (143 verb pairs), with the more recent addition of SimVerb (Gerz et al., 2016) providing pairwise similarity ratings for 3,500 English verb pairs.

While pairwise rating datasets have been ubiquitous in intrinsic evaluation, alternative annotation methodologies and types of datasets have been proposed to address some of their limitations. Examples include *best-worst scaling* (Asaadi et al., 2019; Avraham and Goldberg, 2016; Kiritchenko and Mohammad, 2017, 2016; Louviere et al., 2015; Louviere and Woodworth, 1991), where annotators perform relative judgments of several items to decide which displays a given property to the highest and which to the lowest degree, and *paired comparisons* (Dalitz and Bednarek, 2016), where the task is to determine which of the two items at hand has more of a given property. Further, as an alternative to the words-in-isolation approach, datasets composed of judgments of similarity in context have been constructed (Armendariz et al., 2020; Huang et al., 2012; Pilehvar and Camacho-Collados, 2019), where target words are presented in sentential contexts triggering a specific meaning of each word. Representation models have also been evaluated on synonym detection datasets using English as foreign language test data (Landauer and Dumais, 1997; Turney, 2001), word games (Jarmasz and Szpakowicz, 2003), where the aim is to identify one correct synonym of the target word among 4 candidates, and on analogy (Gladkova et al., 2016; Mikolov et al., 2013a) and semantic relation datasets (Baroni and Lenci, 2011).

The most extensive verb-oriented dataset available to date, SimVerb-3500 (hereafter SimVerb), is a product of a crowdsourcing effort with over 800 raters, each completing the pairwise similarity rating task for 79 verb pairs. In this chapter, I describe an alternative novel approach which enables an annotator to implicitly express multiple

pairwise similarity judgments by a single mouse drag, instead of having to consider each word pair independently. This allowed for scaling up the data collection process and, starting from the same sample of verbs as those used in SimVerb, generated similarity scores for over eight times as many verb pairs. Consideration of multiple items concurrently also provides some context for ambiguous words while not relying on sentential contexts, which give rise to issues of sparsity and coverage, and are therefore less amenable to building larger lexical resources. Moreover, the proposed approach also yields relatedness-based item classes thanks to a precursor semantic clustering method, within which the similarity judgments are made.

4.3 Multi-Arrangement for Semantics

4.3.1 Spatial Arrangement Method (SpAM)

The spatial arrangement method (SpAM) has been previously employed to record similarity judgments through geometric arrangements of visual stimuli in psychology and cognitive neuroscience (Charest et al., 2014; Goldstone, 1994; Hout et al., 2013; Kriegeskorte and Mur, 2012; Levine et al., 1996; Mur et al., 2013). However, its potential and applicability to *semantic similarity between lexical stimuli* has not yet been studied.

In the commonly used pairwise rating method (e.g., employed to produce SimVerb) a rater is presented with a pair of words at a time and the number of possible pairwise combinations of stimuli grows quadratically as the sample size increases. For a sample of n stimuli there are $n(n-1)/2$ pairwise combinations possible. However, in SpAM a subject arranges multiple stimuli simultaneously in a two-dimensional space (e.g., on a computer screen), expressing (dis)similarity through the relative positions of items within that space: Similar items are placed closer together and dissimilar ones further apart. The inter-stimulus Euclidean distances represent pairwise dissimilarities and all stimuli are considered in the context of the entire sample presented to the user. Each placement simultaneously signals the similarity relationship of the item to all other items in the set. Figure 4.1 illustrates this comparison.

SpAM leverages the spatial nature of humans' mental representation of concept similarity (Casasanto, 2008; Gärdenfors, 2004; Lakoff and Johnson, 1999) and allows for a freer, intuitive expression of similarity judgments as continuous distances, rather than necessitating assignment of discrete numerical ratings. The latter, although omnipresent in intrinsic evaluation of representation models as a handy approximation

Word1	Word2	Rating	Word1	Word2	Rating
hammer	pliers	4	platter	pear	0
pliers	cup	1	banana	capsicum	4
platter	cup	4	pear	capsicum	4
hammer	cup	1	hammer	capsicum	0
platter	pliers	1	pliers	capsicum	0
banana	cup	0	platter	capsicum	0
banana	pliers	0	cup	capsicum	0
banana	hammer	0	cup	pear	0
banana	platter	0	hammer	pear	0
pear	banana	4	hammer	platter	1
pear	pliers	0			

Fig. 4.1 Comparison of the SpAM method with visual and lexical stimuli, and the pairwise rating approach, on a toy set of concrete real-world concepts. The 7-item sample generates 21 unique pairings of items in the pairwise rating method (example numerical ratings are given for illustrative purposes). In SpAM, placements of items express relative similarities: artefacts *pliers, hammer, platter, cup* are closer together than the fruit; Within the fruit group, *capsicum* is closer to *pear* than *banana*, while *pliers* and *hammer*, and *plate* and *cup*, form two smaller clusters of similar items. Images used in the diagram courtesy of the MRC Cognition and Brain Sciences Unit (University of Cambridge) and the Open Images Dataset (Kuznetsova et al., 2020).

of the strength of lexical relations, have been shown to have a number of limitations (Batchkarov et al., 2016; Faruqui et al., 2016; Gladkova and Drozd, 2016; Kiritchenko and Mohammad, 2017). Rather than reflecting semantic factors, annotators' judgments of isolated word pairs are often found to be biased by word frequency, prototypicality, order of presentation and speed of association. Moreover, subtle meaning distinctions and degrees of similarity between words are very difficult to quantify and translate onto a discrete scale without context or points of reference, in the form of other related words. As a result, the collected judgments are prone to inconsistencies, both across annotators and within the same annotator. SpAM helps address shortcomings of the absolute pairwise ratings by allowing repeated multi-way, relative continuous similarity judgments, which produce evaluation data capturing the complexity of lexical relations in continuous semantic space.

In this work, I adapt the multi-arrangement method proposed by Kriegeskorte and Mur (2012), which uses inverse multidimensional scaling to obtain a distance matrix from multiple spatial arrangements of subsets of items within a 2D space. The subjects are presented with subsets of items designed by an adaptive algorithm aimed at providing optimal evidence for the dissimilarity estimates. They are asked to drag and drop the stimuli within a circular *arena* on the computer screen, placing items perceived as similar close together and those dissimilar further apart (see Figure 4.1 again). The method has been shown to have high test-retest reliability (Spearman's $r = 0.93$, $p < 0.0001$) and to yield similarity data which strongly correlate with

those acquired by means of the traditional pairwise similarity judgment approach (Spearman's $r = 0.89$, $p < 0.0001$) (Kriegeskorte and Mur, 2012).

The first arrangement of all items within a sample provides an initial estimate of the representational dissimilarity matrix (RDM). Subsequently, the subject continues work on subsets sampled from the entire stimuli set. The adaptive subset selection algorithm elicits repeated judgments on items placed close together in the previous trial to ensure that enough evidence is collected for the relative distances between the similar items and for each possible pairing (Figure 4.6). The process can be terminated at any time after the first arrangement onward, but an earlier termination entails a potentially noisier final RDM. The subject is instructed to use the entire space available for each consecutive arrangement. This allows them to spread out items previously clustered together, thus reducing bias from placement error. The relative inter-item distances, rather than the absolute screen distances, represent dissimilarities between the items from trial to trial. The RDM estimate is updated after each trial and the collected evidence is statistically combined to yield the final RDM (for details of the algorithm see Kriegeskorte and Mur (2012)). The thus obtained pairwise dissimilarity scores for each class are normalised by scaling each distance matrix to have a root mean square (RMS) of 1 to guarantee inter-class consistency (Eq. 4.9).

In order to adapt the multi-arrangement approach for the purposes of the present task I had to address two key challenges, previously unsolved by SpAM-based methods: *scalability* and *semantic ambiguity*. So far, cognitive science research has applied SpAM to fairly small stimuli sets (≈100 items). Moreover, preliminary tests carried out in the early stages of the present work revealed that larger samples are technically and cognitively difficult for human subjects. First of all, the size of the arena within which the items are arranged is restricted by the dimensions of the computer screen (Figure 4.6). With samples larger than 100 items the arena becomes overcrowded, which makes it difficult to distribute the items as needed. What is more, longer sessions increase participant fatigue and thus affect judgment quality and consistency. Second of all, lexical stimuli are semantically ambiguous: Without multiple sense labels, annotators consider different word senses, which results in different similarity judgments.

In the following sections, I describe a new SpAM-inspired framework that resolves both the issue of scalability and semantic ambiguity, and demonstrate how these key challenges are addressed by the proposed two-phase study design.

4.3.2 Two-Phase Design

The annotation process is structured as follows. First, in a *rough clustering phase* (Phase 1), the large starting sample is divided into smaller, broad classes of semantically similar and related verbs. Second, in a *spatial multi-arrangement phase* (Phase 2) the verbs placed in the classes in the previous phase are repeatedly arranged within the 2D space.

The two-phase design enables overcoming the challenges posed by ambiguity and scale discussed in the previous section (Section 4.3.1). It splits the large sample into manageable theme classes, which can be accommodated by most computer screens without negatively affecting legibility. Furthermore, the two-phase solution handles the issue of ambiguity by providing a functionality which enables annotators to copy verb labels to capture different word senses in Phase 1. The rough clustering phase ensures that each verb is presented in the context of related verbs in the arena in Phase 2, a necessary prerequisite for meaningful similarity judgments in psychology (Turner et al., 1987).[5] The sense of any given word is implied by the surrounding related words, which helps prevent discrepancies in similarity judgments between participants for ambiguous verbs. Moreover, it avoids the common issue of ambiguous low similarity scores (Milajevs and Griffiths, 2016) that conflate similarity ratings of antonyms (*agree – disagree*) and completely unrelated notions (*agree – broil*), and elicits judgments between comparable concepts.

4.3.3 Data

In order to evaluate the scaling-up potential of the proposed method and to enable direct comparisons with the standard pairwise similarity rating methods, I chose the 827 verbs from SimVerb (Gerz et al., 2016) as the starting sample (with two verbs, *tote* and *pup*, removed due to their very low frequency as verbs, producing an 825-verb final sample). The sample poses a considerable challenge due to its size, being seven times as numerous as the largest stimuli sets so far used in SpAM research, and spans a wide range of verb meaning, with each top-level VerbNet class represented by three or more member verbs.

[5] According to Turner et al. (1987, p. 46), "stimuli can only be compared in so far as they have already been categorised as identical, alike, or equivalent at some higher level of abstraction."

4.3.4 Participants

The target number of participants for both phases of the study was set to 10, with 10 participants clustering verbs in Phase 1 and the minimum of 10 participants working on the arrangement of each class in Phase 2. This decision was motivated by past practices in NLP dataset creation projects as well as the financial constraints on the study. Encouragingly, the analysis of Snow et al. (2008) showed that collecting annotations from 10 (or fewer) non-expert participants yields results highly correlated with expert judgments for a range of natural language annotation tasks, including word similarity, recognising textual entailment, and word sense disambiguation. This number of participants was subsequently used in studies of Huang et al. (2012) and Luong et al. (2013). Gerz et al. (2016) and recently Vulić et al. (2020a) collected the minimum of 10 ratings per word pair, whereas Pilehvar et al. (2018) relied on 8 annotators in their rare word similarity task.

4.3.5 Interface and Task Structure

The two tasks constituting the proposed annotation design were set up on an online platform which allows users to save progress and resume annotation work after breaks as required.[6] Phase 1 and Phase 2 were set up consecutively as separate studies and participants were recruited for each individually. The guidelines for both phases were embedded in the experiment structure, available both prior to and during the task (Appendix C.1). The annotators' understanding of the instructions for each phase was tested in a short qualification task simulating the full experiment, which consisted in clustering (Phase 1) and spatially arranging (Phase 2) seven verbs. The average time spent on the qualification task was 1.5 minutes for Phase 1 and 7 minutes for Phase 2.

4.4 Phase 1: Rough Clustering

The goal of Phase 1 was to classify 825 English verbs into groups based on their meaning, so as to form broad (thematic) semantic classes. The guidelines instructed the annotators to group similar and related words together. While the exact number and size of the classes were left unspecified, the annotators were asked to aim for broad categories of roughly 30-50 words. Deviations from this guideline were allowed in case some smaller or larger semantically coherent groupings of verbs were identified. The reference range of cluster sizes was established through trial experiments with the same

[6] www.meadows-research.com

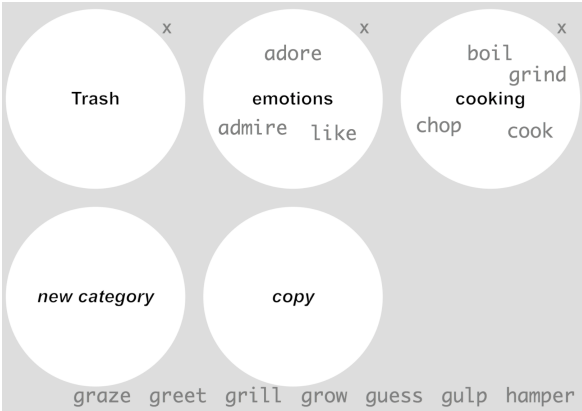


Fig. 4.2 The rough clustering task layout (zoomed in). Verbs can be dragged onto the "new category" circle to create a new grouping, onto "copy" to create a duplicate label, or "Trash" to dispose of the unwanted duplicate.

sample of verbs, taking into account the aforementioned restrictions on the scale of the multi-arrangement task (Phase 2) due to cognitive and technical constraints (i.e., the size of the computer screen). It was meant to guarantee that the size of Phase 2 samples would not significantly exceed 100 verbs, while maximising the coverage of each broad semantic domain in terms of pairwise dissimilarity scores in the final dataset (i.e., judgments in Phase 2 can only be collected on the verbs which appear in the same arenas, having been assigned to the same clusters in Phase 1). Preserving broad clusters in Phase 1 also ensures that there is sufficient context for each arrangement in Phase 2. If much smaller clusters were created in Phase 1, the annotators would have fewer context words against which to calibrate their judgments in Phase 2, yielding potentially more arbitrary scores.

The Phase 1 task interface presents the participants with a scrollable alphabetical queue of 825 verbs at the bottom of the screen and three white circles, "new category," "copy" and "trash" (Figure 4.2). They are instructed to drag and drop the verbs from the list one by one into the circles, creating new ones as they work through the sample. Each circle represents a semantic cluster created by the participant and serves as a container for a single grouping of similar and related verbs. If a single verb fits in more than one group, the guidelines instructed to copy the verb label (as many times as required, by dropping it onto the "copy" circle) and put each in a different circle of related verbs. This was illustrated in the annotation guidelines with the verb *draw*, which could be clustered with art-related verbs (e.g., *paint, design*) or verbs such as *pull* and *drag*. The copying functionality allowed handling of both polysemous and vague verbs.

4.4.1 Participants

Two native English speakers first participated in a test round of the rough clustering task. The clusters they produced showed an encouraging degree of overlap, calculated based on the B-Cubed metric (Bagga and Baldwin, 1998) extended by Amigó et al. (2009) to overlapping clusters and by Jurgens and Klapaftis (2013) to fuzzy clusters, as used in related work (Jurgens and Klapaftis, 2013) and Chapter 3. The B-Cubed metric, based on precision and recall, estimates the overlap between two clusterings X and Y at the item level. Let U represent the collection of items, X_i the set of clusters containing item i in clustering X, Y_i the set of clusters containing i in clustering Y. Let $j \in X_i$ and $j \in Y_i$ be an item, including i, from the set of clusters containing i in clustering X and Y, respectively. B-Cubed precision P and recall R are defined as:

$$P = \frac{1}{|U|} \sum_{i \in U} \frac{1}{|j \in X_i|} \sum_{j \in X_i} \frac{min(|X_i \cap X_j|, |Y_i \cap Y_j|)}{|X_i \cap X_j|} \tag{4.1}$$

$$R = \frac{1}{|U|} \sum_{i \in U} \frac{1}{|j \in Y_i|} \sum_{j \in Y_i} \frac{min(|X_i \cap X_j|, |Y_i \cap Y_j|)}{|Y_i \cap Y_j|} \tag{4.2}$$

Precision and Recall are combined into F-measure as follows, defined as their harmonic mean where $\alpha = 0.5$:

$$F_\alpha(P, R) = \frac{1}{\alpha(\frac{1}{P}) + (1-\alpha)(\frac{1}{R})} \tag{4.3}$$

The obtained B-Cubed Inter-Annotator Agreement (further IAA) score (0.400) compares favourably to the results of the preliminary study in Chapter 3 (B-Cubed IAA scores ranged between 0.172-0.338). It is also promising compared to the results obtained in SemEval (Jurgens and Klapaftis, 2013), where scores ranged between 0.201-0.483, given that cluster labels in that task were selected from a small number of fixed classes per item based on WordNet (Miller, 1995).

Subsequently, a group of 10 English native speakers from the UK and the US, with a minimum undergraduate level of education, who successfully completed the qualification exercise, participated in the task, spending 2.4 hours on average to complete it. The number of the produced clusters ranged between 10-67 (27.5 on average) (Figure 4.3), each with an average of 12.3-82.5 verb members.[7]

[7]Distributions of cluster sizes per annotator are presented in Appendix C.2.

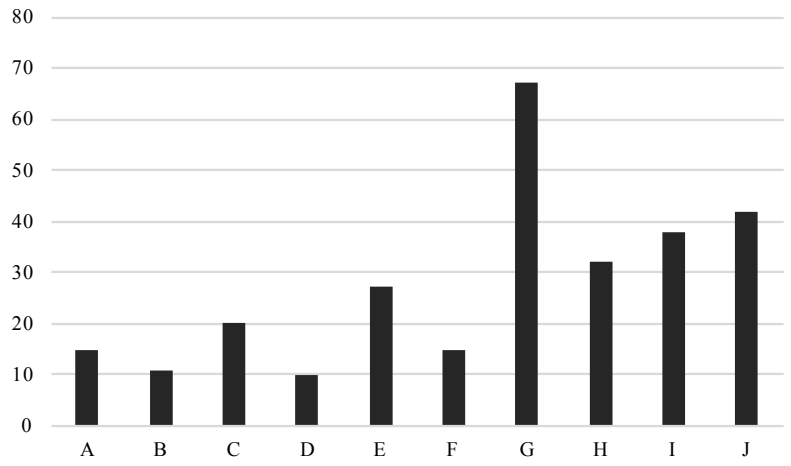


Fig. 4.3 Distribution of the number of clusters in the output of Phase 1 across the 10 annotators (A-J).

4.4.2 Qualitative Analysis

Manual review of the Phase 1 data and the cluster labels provided by the annotators[8] reveal strong similarities in the semantic themes identified by each participant (analysed in depth in section 4.4.3), e.g., "crime", "movement", "negative" or "positive attitude", "cooking", "sound". However, the distribution of cluster sizes across annotators (Figure 4.3) reveals that there is substantial variation in the number of clusters created from the starting sample, as well as the exact cluster membership. To quantify the overlap between individual clusterings, I compute standard cluster evaluation measures Purity and Inverse Purity between the clusterings from all pairings of annotators. Let X and Y be two sets of clusters and N the number of clustered items. Following Amigó et al. (2009), Purity is computed by taking the weighted average of maximal precision values:

$$\text{Purity} = \sum_i \frac{|X_i|}{N} max_j \text{ Precision}(X_i, Y_j) \tag{4.4}$$

where the precision of a cluster X_i with respect to a cluster Y_j is defined as:

$$\text{Precision}(X_i, Y_j) = \frac{|X_i \cap Y_j|}{|X_i|} \tag{4.5}$$

Inverse Purity, focusing on recall, is defined as:

[8]The choice between numerical or descriptive labels for Phase 1 clusters was left to the annotators; 4 out of 10 participants provided descriptive labels.

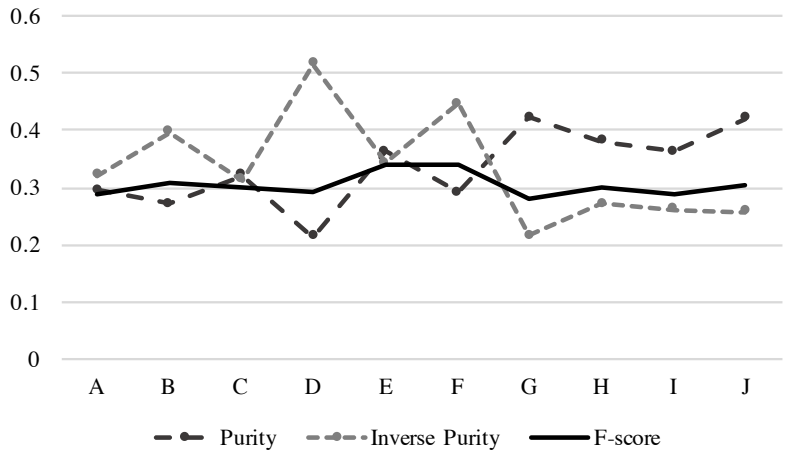


Fig. 4.4 Average pairwise Purity, Inverse Purity and F-score per annotator (A-J), computed between the clustering produced by that annotator and each of the clusterings from other annotators, and averaged across all such pairwise comparisons.

$$\text{Inverse Purity} = \sum_i \frac{|Y_i|}{N} max_j \text{ Precision}(Y_i, X_j) \tag{4.6}$$

Figure 4.4 shows average pairwise Purity, Inverse Purity and their harmonic mean (F-score) for each annotator with respect to the rest. Clustering solutions characterised by fewer, broader clusters score the highest in terms of Inverse Purity (B, D, F), which rewards grouping items together; Conversely, solutions with numerous, fine-grained clusters (G, J) score best in terms of Purity, which penalises noise. Since there is no gold standard classification available against which to evaluate each individual clustering, in this section I qualitatively analyse the Phase 1 data to help uncover the patterns of variation and shed more light on how the participants interpreted the task, as well as the guiding notions of similarity and relatedness of meaning.

Differences in Cluster Granularity

Figure C.1 (Appendix C.2) shows the sizes of individual clusters created by each annotator. Even though the task guidelines provided a reference range for cluster sizes (30-50 verbs), it is apparent that some subjects favoured a more coarse-grained clustering solution (annotators A, B, C, D, and F), while others chose to separate the sample into narrow clusters of similar verbs, which is especially clear in the clustering G. The largest clusters (100+ verbs) produced by annotators B, D, E and F group together verbs related to such meaning domains as movement, violence, crime and

justice, and human communication. In turn, more fine-grained clusterings split these into several subclusters of more closely similar and related verbs. For instance, the partition with the highest number of clusters (G), includes separate groupings of verbs related to physical violence (*beat, hit, kill, kick, punch*) and crime and the justice system (*accuse, acquit, prosecute, sue, abduct, steal*), as well as a separate cluster of verbs describing modes of motion, typically undertaken by an Agent (*glide, jog, jump, pounce, roam*) and verbs describing setting in motion an object (a Theme) (*carry, chuck, dip, haul, toss*).

The differences in cluster granularities between different annotators are a reflection of the inherently hierarchical nature of semantic classes: Each broad meaning domain can be subdivided into a number of smaller sets, and those can be split further to reflect the different degrees of overlap in the meaning of their members. The task guidelines were designed to help annotators arrive at a similar cluster granularity (by providing a reference cluster size range), however, annotators were free to deviate from the general guidance if they thought that a particular meaning domain should be larger or smaller. Depending on the goals of future studies, maximum cluster size could be imposed as a hard criterion. Further, the task interface could be enriched with a 'cluster size tracker' functionality to help annotators monitor cluster sizes as they work on the task. In the present study, the differences in cluster granularity were handled by the class selection protocol (Section 4.4.4), aimed at identifying the overlap between fine- and coarse-grained clusterings based on pairs of verbs which ended up clustered together by the majority of annotators.

Unclassified Outlier Verbs

Manual analysis of the individual clusterings and annotator feedback revealed that some very small clusters (1-3 verbs) were formed to contain the verbs which did not seem to fit in any other cluster. These hard-to-classify cases include verbs *bumble, duck, poise, lounge, engage*. The analysis of clustering results in Chapter 3 indicated that verbs which elude classification are usually those with vague and/or abstract meaning, often with several related senses used in slightly different contexts (e.g., *engage*, which falls in the 'problematic' category also in the present study). This observation similarly applies to the cases listed above. Verbs such as *bumble* or *poise* are less frequent (e.g., *bumble* and *poise* appear only 920 and 844 times, respectively, in the English Wikipedia corpus (as of March 2019[9]), compared to *run* (475,211 times), *take* (527,788) or *include* (809,414)) and polysemous (e.g., *bumble* can refer to the way of acting, walking or

[9]https://dumps.wikimedia.org/enwiki/

speaking, as well as a humming or buzzing sound), which makes them more challenging to quickly assign, and thus more likely to be left unclassified.

Relatedness vs. Association

While the task guidelines encouraged the subjects to place both similar and related verbs in the same clusters, the participants varied in how they interpreted the latter concept and to what extent they let association play a role. In general, the coarser categorisations with larger clusters were more prone to groupings guided by loose association. Examples of such association chains include *whistle, pat, lick, sniff*, evocative of an interaction with a dog, or *visit, cook, knit, pray*, possibly linked through stereotypical associations with the elderly. These purely association-based groupings are specific to individual annotators, which means that they are filtered out from the unified output of Phase 1, based on the agreement from the majority of participants (Section 4.4.4).

Lexical Ambiguity

An important source of variation in the clustering decisions is polysemy. As discussed in Chapter 3, since task participants work on word forms, rather than word senses, they inevitably vary with regard to the number of individual senses considered for any particular word. Six out of 10 annotators used label copying to capture ambiguity and 234 different verbs (out of 825) were assigned to more than one class. Amongst the most highly polysemous verbs (i.e., copied more than once within a single clustering) were *charge* (which the annotators grouped with verbs related to finance/possession, crime and justice, and movement), *angle*, *beat*, followed by *bear*, *lead*, *poach*. In contrast, the four annotators who did not use the copying functionality assigned these ambiguous cases to single clusters, based on the chosen sense (e.g., the sense of *charge* related to the financial domain was selected by three of those annotators). The average pairwise percent agreement on ambiguity decisions (i.e., a binary choice whether a verb is ambiguous or not) was 91.1%. While this thesis focuses on the collection and analysis of human judgments on basic word forms, future work could explore using sentential contexts to help disambiguate word labels (see Chapter 7 for further discussion).

Errors and Bias from Fixed Order

Lastly, some discrepancies in cluster membership across annotators stem from assignment errors. Although the interface permitted modifying the composition of each

cluster throughout the task, the large scale of the endeavour and annotator fatigue made careful cluster verification a challenging task. However, errors can be easily identified post hoc through cross-subject comparisons: Usually, these erroneous cases are specific to single annotators. A special case of errors are misassignments due to the bias from the order of presentation. Given that the verb sample was presented to annotators in alphabetical order to help them locate specific verbs on the scrollable list, annotators could potentially be more liable to clustering orthographically similar (and therefore adjacent) verbs together, regardless of their dissimilar meanings. Two such cases identifiable in the data are the pairs *boom* and *boost*, and *adore* and *adorn*, each placed in the same cluster by two different annotators, which suggests that orthography and/or adjacency, rather than semantics, likely played a role. Further, a consequence of a fixed order of presentation is that annotators may be influenced to make similar choices based on the fact that the words are considered in the same sequence. The annotators were free to edit the clusters and change their choices throughout the clustering process as they encountered new verbs, which helped mitigate the effect of the order of presentation. However, future analyses comparing the clusterings from Phase 1 with the output of a fully randomised experiment may shed more light on the impact of the study design on the resultant clusters. In Chapter 7, I discuss what modifications to the task interface could help make a randomly ordered word list easily searchable for task participants.

In light of the ambiguity inherent to semantic tasks and the flexibility of the experiment setup, where no fixed cluster number or cluster size were imposed on the annotators, inter-subject variability in the clustering output is inevitable. To what extent its sources lie in individual differences between the subjects, the properties of the verb sample and the study design, or the differences in the saliency of the semantic features used as classification criteria, merits further inquiry in a larger-scale psycholinguistic study, which is outside the scope of this thesis. In this work, the focus is on the intersection of individual clusterings, i.e., the shared themes and class members on which the annotators agree. Identifying the areas of overlap in clustering decisions allows retrieving the main meaning domains represented in the verb sample and the core members of each semantic class, which undergo further refinement in the semantic multi-arrangement task in Phase 2. In the following section, I examine the verb groupings emerging from majority decisions and discuss the most salient semantic themes recognised by the participants, before producing the final average classification as input for Phase 2 (Section 4.4.4).

4.4.3 Cluster Analysis

Before unifying the clusterings from individual annotators for the second phase (see Section 4.4.4), I applied network analysis to further scrutinise the rough clustering data. The ultimate goal is to obtain an average classification where membership and size of each class is determined by the intersection of the classes from all annotators (the core), extended by additional valid member verbs on which there was partial agreement. From the entire set of all clusters (G) produced by the 10 annotators, I extracted all pairs of verbs put together in a cluster g by an annotator, that is $v_1v_2 \in P_g$ where P_g is the set of all verb pairings in cluster g. Let P_G be the multiset of all such pairs from $\{v_1v_2 \in P_g : g \in G\}$. Each verb in the pair represents a node in the network, linked by an edge weighted according to the number of annotators clustering them together, that is, the number of occurrences in the multiset P_G, denoted as $N(P_G, v_1v_2)$. Thus, the edge weight is calculated as $w(v_1v_2) = N(P_G, v_1v_2)$.

I applied a weight-based threshold to eliminate weak ties (where $w(v_1, v_2) < 6$, that is, there is no majority from the 10 annotators[10]) and reduce computational burden for network processing, given that the full graph had approximately 285,000 edges. I then used Cytoscape open source software (Li et al., 2017; Shannon et al., 2003) for analysis and visualisation (see Figure 4.5).

To identify higher density areas, corresponding to groupings of similar verbs, I performed cluster analysis with a selection of graph clustering algorithms designed for detecting overlapping and non-overlapping clusters (Li et al., 2008; Nepusz et al., 2012; Wang et al., 2011a, 2012b). Table 4.1 presents the results of this analysis. The labels are given for descriptive purposes alone. All four approaches identified the same largest area of high density of links, formed by the "movement" verbs (e.g., *move, fly, swim, walk*). Other large areas of interconnected nodes include, for example, "communication" verbs, verbs related to crime and violence, "negative emotions" and "cognitive" verbs (Table 4.1).[11] I explored the clusters with two network analysis metrics as follows: closeness centrality (Newman, 2005)

$$C_c(n) = \frac{1}{avg(L(n,m))} \tag{4.7}$$

[10] I experimented with different thresholds and settled for 6 as the value representing the actual majority of annotators and a good compromise between comprehensiveness and computational efficiency. At this point, I wanted to include as much variation as possible in the graph (to also see edges weaker than those representing perfect agreement), while discarding the pairings on which annotator consensus was below the minimum majority threshold.

[11] I manually analysed the clusters output by the four algorithms to identify the main areas of agreement without imposing strict overlapping membership criteria.

Cluster label	**Example verbs**	N
movement	*wander, swing, fly, glide, roam*	4
communication	*persuade, command, tell, ask, say*	4
crime & violence	*beat, abduct, abuse, shoot, kill*	4
negative emotions	*offend, aggravate, enrage, disgust*	4
positive emotions	*admire, respect, adore, like, approve*	4
cognitive processes	*suppose, assume, realize, know*	4
cooking	*cook, slice, grind, stew, boil*	4
possession	*belong, accumulate, obtain, acquire*	4
physiological processes	*perspire, sweat, vomit, inhale*	4
perception	*glance, observe, stare, look*	4
destruction	*perish, demolish, decompose*	4
accomplishment	*accomplish, succeed, excel*	4
construction	*repair, fasten, mend, fit, fix*	2
sound	*hoot, roar, crackle, rattle, hum*	2
rate of change	*boost, raise, accelerate*	3

Table 4.1 Main clusters identified by N graph clustering algorithms in the network created from the 825-verb manual clustering data and example member verbs. Cluster labels are given for descriptive purposes.

and betweenness centrality (Brandes, 2001):

$$C_b(n) = \sum_{s \neq n \neq t} \frac{\delta_{st}(n)}{\delta_{st}} \tag{4.8}$$

$L(n, m)$ is the length of the shortest path between two nodes n and m, and $C_c(n)$ of n is the reciprocal of the average shortest path length. s and t are nodes different from n, δ_{st} is the number of shortest paths from s to t, and $\delta_{st}(n)$ is the number of shortest paths from s to t on which lies n (i.e., the number of paths equal to the shortest length overall).

The closeness centrality measure identifies the nodes with the shortest total distance to all other nodes, that is, the prototypical member of a class. The verbs with highest $C_c(n)$ values can be seen as representing the underlying common "theme" of the cluster. For example, verbs with the highest $C_c(n)$ score are *speak* for communication verbs, *annoy* for the "negative emotions," and *destroy* for "destruction" verbs. Betweenness centrality, on the other hand, quantifies the number of times a node acts as a bridge along the shortest path between two other nodes. Thus, it can be used to identify the verbs which act as connectors between different clusters, such as ambiguous verbs whose different senses belong to different groups. One example is *cry*, which connects the "sound" cluster comprising verbs such as *scream, holler, squeal*, and the "physiological processes" verbs, like *breathe, cough, sneeze* (Figure 4.5). Identifying prototypical

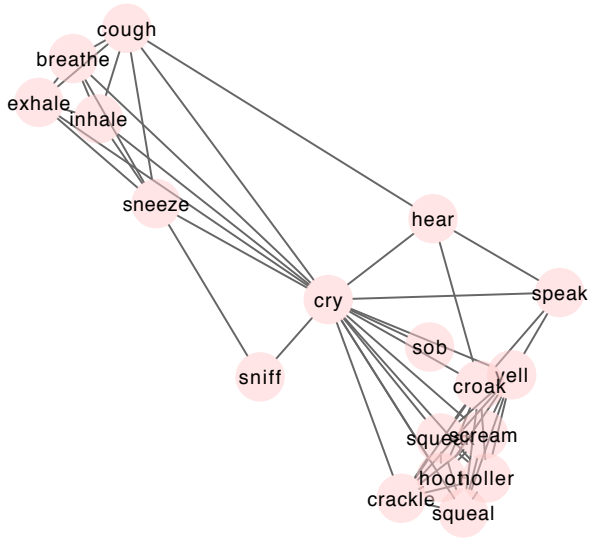


Fig. 4.5 Visualisation of a fragment of the network with the verb *cry* acting as a connector node.

members and verbs acting as inter-class links can be especially useful for creation of a comprehensive verb resource.

In the next section (4.4.4), I outline the protocol for selection of classes for Phase 2. While cluster analysis was used primarily as an exploratory means, allowing for an examination of annotator decision patterns on the rough clustering task and the emerging semantic categories, it also served as a preliminary step that informed the decisions relative to Phase 2 class selection. I kept the minimum majority threshold of 6 annotators, as high enough to ensure semantically coherent classes and discard noise, but also comprehensive enough to leave room for some degree of variation in clustering decisions, reflecting their inherent flexibility (i.e., there is no single perfect clustering solution). The chosen threshold also guaranteed the desired granularity and nature of resultant classes. While higher thresholds produced many narrow clusters of synonyms or close-synonyms (e.g., *join, connect, associate*, or *forbid, deny, disallow, refuse*) lowering the cut-off value yielded broader semantic classes including the less prototypical members (on which there was partial agreement), which was the intended output of Phase 1.

4.4.4 Class Selection for Phase 2

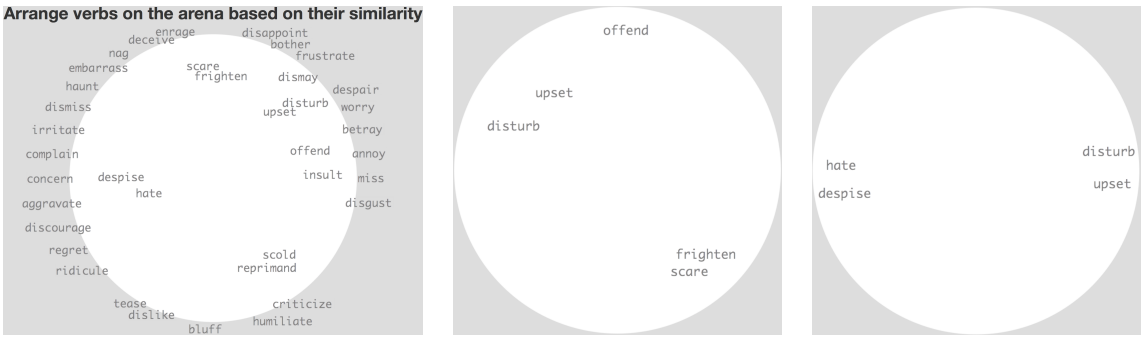
The class selection protocol which determined the classes to be used in Phase 2 was the following. Clusters obtained from the verb pairings on which any 6+ participants (the majority) agreed were used as a starting point and determined the broad semantics

of the classes (e.g., "perception," "movement," "communication"). Post-processing was limited to (1) merging smaller semantically related clusters to produce large, all-encompassing classes based on semantic relatedness of class members, and (2) populating the thus created sets with the verbs missing from the majority classes based on their relatedness to the already classified members (i.e., these lower-agreement verbs were reviewed and manually added to related classes). Clusterability of Phase 1 verbs (i.e., the SimVerb sample) was guaranteed by balanced sampling from across different VerbNet classes (Gerz et al., 2016). Ambiguous verbs could be placed in several classes with semantically related members by means of the copying functionality described above (Section 4.4). The final number of produced classes was 17.

The main clusters identified by the clustering algorithms in Section 4.4.3 overlap very closely with the final classes used for Phase 2. All the semantic areas (see descriptive labels in Table 4.1) recognised through network clustering are represented in Phase 2. The only discrepancies lie in the granularities, for instance, while the clustering algorithms unify all verbs related to motion, the class selection protocol produced two different classes split along a line mirroring the intransitive/transitive distinction, that is, movement verbs where the intransitive sense is predominant (*crawl, dash, fly*), and transitive verbs describing causing something to move (*drag, tow, fling*). Similarly, the broad cluster related to "crime and violence" is split into two Phase 2 classes: verbs of physical contact (*beat, kill*) and verbs describing criminal acts and legal terms (*kidnap, abuse*). Amongst the smaller areas of higher density identified by cluster analysis that are not represented as separate Phase 2 classes were narrow semantic groupings, usually of synonyms or close-synonyms, such as (*imitate, mimic, impersonate*), (*crave, yearn, want*), or (*help, assist, aid, rescue, protect*), as well as few examples of small clusters based on association (e.g., *embarrass, worry, weep, regret, sprain*, or *stop, withdraw, unload*).

4.5 Phase 2: Multi-Arrangement

In Phase 2, participants performed the spatial multi-arrangement task. Each of the 17 verb classes output by Phase 1 was individually displayed on the computer screen, in random order, around a *circular arena* (Figure 4.6). The participants were instructed to arrange verbs based on similarity of their meaning, dragging and dropping the labels one by one onto the circle, so that similar words ended up closer together and less similar ones further apart, while the relative positions and distances between them reflected the degree of similarity.



(a) trial 1 (b) trial 2 (c) trial 3

Fig. 4.6 Consecutive Phase 2 trials on a single class (zoomed in). In the first trial (a), the whole class is presented around the arena and words are dragged and dropped one by one, with their relative distances representing the degree of similarity. Words put closer together in the first trial are subsampled in the subsequent trials (b)-(c), and arranged again in a less crowded space, which ensures a higher signal-to-noise ratio (i.e., since annotators use the whole space available in each trial, the items are more spread out and placement error is a smaller proportion of the dissimilarity signal). The RDM estimate is updated after each trial and the evidence from consecutive 2D arrangements is combined to produce the final pairwise dissimilarities for the entire word set.

4.5.1 Participants

The minimum number of annotators to work on each class was set to 10. Each annotator was asked to arrange at least 3 classes, presented in random order, and permitted rest breaks between classes. Annotator recruitment continued until the minimum number of annotators per class was satisfied. Overall, 40 native English speakers from the UK and the US, with a minimum undergraduate level of education, took part in the multi-arrangement task, producing ultimately a total of 314,137 individual pairwise scores. For each class and annotator, the time spent on each individual trial was recorded (i.e., each consecutive arrangement of subsets of a single class). The average total time spent completing the task for all 17 classes was 735 minutes, with the average time spent on a single task (equivalent to arranging one class) ranging from 15.5 minutes (for the smallest class) to 60 minutes (for the largest class).

4.5.2 Post-Processing

I employed the following steps to ensure high quality of the resultant data. First, I discarded annotations where word placements were executed too fast in the first arrangement of each class, i.e., where the average time spent on dragging and dropping

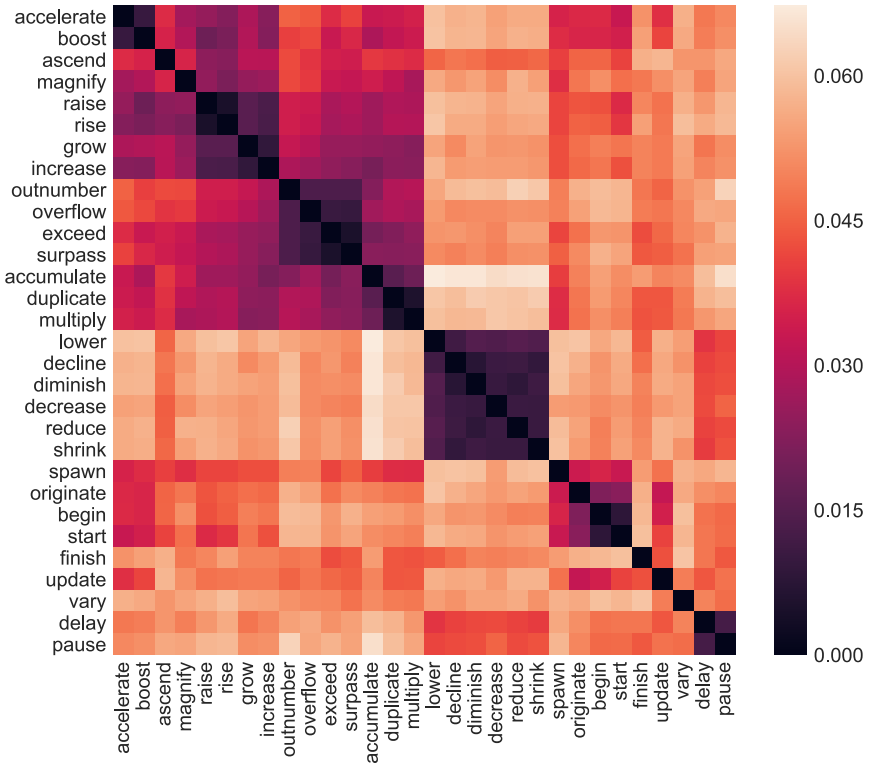


Fig. 4.7 Average ordered dissimilarity matrix for one of the verb classes (dark-to-light colour scale for small-to-large dissimilarities), with dark areas corresponding to clusters of similar verbs (e.g., *lower, decline, diminish, decrease, reduce, shrink*).

a single verb label was less than 1 second (<2% of responses). This heuristic allowed for quickly identifying and eliminating rogue annotators. Trial experiments showed that users spend much longer on the first arrangement than the consecutive ones for that class, given that it is the first time they see a given word sample and extra time is needed to familiarise oneself with the set. Extremely short times spent on word placements in the first trial were therefore a clear indicator of low-effort responses. Second, for each class I excluded outlier annotators for whom the average pairwise Spearman correlation of arena distances with distances from all other annotators was more than one standard deviation below the mean of all such averages. The same criterion was adopted as the acceptability threshold in the creation of SimLex (Hill et al., 2015). This excluded 18% of submissions, leaving 8-13 accepted annotators per class.

For each class, I computed the average of the Euclidean distances from all accepted annotators for each verb pair and obtained an average RDM (as shown in Figure 4.7). The averaged pairwise distances (= dissimilarity scores) in each class were then scaled

to have a root mean square (RMS) equal to 1, as done in previous work using inverse MDS (Kriegeskorte and Mur, 2012; Mur et al., 2013), to ensure inter-class consistency. For each class, the scaled distances $d'_i, ..., d'_N$ were thus obtained for N pairs by dividing each pairwise distance d_i by the square root of the mean of N distances squared (d_i^2):

$$d'_i = \frac{d_i}{\sqrt{\frac{1}{N}\sum_{i=1}^{N} d_i^2}} \tag{4.9}$$

The final dataset, SpA-Verb, collates the thus obtained scaled averaged pairwise distances for each of the 17 verb classes, comprising (dis)similarity scores for the total of 29,721 unique verb pairs.

4.6 Inter-Annotator Agreement

I measure inter-annotator agreement in Phase 2 based on Spearman's rank correlation coefficient (ρ): For each class, I calculate the average correlation of an individual annotator with the average of all other annotators (*mean* Spearman's ρ) (Gerz et al., 2016; Hill et al., 2015) (see Table 4.2).[12] I do not calculate IAA over the entire dataset as different groups of annotators worked on different classes.

The characteristic flexibility offered by the drag-and-drop interface, where similarity judgments expressed through word placements produce fine-grained pairwise similarity scores differing by fractions, based on the words' relative positions in the circular space, leaves a lot of room for divergence in scores across annotators compared to discrete ordinal rating scales. Nonetheless, the resultant IAA scores (ρ^A) are promising. In particular, they compare favourably with inter-subject correlations reported in cognitive neuroscience research for spatial multiple arrangements of concrete visual stimuli (real-world objects like in Figure 4.1): e.g., Mur et al. (2013) report an average total pairwise inter-subject Spearman's ρ correlation of 0.32, Cichy et al. (2019) report scores in the range of approximately $0.12 - 0.21$ ($p < 0.001$).

[12]Rank correlation metrics like Spearman's ρ, which measure the correlation between rankings of (dis)similarity scores, rather than the absolute scores, are recommended for comparing RDMs (Nili et al., 2014). Given the free nature of the arrangement task, some degree of inter-subject variability in the usage of the arrangement space and the raw inter-item distances in each trial is expected, regardless of the degree of consensus on the relative similarities of word pairs in the arena, which is of interest in this study; Therefore, comparing rank orders of the dissimilarities, rather than the variance of their raw values, provides an informative measure of agreement on similarity judgments. Note also that there is no fixed relationship between screen distance and dissimilarity that holds across trials: Since participants "zoom in" on items previously clustered together by spreading them out upon successive trials (Figure 4.6), it is the relative screen distances (i.e., screen distance ratios) that reflect the relative dissimilarities on each trial. See Section 4.6 for more discussion.

Effect of Class Size and "Clusterability"

The main factor affecting the difficulty of the task was class size, as reflected in the differences in agreement scores reported in Table 4.2. We can observe negative correlation between inter-annotator agreement and the number of verbs in a class (Spearman's $\rho = -0.67$). However, the semantics of the classes seem to play a role as well. For instance, the agreement on the largest 100-verb class of movement verbs (*#10*) is higher than could be expected based on its size alone ($\rho = 0.26$), compared to the smaller Class 15, where the agreement is the lowest. We can observe that class "clusterability" is an important factor, i.e., the availability of underlying structure within a bigger class, where words cluster into balanced sub-groups with clearly defined shared semantics.[13] For instance, many movement verbs have well-defined, concrete meanings, clusterable into smaller groupings, for instance, based on the medium in which the movement takes place (on land (*walk, crawl*), in water (*swim, dive*), in the air (*glide, fly*)). The lowest-agreement Class 15, comprising verbs of motion undergone by the verb's *object*, such as *add, dip, flush, spread*, is more heterogeneous, i.e., there is more variety in verbs' semantic properties and the dimensions along which the class members differ are less clearly defined, which means there are many equally valid arrangements possible. As indicated in annotator feedback, this characteristic made it harder to identify the potential groupings and subcategories into which words could be classified; Consequently, their relative positions varied by participant.

In order to examine the impact of sample size on IAA, I carried out a follow-up experiment on the lowest-IAA class (*#15*). The goal was to verify if higher IAA scores can be obtained on the same verb pairs split into smaller samples. Five new annotators subsequently arranged three equal 29-word subsets randomly sampled from the entire 87-word class (*#15*), each working on the three subsets one by one, with breaks in between. The IAA computed for the smaller sets proved lower than in the full-class (87-word) setting, producing an average across the three subsets of $\rho = 0.098$, compared to $\rho = 0.19$ on the full class. This analysis suggests that while big samples are generally more challenging, the task's difficulty very much lies in the verbs included in the sample, and this class proves particularly difficult due to its heterogeneity. While annotators consistently place similar verbs close together (e.g., *smear* – *smudge*, *seize* – *snatch*), there is greater variability in the distances between the less similar words.

[13] In the preliminary trial experiments, annotators reported that the availability of words which naturally group together within a bigger class based on some criterion (e.g., animal sounds, human sounds) significantly facilitated the spatial arrangement task, in contrast to having small but randomly sampled sets of words to arrange, with many semantic "isolates," i.e., words which were dissimilar from all others.

In the follow-up study, this issue was further aggravated by randomly splitting the coherent big set and potentially separating verbs naturally clusterable together. These findings also indicate that simply reducing the number of words to be arranged in the arena does not guarantee higher agreement: Being presented with a semantically clusterable bigger set of words (like those produced in Phase 1) may be preferable to imposing an arbitrary limit on class size. The greater difficulty of some verb sets resulted in inter-annotator agreement scores for some classes showing low positive correlation. Therefore, evaluation of representation models should best be focused on classes with higher inter-annotator agreement and consequently clearer semantics.

Inter-Subject Variability in SpAM

While class size and heterogeneity account for most of the variation found across annotators, taking a closer look at the data collected for individual annotators and comparing them may shed light on other sources of variation, and any differences in how the participants tackled the task. It is important to remember that the data collected in Phase 2 are in the form of average representational dissimilarity matrices (i.e., matrices of pairwise distances between all verbs in a class over multiple arrangements, statistically combined), rather than visual snapshots of actual consecutive item arrangements. Therefore, in order to visually inspect the data from different annotators, dimensionality reduction and visualisation techniques, such as Principal Coordinates Analysis (PCoA) (Gower, 1966), need to be used to represent the data in two dimensions, bearing in mind that their output is an approximation of the relations encoded in the data.

Figures C.2 and C.3 (Appendix C.3) compare the semantic spaces reconstructed from the Phase 2 data from pairs of annotators with the highest (top) and lowest (bottom) average pairwise agreement with the rest of annotators, to highlight the ways in which the most divergent solutions differ from the mean.

Figure C.2 visualises the distances between the verbs belonging to the 'cooking' domain (Class 11). The main groupings which emerge in the top arrangement include: verbs describing ways of drinking (e.g., *swallow, gulp, drink, slurp, sip*) and eating (e.g., *bite, chew, munch*), actions related to liquids (e.g., *drain, rinse, drip, wash, pour, sprinkle*) and application of heat to food (e.g., *broil, cook, bake, fry*), as well as methods of preparation of food, split between mixing and whipping (e.g., *mix, combine, bend, whip*) and cutting (*slice, chop*). The upper region is occupied by verbs of change of state, with a separation between *set* and *freeze*, and verbs such as *melt, thaw* and

dissolve. Within these groupings, synonymous and near-synonymous words appear close together (*soak* and *drench*, *defrost* and *thaw*, *mix* and *combine*, *toast* and *grill*).

Similar clusters emerge in the bottom figure, even if less clearly delimited: verbs related to liquids in the centre (*drain, strain, drench, wash*), 'heat' verbs towards the bottom (*broil, boil, fry, poach, bake*), verbs describing ways of eating and drinking on the right (*drink, bite, chew, suck*) and a separate cluster of verbs of change of state (*thaw, defrost, melt, dissolve*) on the left. Again, pairs of synonymous verbs, such as *thaw – defrost*, *combine – blend*, appear close together.

While there are differences between the two arrangements with regard to the exact composition of the clusters, the participants seem to have considered similar criteria when judging word similarity, paying attention to such features as temperature, manner, and the physical state of the verbs' arguments. However, a few cases can be found where association seems to have overridden the considerations of shared semantic features in one or both participants. In the top arrangement, the strongly associated but dissimilar pair *starve – feed* appear close together. In the bottom arrangement, the placement of *feed* close to *slurp* may be similarly motivated by association (e.g., feeding a child). Both, however, separate strongly associated but antonymous *freeze* and *thaw*.[14] In some cases, the reason behind association-based choices seems to be the lack of similar verbs with which a given verb could cluster, as in the case of the verb *intoxicate* appearing close to the verbs of drinking in the top arrangement. This case is representative of a pattern observed across different classes: Where a verb is an outlier in terms of semantic similarity, the annotators choose relatedness or association as a secondary criterion; If no such link can be found, the verb is usually a true outlier, consistently placed far away from the rest.

Figure C.3 visualises the verbs of rate of change from Class 3. The underlying RDMs are two of the least correlated with each other in the set ($\rho = 0.25$). However, manual inspection of the two images reveals that, despite the relatively weak correlation, the annotators largely agreed on the relative similarities of the verbs in the sample, as reflected in the created groupings. Both include a separate grouping of verbs describing growth (e.g., *increase, grow*) and a separate cluster of their antonyms (e.g., *decrease, shrink*); the verbs related to time form separate groupings (*pause* and *delay*, *start* and *begin*), whereas *update* and *vary* are kept separate from all the rest. Although there is variability in the relative placements of the groupings themselves within the arena (as a consequence of the individual differences in the usage of the arrangement space, as well

[14]The location of *freeze* in the vicinity of *burn* in the bottom arrangement is a misrepresentation due to dimensionality reduction: their dissimilarity scores reveal that they are distant from the rest of verbs in the arena, as well as from each other.

as accidental differences due to PCoA), there is substantial consensus on which verbs should be clustered together, with pairs of synonyms and near-synonyms consistently placed next to each other by both participants (e.g., *start – begin*, *increase – grow*, *decrease – decline*).

These comparisons provide important context for the interpretation of Spearman's IAA results in terms of 'low' and 'high' agreement. Contrary to rating-based datasets, where participants assign scores to word pairs, in the spatial method, pairwise scores are a product of statistically combining the evidence from consecutive arrangements to obtain the RDM estimate. The final complete RDM for a given word set for a given annotator aggregates dissimilarity signals over multiple arrangements, preserving the meaningful relative differences in pairwise distances. Given the inherent flexibility of item placements, inter-subject variability in how the arrangement space is used from trial to trial is expected (see also Footnote 12). However, although the correlations obtained in the spatial method are on average lower than in the pairwise rating method (Cichy et al., 2019; Mur et al., 2013), the two methods have been shown to produce highly correlated results, with the spatial method surpassing the pairwise rating approach in test-retest reliability (computed as Spearman's correlation): $r = 0.93$ ($p < 0.0001$), compared to $r = 0.92$ (Kriegeskorte and Mur, 2012).

What is more, a comparative study of the two methods by Hout et al. (2013) (who used, however, a single-arrangement implementation of SpAM, in contrast to the more expressive and robust multi-arrangement method used in this work (Kriegeskorte and Mur, 2012)) demonstrated that individual variations in the salience of dimensions used by participants in their arrangements may produce a higher-quality aggregate solution which appreciates the full set of dimensions in high-dimensional space. In their analysis, they found that while SpAM offers more freedom for the expression of similarity judgments than pairwise ratings, and more room for individual differences in terms of the use of the geometric space, the aggregate results are resistant to outliers (identified as the 25% annotators with (a) the lowest average correlation with others or (b) the lowest proportions of reliable correlations). They evaluated the degree to which the irregular participants skew the aggregate results by computing the correlation of (i) the average of scores from the irregular solutions, and (ii) the average of scores from the regular solutions, with the average aggregate scores from all annotators. They found that removal of irregular solutions had little effect and the regular scores (ii) were in high agreement with the aggregate scores ($r = 0.91$). In contrast, the irregular solutions were weakly correlated with the aggregate ($r = 0.12$). Interestingly, a similar analysis of the pairwise rating method showed that while regular annotators still highly

#	Example verbs	N	N^A	ρ^A	ρ^{SV}	N^{SV}
1	*beat, punch, smash, slap*	48	1128	0.53	0.50	92
2	*accuse, condemn, forbid, blame*	80	3160	0.27	0.61	134
3	*accelerate, decrease, shrink, increase*	30	435	0.64	0.71	38
4	*achieve, aim, tackle, accomplish*	57	1596	0.34	0.41	98
5	*acquire, have, keep, borrow*	47	1081	0.40	0.50	102
6	*dismay, frustrate, upset, irritate*	38	703	0.24	0.35	73
7	*ask, confess, discuss, inquire*	85	3570	0.27	0.30	194
8	*approve, desire, prefer, respect*	23	253	0.41	0.33	31
9	*calculate, analyze, predict, guess*	75	2775	0.31	0.51	159
10	*climb, jump, roam, slide*	100	4950	0.26	0.48	253
11	*bake, grate, slice, broil*	53	1378	0.52	0.66	85
12	*cough, gulp, inhale, sniff*	56	1540	0.29	0.69	52
13	*chirp, hoot, roar, whistle*	34	561	0.53	0.65	51
14	*build, fasten, mend, restore*	62	1891	0.24	0.46	89
15	*drag, fling, haul, toss*	87	3741	0.19	0.36	129
16	*demolish, erode, wreck, disintegrate*	27	351	0.46	0.62	51
17	*glance, observe, perceive, look*	41	820	0.43	0.71	76

Table 4.2 IAA (mean Spearman's ρ) by verb class (ρ^A) of N verbs and N^A unique verb pairs and set of N^{SV} verb pairs shared with SimVerb in that class (ρ^{SV}), and examples of verbs in each class.

correlated with the aggregate ($r = 0.85$), the irregular pairwise solutions also showed moderate correlation with the aggregate data ($r = 0.42$). It is worth emphasising that in the multi-arrangement approach used in the present study, high-dimensionality is achieved already at the level of an individual participant, thanks to the multiple successive arrangements, aggregated into the single participant RDM. Finally, Hout et al. (2013) also evaluated the robustness of SpAM to a reduction in the number of annotators (from 80-90 to 10-20 participants; note that each participant arranged the stimuli only once) and its impact on the quality of the aggregate solution. They found single-trial SpAM to consistently correlate highly with the solutions from the pairwise technique, regardless of the number of aggregated solutions.

SpAM vs. Pairwise Ratings

Since the verb sample used in this study is the same as SimVerb's, we can directly compare IAA recorded for each class with the IAA on the verb pairs in that class also occurring in SimVerb. The results of this analysis are shown in ρ^{SV} of Table 4.2. In what follows, I use this comparison to highlight the main similarities and differences between the output of the Phase 2 method and the pairwise rating approach used with

SimVerb, which due to its scale and sole focus on verbs is the most similar resource currently available.

Even though the two resources share the starting verb sample, the number of overlap pairs in each class (as shown in column N^{SV} of Table 4.2) is reduced due to the differences between the annotation paradigms used in SimVerb and SpA-Verb. In SimVerb, pairs which end up in the final dataset were selected to cover different degrees of relatedness, including completely unassociated pairs. Whereas the rough clustering phase (Phase 1) divides the sample into classes based on relatedness, therefore the possible pairwise combinations of verbs are limited to related verbs. These discrepancies are reflected in the different score distributions in both datasets, as illustrated in Figure 4.8. SimVerb scores show a peak at the 0-1 unrelated end of the distribution. The most numerous are the easy to annotate unrelated verb pairs, which are filtered in Phase 1 of the approach presented in this chapter (see Section 4.8.6 for a discussion of the implications of these differences for intrinsic evaluation).

The sets of shared pairs are on average over one order of magnitude smaller than the respective complete classes (N^A in Table 4.2). What is more, the overlap pairs are more spread out in terms of degree of similarity compared to the complete classes, which comprise very many nearly equidistant verb pairs. Crucially, for each cluster of similar verbs in a Phase 2 arena, the SpA-Verb dataset includes all the possible unique pairwise combinations; Consequently, many scores differ by small amounts. Only some of those pairs appear in SimVerb, for example, out of Class 9 pairs *decide – choose*, *decide – select*, *decide – elect*, *decide – pick*, only the first one is present. These highlighted differences explain the lower correlation scores obtained on most of the entire classes compared to overlap pairs (ρ^A vs. ρ^{SV}), which, in turn, reflect the greater difficulty in making subtle distinctions between very many semantically related words appearing in the same arena in the spatial arrangement task.[15] While many concurrent decisions make judgments in the arena harder, the resultant scores are more thorough, offering a more comprehensive coverage of a given semantic domain (i.e., a complete pairwise similarity matrix for each arena).

Even though SimVerb and SpA-Verb are produced by different paradigms, there is a reasonable level of correlation between the two resources on all the 1,682 shared pairs: $\rho = 0.62$. Crucially, by eliciting simultaneous judgments on multiple lexical items the approach presented in this chapter significantly speeds up the data collection process. As an example, with the presented SpAM-based method 60 minutes of work of a

[15]The ρ^{SV} scores computed for the overlap pairs are promising compared to the $\rho = 0.612$ correlation reported for SimVerb (Pilehvar et al., 2018), especially in light of the fact that the easy cases of pairs of very disparate verbs (split into different classes in Phase 1) are not included in the present results.

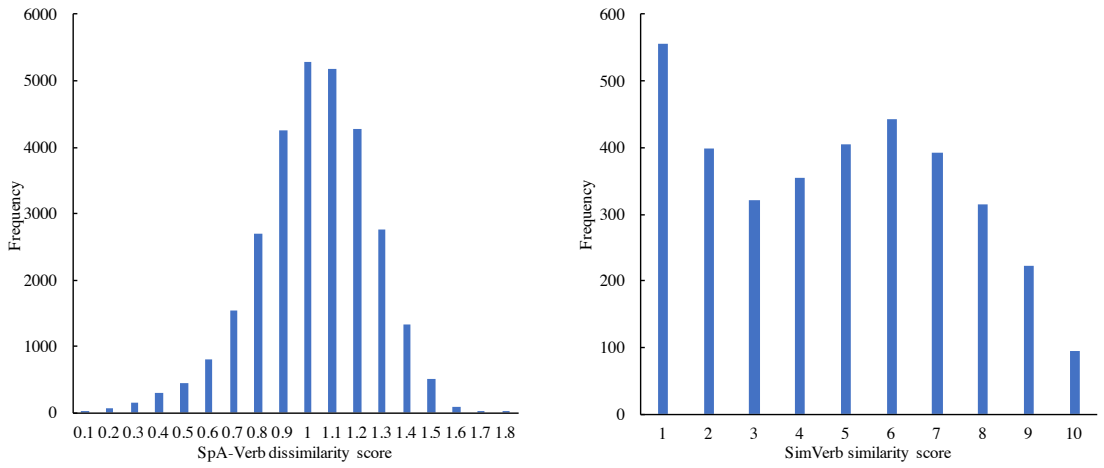


Fig. 4.8 Score distribution for SpA-Verb (dissimilarities are scaled Euclidean distances (Eq. 4.9)) and SimVerb (ratings on a 0-10 interval) in terms of frequency of each score interval (i.e., the number of individual ratings belonging to a given score interval in each dataset). Each score interval label gives the upper bound.

single annotator produces pairwise similarity scores for 4,950 unique verb pairs. In the pairwise rating approach used for SimVerb, the same number of similarity judgments would take a single rater over 8 hours to record (requiring approximately 8 minutes to complete 79 questions by a single participant (Gerz et al., 2016)). The two-phase design introduced in this chapter and the modular nature of its annotation pipeline make it particularly suitable for crowdsourcing. In the following section, I explore in detail the properties of each of the two phases, highlighting their benefits beyond what is offered by pairwise rating datasets.

4.7 Phase 1 and Phase 2 Analysis

The two tasks constituting the annotation design target two types of closeness of meaning: broad topical similarity in Phase 1, represented as semantic clusters, and similarity of meaning in Phase 2, understood in terms of shared semantic properties and represented as varying distances between related words. In order to examine how the collected human judgments reflect these different assumptions, I carry out a comparative analysis of the data with two lexical resources, FrameNet (Baker et al., 1998) and VerbNet (Kipper Schuler, 2005). While the former has a semantic focus, the latter captures the interrelatedness of verb meaning and syntactic behaviour. Measuring the overlap between VerbNet classes and human judgments of semantic

similarity will therefore cast light on one of the research questions of this work: If verbs' semantics is a strong predictor of their structural properties, how similar will a purely semantically driven classification be to one that is also syntactically informed? Following a quantitative analysis, I zoom into selected Phase 1 classes for a qualitative comparison with VerbNet classes to gain insights about their similarities and differences, relying on different methods of cluster analysis. Next, I investigate whether the perceived semantic distances between related words recorded in the spatial arrangement task reflect finer-grained lexical relations like synonymy or hypernymy by comparing average similarity scores collected for verb pairs grouped according to the WordNet relations in which they participate. I conclude the analyses with a closer inspection of the distribution of verbal concepts within a given semantic domain by means of Principal Coordinates Analysis on the spatial dissimilarity data.

Comparison with FrameNet and VerbNet

Based on Frame Semantics theory (Fillmore, 1976, 1977, 1982), FrameNet comprises over 1220 semantic frames, that is, descriptions of types of events, relations, or entities, and their participants. Each frame records the semantic type of a given predicate, usually a verb, and the semantic roles and syntactic realisations of its arguments. The lexical units associated with it share similar semantics and argument structures. On the other hand, VerbNet extends Levin (1993)'s taxonomy and groups verbs into classes based on shared semantic and syntactic properties. Each class is described by thematic roles, selectional restrictions on the arguments, and frames including a syntactic description and a semantic representation.

For each phase of the proposed design, I compute the overlap between the resultant human classes/clusters and classes/frames extracted from both of these resources. For Phase 1, I compute the B-Cubed metric directly between the 17 classes and FrameNet parent frames (i.e., one level up in the hierarchy from fine-grained FrameNet frames) and top-level VerbNet classes, extracted for the 825 verbs in the present sample.[16] For Phase 2, for three selected verb classes (the largest (*#10*), the lowest IAA (*#15*) and highest IAA class (*#3*)) I first extract K_{Gold} FrameNet frames and parent frames, and K_{Gold} top and 1st level VerbNet classes (e.g., 17.1). K_{Gold} is the number of frames or classes in which the verbs in a given Phase 2 sample (*#10*,*#15*,*#3*) participate in

[16] I selected the hierarchy levels in FrameNet and VerbNet for comparison with the Phase 1 classes and Phase 2 clusters aiming to compare similar granularity levels, comparing broader Phase 1 classes with higher levels of each hierarchy. However, there is still a major difference in the number of classes in Phase 1 (17) and FrameNet parent frames (128) and VerbNet top-level classes (101) (for the shared verbs).

these resources (see Table 4.2 for examples of verbs in each). I then apply hierarchical agglomerative clustering with average linkage (Day and Edelsbrunner, 1984) (Appendix A) on top of the distance matrices produced by Phase 2. I calculate B-Cubed for the optimal clustering solution, defined in terms of highest value of F1 score, and for $k = K_{Gold}$, where the number of target clusters is fixed and equal to the number of classes identified by experts for the same set of verbs in each resource. While keeping k free allows for optimising the clustering solution to maximise the overlap score with the gold classes (i.e., the optimal k does not necessarily equal K_{Gold}), the harder $k = K_{Gold}$ setup constrains the algorithm to partition the verbs into the same number of classes as those which accommodate them in each lexicon. This directly addresses the question: If verb sample S were to be divided into K_{Gold} classes automatically based on spatial arrangement data, how closely would they resemble those created by experts in VerbNet or FrameNet? To contextualise the B-Cubed scores and provide a framework for their interpretation, I also compute the overlap between the two reference resources themselves, FrameNet and VerbNet, at two levels: comparing the top-level VerbNet classes with parent frames in FrameNet, and the first-level VerbNet classes and FrameNet frames, on the same sets of verbs (i.e., the whole sample (Phase 1), and the three Phase 2 classes). The results are reported in Table 4.3.

The limited degree of overlap between Phase 1 and the two resources is understandable due to the different granularity: Phase 1 includes only 17 classes, compared to 128 FrameNet parent frames[17] and 101 VerbNet top-level classes in which the 825 verbs in the present sample participate. However, the overlap with VerbNet top-level classes is only marginally lower than that found between VerbNet and FrameNet at the same hierarchy level, despite the similar number of classes/frames in those resources and the known commonalities between the two classifications (Baker and Ruppenhofer, 2002).

For Phase 2, the fact that the two resources include overlapping classes, while the clusters extracted from the distance matrices are exclusive, negatively affects the overlap scores. Nevertheless, they are again within the range of baseline scores measuring the overlap between FrameNet and VerbNet on the same subsets of verbs. At the top hierarchy level, the clusters derived from the Phase 2 distance matrix for Class 10 movement verbs align more strongly with each expert resource than these resources align with each other (0.527, 0.666 > 0.504 (FNxVN)). The encouraging B-Cubed results against VerbNet classes (> 0.6 for classes #10 and #3) suggest that the proposed arena-based approach allows annotators to intuitively differentiate between degrees of overlap in verbs' properties and create, by deliberate word placements, clusters of

[17]Further analysis could also explore indirect inheritance.

	FrameNet				**VerbNet**				**FNxVN**	
	parent		frame		top		1st		top	1st
P1	0.247		-		0.302		-		0.313	0.428
P2 $k =$	*optimal*	*gold*	*optimal*	*gold*	*optimal*	*gold*	*optimal*	*gold*		
#10	0.527	0.247	0.407	-	0.666	0.259	0.481	0.337	0.504	0.596
#15	0.470	0.289	0.448	-	0.407	0.324	0.449	0.388	0.453	0.533
#3	0.546	0.501	0.578	-	0.642	0.566	0.618	0.616	0.668	0.720

Table 4.3 Comparison of Phase 1 (P1) classes and clusters extracted from Phase 2 (P2) distance matrices for classes 10, 15, and 3 against FrameNet fine-grained frames (**frame**) and parent frames (**parent**), and VerbNet top-level (**top**) and first-level (**1st**) classes. For context, I also compute baseline overlap scores between FrameNet and VerbNet (**FNxVN**), at two levels: top-level VN classes against parent FN frames (**top**) and first-level VN classes against FN frames (**1st**). All scores are B-Cubed F-scores, measuring the overlap between P1 classes/P2 clusters and the FrameNet and VerbNet classes for the shared verbs, and between FrameNet frames and VerbNet classes themselves, on the same sets of shared verbs (higher score = greater overlap). For Phase 2, *optimal* columns show scores obtained for the optimal clustering solution in terms of F1 score, determined iteratively over values of $k = \{1, ..., N\}$, where N is the size of each class (#10 – 100, #15 – 87, #3 – 30); *gold* columns show scores for clustering solutions with $k = K_{gold}$, where K_{gold} is the number of classes in FrameNet or VerbNet in which the shared verbs participate. I do not report *gold* values where the number of gold classes was larger than the number of verbs in a given P2 sample ($K_{gold} > N$), due to multiple class membership of individual verbs in *gold* resources.

similar verbs within a broader related set that reflect some of the fine-grained class divisions in the semantic-syntactic expert-created lexicon. This is also observable when comparing the similarity judgments from Phase 2 (i.e., distances in the arrangement space, therefore smaller scores signify greater similarity) to the taxonomic distance within the VerbNet class hierarchy between the same pairs of verbs, where the growing taxonomic distance is reflected in the growing dissimilarity scores for verbs (a) belonging to the same low-level VerbNet subclass (17.1-1-1: *throw* – *toss*, dissimilarity $d = 0.273$), (b) verbs in a class-subclass relation (17.1-1 *fling* – 17.1-1-1 *toss*, $d = 0.350$), (c) verbs sharing the same first-level class (11.4: *tow* – *haul*, $d = 0.421$), or (d) the same top-level class (11.4 *tow* – 11.3 *take*, $d = 0.584$), or (e) belonging to different top-level classes (11 *tow* – 17 *chuck*, $d = 0.943$).

Semantic vs. Semantic-Syntactic Classes: Qualitative Comparison

Given the substantial overlap between VerbNet classes and the clusters emerging from the spatial data, it is worth examining where they correspond and where they diverge

Class 7	Class 13
{ *accept, agree, approve, concur* }	{ *boom, erupt, explode, pop* }
{ *ask, inquire, request* }	{ *squeak, squeal* }
{ *advise, clarify, educate, explain, inform, teach* }	{ *chirp, hum*}
{ *comfort, console, protect, soothe* }	{ *holler, roar, scream, yell* }
{ *collaborate, cooperate* }	{ *discover, find* }
{ *depend, rely* }	{ *knock, rattle, tap* }
{ *debate, disagree, protest* }	{ *crack, crackle, crunch, snap* }
{ *say, speak, talk* }	{ *giggle, laugh* }
{ *reply, respond* }	{ *croak, whisper* }

Table 4.4 Examples of fine-grained DBSCAN clusters extracted from dissimilarity matrices of two classes, #7 (left) and #13 (right).

in more detail. The availability of complete distance matrices for each Phase 1 class enables clustering analyses aimed at producing fine-grained subclasses within each semantic domain. Table 4.4 shows examples of narrow semantic clusters output by DBSCAN algorithm (*Density-Based Spatial Clustering of Applications with Noise*) (Ester et al., 1996),[18] which detects clusters by identifying high concentrations of data points in the data space; here, it was run on top of two of the Phase 2 distance matrices (Class 7 and Class 13).

Placing this analysis in a larger context of previous work on semantic-syntactic verb classes, several interesting patterns can be observed. For example, as far as attitude verbs are concerned, in Phase 1 we can observe certain divisions corresponding to splits posited by Bolinger (1968), i.e., between representational and preferential verbs. The first group concerns expression of judgments of truth (e.g., *think, believe*), while the second includes verbs which express preferences (e.g., *desire, wish*), and the separation is reflected in the verbs' ability to be postposed (in English) and mood selection (in Romance languages) (Anand and Hacquard, 2008; Bolinger, 1968; Searle and Vanderveken, 1985; Villalta, 2000, 2008; White et al., 2014). In the semantic clustering task, the participants split these two types of verbs between classes #9 and #8, respectively, the first including cognitive verbs (e.g., *believe, accept, understand*) and the latter verbs of positive attitudes or emotions (e.g., *crave, want, prefer*).

Attitude verbs have been much studied in the context of language acquisition (Fisher, 1996; Gleitman, 1990; Harrigan et al., 2016; Papafragou et al., 2007). Given that they lack direct correlates in the physical environment, it has been argued that

[18] I selected DBSCAN as it does not require specifying the value of k upfront and thus avoids explicitly imposing a predetermined cluster granularity. The algorithm finds high-density areas separated by lower density areas, with a tunable ϵ parameter which represents the radius within which two points can be neighbours (see Appendix A for details).

it is the sentence structure and the number of nominal arguments accompanying the verb that help narrow down its meaning. Patterns of syntactic behaviour have been empirically shown to correspond to coarse-grained semantic distinctions such as the split between physical change-of-state verbs (e.g., *break*) and mental state verbs (e.g., *believe*) (Fisher et al., 1991; Lederer et al., 1995). Whereas the exploration of semantic correlates of subcategorisation frames led Gleitman (1990) to propose categorising verbs into types such as 'perceptual', 'mental', or 'transfer', each with a characteristic set of syntactic patterns. We can find correspondences between these three categories and Phase 1 classes #17, including perception verbs like *hear, see, stare, watch*, #9 with cognitive verbs such as *think, believe, analyze, examine* and #5 with transfer verbs such as *give, get, lend, receive.* Nevertheless, at higher granularity levels, we can expect to find purely semantic distinctions without relevance for the verbs' syntactic behaviour, given its sensitivity only to certain semantic contrasts (Grimshaw, 1979; Jackendoff, 1972; White et al., 2014). Antonyms are a canonical example, given their paradigmatic similarity: verbs *decrease* and *increase*, and *raise* and *lower*, share syntactic behaviour and therefore are classified together in VerbNet (Class 45.6.2), while being separated in the spatial arrangements in Phase 2 (see Figure 4.7 again). Analogously, while VerbNet class 'admire-31.1' includes verbs expressing both positive and negative subjective judgments, these are separated already in Phase 1 based on their opposite polarity, forming a separate class of positive (#8) and negative (#6) emotion verbs. Other examples can be found in the 'manner of speaking' class in VerbNet, which includes verbs such as *yell, holler, whisper, croak, chirp.* These all fall within Class 13 (Table 4.4), within which Phase 2 participants record finer-grained semantic distinctions based on the quality and intensity of sound, separating [*holler, roar, scream, yell*] from [*croak, whisper*] and [*chirp, hum*].

One advantage of the spatial similarity dataset is that it allows for flexible tuning of cluster granularity, based on the intended usage. Starting from the same symmetric matrix of distances, we can obtain a cluster hierarchy, which could provide a starting point for manual annotation with additional class-specific information, such as semantic roles or syntactic realisations of the verbs' arguments. For instance, Figure 4.9 presents the hierarchical structure yielded by agglomerative clustering with complete linkage (Defays, 1977) for Class 1 of physical contact verbs. The dendrogram traces the sequence in which clusters merged and visualises the distance at which each fusion took place, starting from the bottom, with each word in its own cluster. We can see that the primary distinction made by the human judges, represented by the highest-level split, separates verbs describing nonviolent (*touch, hug, cuddle, brush*) and violent physical

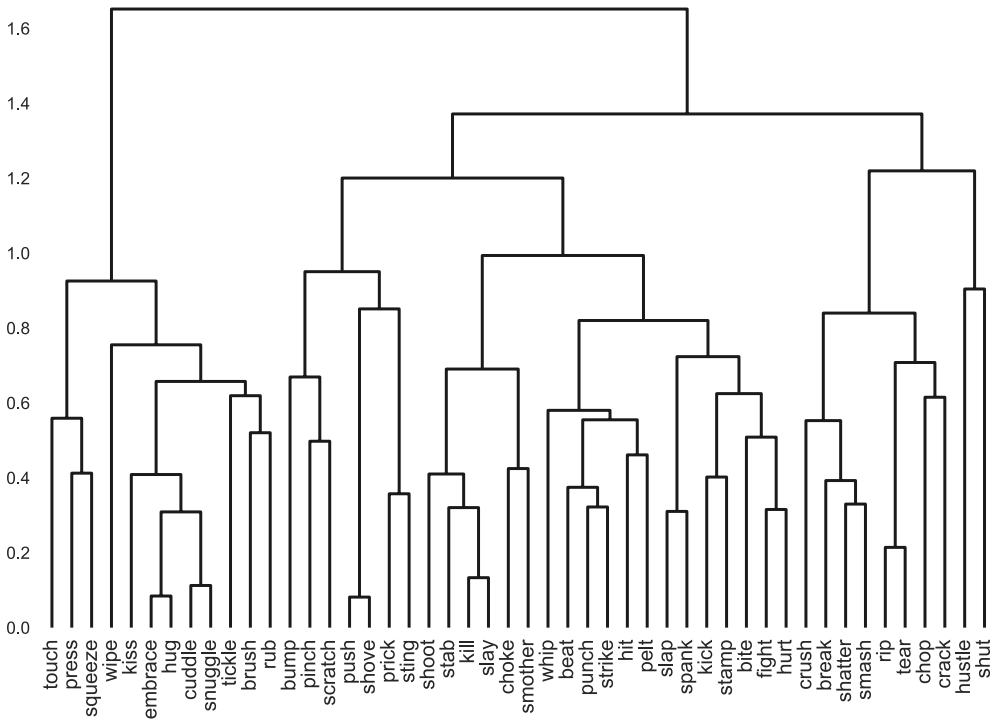


Fig. 4.9 Hierarchical agglomerative clustering output for Class 1.

contact verbs (*shove, prick, choke, beat*). Within the first group, we can observe finer-grained splits reflecting differences in manner, e.g., *tickle, brush, rub*, which involve the component of repeated sweeping or back-and-forth motion, or the cluster of affectionate physical contact, *embrace, hug, cuddle, snuggle*. In the second group, we observe a high-level split which separates verbs of destruction (e.g., *crush, break, shatter, smash*) (with the outliers *hustle* and *shut* on the right) from the remainder of verbs. Bottom-up inspection of the structure shows low-level links joining pairs of synonyms (e.g., *kill – slay*, *push – shove*, or *rip – tear*). There is also a separate cluster of verbs incorporating the pair *kill – slay*, which includes verbs referring to violent acts typically leading to death (e.g., *shoot, stab, choke, smother*).

As an exploratory approach to data analysis, hierarchical clustering enables in-depth examination of the patterns of semantic judgments and identification of meaning components associated with each split. By cutting the dendrogram at different heights, different cluster granularities can be derived, yielding a nested structure of classes and subclasses. These, after manual review, could be employed to produce evaluation data allowing for testing the models' capacity to create fine-grained semantic classifications and taxonomies automatically.

Spatial Similarity vs. WordNet Relations

Phase 1 produces classes encompassing a range of finer-grained lexical relations within related words, including synonymy (*try – attempt*), hyper/hyponymy (*think – rationalize*), cohyponymy (*suck – sip*), or antonymy (*appear – disappear*). In Phase 2, annotators differentiate between these relations, deciding on relative semantic distances between words participating in them. The collected distance matrices include pairwise distances between all possible pairings of items within a class, and hence encode all pairwise relations within that set. This allows for zooming in on a particular relation type and observing how it is reflected in the pairwise scores for word pairs which exemplify it.

To illustrate this, I compute average dissimilarity scores for pairs exhibiting the four relations (extracted from WordNet) in the present dataset and compare them to average SimVerb similarity scores and USF association scores for the same pairs (Table 4.5). We can see smallest dissimilarity scores for synonyms, which increase through hyper/hyponyms and cohyponyms as the degree of association decreases. Antonyms are furthest apart, despite their relatively high (compared to synonyms) USF association score. This supports the hypothesis that the Phase 2 multi-arrangement task setup allows annotators to differentiate between similarity and association, as well as a range of fine-grained lexical-semantic relations.

These findings are noteworthy in light of the different strategies employed by the two paradigms that produced SimVerb and SpA-Verb (each characterised by a different means of expressing similarity judgments) to handle the relatedness/similarity distinction and antonymy. In SimVerb, the guidelines instruct annotators to assign low numerical scores both to antonyms (*stay – leave*) and to related but non-similar words, for example, *walk – crawl* (Gerz et al., 2016). In the present approach, this is handled by means of the two-phase design. First, similar and related verbs are grouped together to form broad semantic classes. Then, fine-grained similarity judgments are made amongst already related verbs. As emphasised in Section 4.3.2, this avoids conflating scores for unrelated and antonymous pairs: the former are split into distinct Phase 1 classes, which reserves low similarity scores (= large distances in the arena) for the latter. This is confirmed by manual inspection of the outputs of both phases: in Phase 1, antonymous words are found in the same broad groups, based on their relatedness (e.g., antonymous pairs *stay* and *leave*, and *lose* and *gain* end up clustered together),[19]

[19]However, there are exceptions: positive (e.g., *love*) and negative (e.g., *hate*) emotion verbs form two different classes. There are also separate classes of 'construction' and 'destruction' verbs. See Table 4.2.

	SpA-Verb	**SimVerb**	**USF**
Synonymy	0.482	6.79	0.190
Hyper/hyponymy	0.593	4.00	0.120
Cohyponymy	0.686	2.79	0.060
Antonymy	1.019	0.54	0.154
Score range	0.065 - 1.720*	0 - 10**	0 - 1

Table 4.5 Average SpA-Verb dissimilarities, SimVerb similarity ratings and USF free association scores across shared pairs representing four semantic relations (extracted from WordNet): synonymy, hyper/hyponymy, cohyponymy, antonymy. Score ranges represent the actual interval of scores in each source (*SpA-Verb scores are based on Euclidean distances scaled to have an RMS of 1 (Eq. 4.9) to guarantee inter-class consistency, as detailed in Section 4.3.1. **SimVerb scores were originally collected as 0-6 ratings and scaled linearly to the 0-10 interval by Gerz et al. (2016)).

while in Phase 2, antonyms are placed far apart in the arena: out of 67 antonymy pairs shared with SimVerb (i.e., labelled ANTONYMS in SimVerb), only 2 are placed closer in the arena (*inhale – exhale* and *sink – swim*). This is also illustrated by the RDM in Figure 4.7, where separate clusters (dark areas) are formed by verbs such as *raise, rise, grow* and *diminish, decline, lower*, whereas *finish* is kept separate from *begin* and *start*.

SpAM for Graded Multi-Way Lexical Relations

Despite the parallels in the treatment of semantic relations in SpA-Verb and SimVerb, comparative analyses shed light on some important differences between judgments yielded by the proposed SpAM method and pairwise rating-based methods, revealing potential benefits not offered by pairwise datasets. As all verbs are simultaneously judged in the context of all other related verbs, not only pairwise but also multi-way relations can be captured, reminiscent of lexical taxonomies. Degrees of similarity can be recorded in a meaningful way and adjusted in the presence of another word, distinguishing between dissimilar unrelated words and words which, despite their lack of similarity to the target word, nonetheless stand in some lexical-semantic relation to it. Such relations are exemplified, for instance, by lexical triplets (e.g., *try – succeed – fail*), where the first element expresses the necessary presupposition for the pair of complementaries (i.e., words which divide some conceptual domain into two mutually exclusive parts) (Cruse, 1986). According to Cruse (1986), the binary relation between satisfactives *try* and *succeed* (an attempt vs. successful performance) is a weak form of oppositeness, while *succeed – fail* present a strong oppositeness.

The two-phase design allows capturing these three-way relations simultaneously, grading similarity and oppositeness: synonyms *try – attempt* receive a low 0.283 score (smaller score = greater similarity), satisfactives *try – succeed* and *attempt – succeed* 0.891 and 0.861, respectively, *try – fail* 0.960 and antonyms *succeed – fail* 1.063. To compare, in SimVerb, where pairs are judged independently, *attempt – succeed* receive a 2.16 score on the 0-10 scale. This score is lower than those for dissimilar and unrelated pairs such as *perish – sob* and *blur – rush* (2.49), so the information about *attempt – succeed* standing in some meaningful relation as opposed to the unrelated pairs is not captured. There is no consensus on what the best treatment of such cases is, but distinguishing between weak oppositeness, strong oppositeness and unrelatedness and capturing meaningful degrees of dissimilarity (e.g., '*attempt – succeed* is less dissimilar than *perish – sob*') may be beneficial for reconstruction of complex semantic hierarchies and bottom-up creation of lexical taxonomies which go beyond pairwise similarity.

Similar comparisons can be made about capturing troponymy. In WordNet, *jump* forms a synonym set (synset) with *leap, bound, spring*, and their troponyms include *hop, skip* and *bounce*, which describe a specific manner of jumping. The troponymy relation is reflected in small distances in the arena: for example, *spring – bounce* 0.373 and *jump – skip* 0.277. In SimVerb, *spring – bounce* have a high 8.80 rating, but *jump – skip* receive a much lower 5.48 score, despite holding an analogous troponymy relation. This is a score equal to the similarity rating of *embarrass – blush*, which are strongly associated, but dissimilar. Meanwhile *jump* and *skip* display a high degree of semantic overlap (i.e., describe a similar kind of motion). The availability of scores for all possible pairings allows for tracing and reconstructing semantic and taxonomic links like those in WordNet: for the 'choose, select' WordNet synset, synonymous *choose – select* are very close together (0.121), close but slightly further away from their direct hypernym *decide* (*choose – decide* 0.283, *select – decide* 0.216), and still further away from their troponym *elect* (*elect – choose* 0.544, *elect-select* 0.512). The distance grows slightly with inherited hypernymy across three levels (*elect – decide* 0.592) and co-hyponymy (*elect – pick* 0.556). In SimVerb, *elect – choose* receive score 8.47, but *elect – select* 5.15, despite standing in analogous relations. Such discrepancies in scores for similar relations may be a consequence of judging pairs in isolation in SimVerb: when simultaneously presented with all verbs belonging to the 'jumping' or 'choosing' domain (in a given sample), it should be easier to record consistent similarity judgments across relations of the same kind and degree (e.g., troponymy), which is a considerable benefit of the proposed SpAM approach.

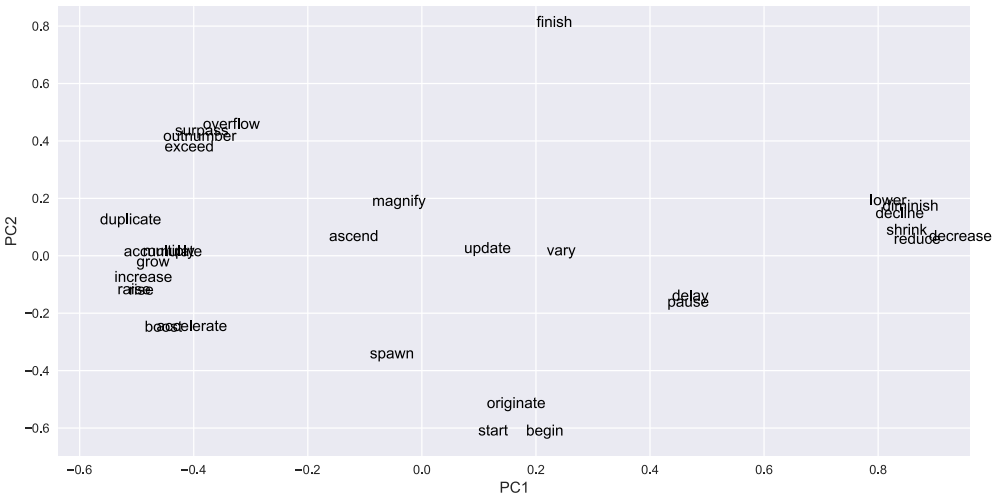


Fig. 4.10 Visualisation of PCoA applied to the Phase 2 distance matrix for rate of change verbs (Class 3).

SpAM for In-depth Exploration of Semantic Domains

The complete RDMs from Phase 2 permit in-depth analyses of the resultant nets of semantic relations. In order to understand better what information is being captured and what underlying features and dimensions inform human similarity judgments, I applied Principal Coordinates Analysis (PCoA) (Gower, 1966) to each distance matrix to examine the dimensions and meaning components characterising the semantic space in question (Gärdenfors, 2004). Figure 4.10 provides an example visualisation of the main axes (PC1, PC2) for Class 3. The first dimension (PC1), which explains 50% of the variance, roughly corresponds to word polarity: positive rate of change verbs are clustered on the negative side of the PC1 axis (e.g., *accelerate, increase, raise* and *exceed, surpass, overflow*), less strongly polarised verbs closer to the middle (e.g., *update, vary*), and the negative rate of change verbs on the far right (e.g., *lower, decline, shrink*). Whereas the second axis (PC2) constitutes the dimension of difference for verbs expressing inception and termination, with *start, begin, originate* and *finish* occupying its opposing poles.

An analogous inspection of Class 10, comprising verbs of motion and position (e.g., *lounge, sit*), in three dimensions (Figure 4.11; note that the projection was rotated during the analyses; see Figure C.4 for two other views of the projection) revealed that the most salient dimension roughly corresponds to dynamicity: verbs expressing fast and/or abrupt motion cluster towards one end of the PC1 axis (e.g., *flutter, pounce, leap, spring, soar, glide*) and those conveying a motion towards a stationary state or being in a static position at the opposite end (e.g., *lounge, poise, sit, retire, stop,*

settle). Interestingly, the second dimension reflects some of the differences in the degree of verb agentivity: verbs which select Agent subjects gravitate towards the positive end of the second axis (e.g., *chase, enter, ride, drive, hunt*) while the negative side of the axis is populated by verbs with Patient subjects (e.g., *slip, tumble, fall, slide, sink*). Finally, we can trace the changing medium of motion along the third dimension, from verbs describing motion in water (*sink, flow, float, swim*) at the negative end, to the more prototypical verbs describing movement or position on land at the opposite end (e.g., *trail, walk, run, stand, lean*), with verbs describing motion in the air (e.g., *fly, soar, glide*) grouped closer to the middle.

Similar analyses can be employed to identify the most salient semantic features for each class and to gain a deeper understanding of the implicit meaning dimensions underlying human similarity judgments, and by extension, the organisation of human lexical semantics and the representations constituting a conceptual space (Gärdenfors, 2004; Hollis and Westbury, 2016). In Chapter 5, I demonstrate how such visualisations of the spatial similarity data can serve investigations of cross-lingual similarities and variation.

4.8 Evaluation with Representation Learning Architectures

In order to fully analyse the properties of the proposed two-phase evaluation dataset creation method, and the dataset's potential as an evaluation resource, I evaluate a representative selection of state-of-the-art representation models on two tasks, corresponding to the two phases of the proposed design: (1) clustering, using Phase 1 classes as gold truth, and (2) word similarity, using pairwise scores from the entire SpA-Verb (29,721 pairs) and the thresholded subset (SpA-Verb-THR), including 10,371 pairs from the classes with Spearman's IAA ≥ 0.3, as well as chosen subsets with different semantic characteristics. Several different reference scales have been proposed for the interpretation of the Spearman's correlation coefficient in terms of descriptors such as "strong," "moderate," or "weak" (Akoglu, 2018; Chan, 2003; Schober et al., 2018), and it has been noted (Schober et al., 2018) that the range of values being assessed should be considered in the interpretation (i.e., a wider range of values tends to show a higher correlation than a smaller range, as is the case for the similarity data collected in the present work, see Figure 4.8). I choose $\rho = 0.3$ as a confidence threshold in light of these considerations and exclude the classes where IAA results show low positive correlation from the thresholded dataset. This subset comprises data from 10 of the 17 classes (see

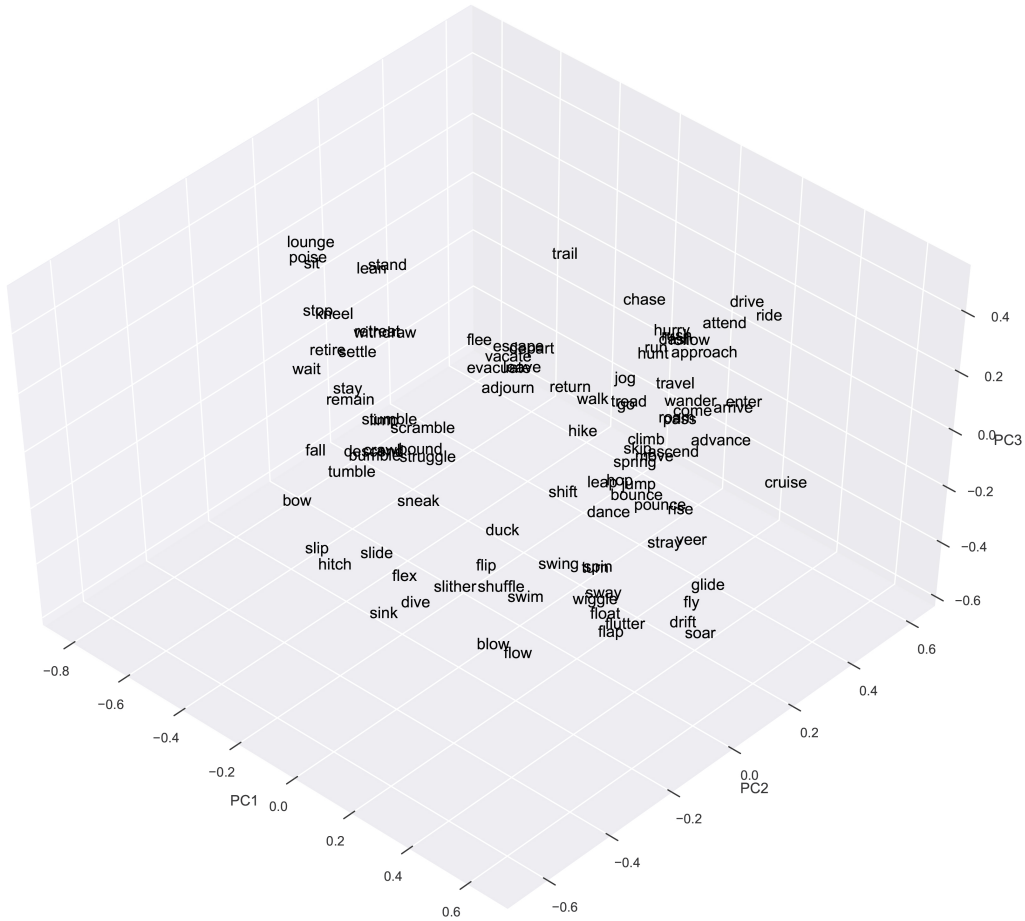


Fig. 4.11 Visualisation of PCoA applied to the Phase 2 distance matrix for verbs of motion (Class 10).

Table 4.2). In the analyses presented in this chapter, I also compare model performance on the subset of pairs from SimVerb-3500 within the SpA-Verb dataset (1,682 pairs) and the original SimVerb-3500 ratings. The selected architectures represent different modeling assumptions, data requirements, and underlying methodologies, which I briefly describe below. For all models, d refers to the embedding dimensionality, and ws is the window size in case of bag-of-words (BOW) contexts.

4.8.1 Representation Models

Included in the selection is an unsupervised model which learns solely from distributional information in large text corpora, the skip-gram with negative sampling (SGNS) (Mikolov et al., 2013b) with BOW contexts trained on the English preprocessed Polyglot Wikipedia (Al-Rfou et al., 2013) by Levy and Goldberg (2014) (SGNS-BOW2; $d = 300$

and $ws = 2$ as in prior work).[20] I also include models using subword-level information. The first is an extension of the original CBOW model (Mikolov et al., 2013b) (CBOW-CC) with position weights and subword information, introduced by Grave et al. (2018). Before taking the sum of context words, each word vector is element-wise multiplied by a position dependent vector as done in Mnih and Kavukcuoglu (2013). Word vectors are sums of their constituent n-grams as in Bojanowski et al. (2017) and Mikolov et al. (2018). The method is trained on deduplicated and tokenised English Common Crawl corpus.[21]

I also experiment with a more recent representation model that computes dynamic word representations conditioned on the surrounding word context (Peters et al., 2018) (ELMo-Static, $d = 300$). The model is based on a deep character-level language model implemented as a bidirectional LSTM. To be comparable to other static word embeddings, I use the static context-insensitive fully character-based type layer to obtain static ELMo vectors; the same technique was used by (Peters et al., 2018). I use the ELMo variant pretrained on the 1 Billion Word Benchmark.[22] I also evaluate two approaches to extracting word-level representations from pretrained Transformer-based BERT models (Devlin et al., 2019), whereby words are fed into the model (i) *in isolation* or (ii) *in context*. To obtain lexical-level representations with the first method (i), I follow prior work (Liu et al., 2019b; Vulić et al., 2020a) and compute each verb representation by (1) feeding it to a pretrained BERT model *in isolation*; and then (2) averaging the H hidden representations (bottom-to-top) for each of the verb's constituent subwords. I then (3) average the resulting subword representations to produce the final d-dim vector. This approach does not require any additional external corpora for the induction of such BERT-based embeddings. I experimented with different values of $H = \{4, 6, 8\}$, as well as an alternative approach where only the representation from the input embedding layer is used, without layer-wise averaging, as done in prior work (Conneau et al., 2020b; Wang et al., 2019b). I also examined two approaches to subword representation averaging, one where special tokens (`[CLS]` and `[SEP]`) are included and one where they are excluded from the averaging step. I found that exclusion of special tokens results in consistently stronger performance across models and values of H. I report results for the strongest performing configuration

[20] https://levyomer.wordpress.com/2014/04/25/dependency-based-word-embeddings/
[21] https://fasttext.cc/docs/en/crawl-vectors.html
[22] https://allennlp.org/elmo

across tasks, averaging representations from the first $H = 6$ layers and excluding special tokens from subword averaging.[23]

The second method (ii) allows for encoding word meaning *in context*, using external text corpora, by first learning N token-level representations for each word and subsequently aggregating them into a static type-level representation as before. I chose English Europarl (Koehn, 2005) as the external corpus, from which I (1) randomly sampled N sentences containing each word in the sample; Then, (2) I computed each word's representation in N sentential contexts (averaging over constituent subword representations and hidden layers as in steps (2)-(3) of method (i) above), and finally (3) averaged over the N representations to obtain the final representation for each word. I report results for three values of N: 10, 100, and 500 (CONTEXT-10, CONTEXT-100, CONTEXT-500). I probe three different variants of English BERT: BERT-BASE ($d = 768$), BERT-LARGE ($d = 1,024$), and BERT-LARGE with whole word masking (BERT-LARGE-WWM, $d = 1,024$), available in the Transformers repository (Wolf et al., 2019).[24]

Furthermore, I test architectures that leverage linguistic information available in external semantic resources. I include sparse binary high-dimensional vectors ($d = 172,418$) proposed by Faruqui and Dyer (2015) (NON-DISTRIBUTIONAL vectors). The vectors are based on a wide variety of hand-crafted linguistic resources such as WordNet, Supersenses, FrameNet, Emotion and Sentiment lexicons, Connotation lexicon, among others.[25] Moreover, I evaluate a retrofitting method that generalises the model of Wieting et al. (2015) and counter-fitting of Mrkšić et al. (2016) (SGNS+ATTRACT REPEL; AR). It fine-tunes any word vector space by pulling words standing in desirable (i.e., ATTRACT) relations closer together, while simultaneously pushing words in undesirable relations (i.e., REPEL) away from each other (Mrkšić et al., 2017). I evaluate best-performing AR-specialised vectors, reaching human performance on SimLex and SimVerb, introduced by Vulić and Korhonen (2018): They use SGNS-BOW2 as the starting space ($d = 300$), and WordNet and Roget's Thesaurus (Kipfer, 2009) as the source of external knowledge (see Section 2.3.1). Additionally, I evaluate two collections of BOW2 distributional vectors specialised by Attract-Repel (as in Vulić et al. (2017b)) using constraints drawn from VerbNet (Kipper et al., 2006) (BOW2-VN) and FrameNet (Baker et al., 1998) (BOW2-FN) to reflect the semantic

[23]While exclusion of special tokens produced consistently stronger embeddings, I observed more variation in scores for different values of H, suggesting careful tuning of this parameter is necessary to achieve optimum performance. Supplementary results for the different word embedding extraction configurations from BERT models are included in Appendix C.5.

[24]github.com/huggingface/transformers

[25]https://github.com/mfaruqui/non-distributional

(shared FrameNet frames) and syntactic-semantic (membership in VerbNet classes) relationships between verbs encoded in those resources.

4.8.2 Clustering

For each collection of distributional or specialised vectors, I apply a choice of three off-the-shelf clustering algorithms to group the 825-verb sample (used in Phase 1) into classes based on their similarity: the MNCut spectral clustering algorithm (Meila and Shi, 2001), as in prior work (Brew and Schulte im Walde, 2002; Sun and Korhonen, 2009; Sun et al., 2010), the k-means clustering (Brew and Schulte im Walde, 2002; Sun et al., 2010), and agglomerative clustering with average linkage (see Appendix A for details of clustering algorithms).

I apply standard evaluation metrics from previous work on verb clustering (Falk et al., 2012; Ó Séaghdha and Copestake, 2008; Sun and Korhonen, 2009; Sun et al., 2010; Vulić et al., 2017b). Modified purity (mPur), that is, the mean precision of automatically induced verb clusters, is calculated as:

$$\text{mPur} = \frac{\sum_{C \in \textbf{Clust}, n_{prev(C)} > 1} n_{prev(C)}}{\#test\_verbs} \tag{4.10}$$

where each cluster C from the set of all K_{Clust} induced clusters **Clust** is associated with its prevalent gold class (from Phase 1), and $n_{prev(C)}$ is the number of verbs in an induced cluster C taking that prevalent class, with all other verbs considered errors. $\#test\_verbs$ is the total number of test verb instances.[26] Weighted class accuracy (wAcc), which targets recall, is computed as:

$$\text{wAcc} = \frac{\sum_{C \in \textbf{Gold}} n_{dom(C)}}{\#test\_verbs} \tag{4.11}$$

where for each class C from the set of gold standard classes **Gold** (Phase 1 classes, $K_{Gold} = 17$) I identify the dominant cluster from the set of induced clusters having most verbs in common with C ($n_{dom(C)}$). I combine the two metrics into an F1 score, calculated as the balanced harmonic mean of mPur and wAcc.

[26] As in prior work, I discard clusters with $n_{prev(C)} = 1$ from the count to avoid bias from singleton clusters (Sun and Korhonen, 2009; Sun et al., 2010; Vulić et al., 2017b).

Results and Discussion

The strongest results are obtained from spectral clustering (as previously in Scarton et al. (2014); Vulić et al. (2017b)), 0.01 points on average ahead of k-means and 0.02 in favour of agglomerative clustering. Table 4.6 summarises the F1 spectral clustering scores for the chosen vector collections, for the optimal value of k (optimal) and for $k = K_{Gold}$ (gold). We can note strongest results from the FrameNet-specialised vectors (BOW2-FN), which is an outcome attributable to the nature of the Phase 1 classes, characterised by thematic similarity, which also underlies FrameNet frames. While the absolute scores are not high (F1 < 0.5 for all vector collections), the relative scores are informative. Progressive improvement in performance can be observed across the different types of external knowledge used for vector space fine-tuning, from the WordNet- and Roget's Thesaurus-specialised SGNS+Attract-Repel vectors, through VerbNet-specialised BOW2-VN embeddings, up to the top-performing BOW2-FN. FrameNet is a fine-grained resource including 1224 semantic frames, some of which describing very specific semantic scenarios. It is therefore quite different in structure from the broad Phase 1 classes. However, the fact that FrameNet knowledge boosts clustering performance suggests the rationale behind human judgments in the Phase 1 rough clustering task aligns somewhat with the hypothesis underlying the organisation of verbs into FrameNet frames.

Overall, stronger performance can be observed from the static distributional models compared to the Transformer-based BERT architectures. Manual inspection of clusters output by the latter systems reveals groupings biased by subword information (e.g., BERT-BASE clusters words *nag, wag*; *thaw, yawn*, and *soar, soak* together), rather than reflecting semantic overlap (e.g., as in the "sound" cluster (*cry, squeal, squeak, roar, rattle, hoot, scream*, etc.) produced by BOW2-FN). The extracted BERT word-level representations capture substantial surface-level information which impacts cluster assignments; However, the embeddings computed *in context* offer improvements over their *in isolation* counterparts. We can also note that the number of contextual representations (N) aggregated into the final word-level embeddings which yield strongest results varies between models and the values of k. Contextualised BERT-LARGE embeddings achieve the highest scores across BERT models, producing clusters characterised by greater semantic coherence (e.g., "possession" verbs including *gather, buy, collect, possess, obtain, steal, borrow, earn, get*, and "cognitive" verbs like *assume, realize, examine, compute, analyze, doubt, guess, understand*). In the next section, I further examine the impact of computing representations in context rather than in

Model (Dimensionality)	**F1 optimal**	**F1 gold**
SGNS-BOW2 (300)	0.355	0.326
CBOW-CC (300)	0.426	0.383
ELMo-Static (300)	0.394	0.387
NON-DISTRIBUTIONAL (172,418)	0.391	0.360
SGNS+ATTRACT-REPEL (300)	0.392	0.354
BOW2-VN (300)	0.416	0.404
BOW2-FN (300)	**0.444**	**0.429**
BERT-BASE (768) (ISO)	0.338	0.310
CONTEXT-10	0.338	0.312
CONTEXT-100	0.340	0.322
CONTEXT-500	0.332	0.309
BERT-LARGE (1,024) (ISO)	0.297	0.269
CONTEXT-10	0.339	0.325
CONTEXT-100	0.334	0.304
CONTEXT-500	0.350	0.323
BERT-LARGE-WWM (1,024) (ISO)	0.323	0.308

Table 4.6 F1 scores obtained by representation models on the clustering task, for the optimal value of k (F1 optimal) and for $k = K_{Gold}$ (F1 gold), evaluated against Phase 1 classes. For BERT-BASE and BERT-LARGE models, I evaluate both the embeddings computed *in isolation* (ISO) and *in context*, for three values of N (10, 100, 500), corresponding to the number of contextualised representations aggregated into the final word-level embedding. Numbers in brackets refer to vector dimensionality.

isolation on the quality of the semantic information captured in the word similarity task.

4.8.3 Word Similarity

Table 4.7 reports the results of evaluation of the chosen models on SpA-Verb (29,721 pairs) and the thresholded subset (SpA-Verb-THR), and the subset of pairs shared with SimVerb-3500 (1,682 pairs), using both the original SimVerb scores and those obtained via the proposed arena-based method. The reported scores are Spearman's ρ coefficients of correlation between the ranks derived from models' similarity scores (i.e., cosine distances in the embedding space) and from human similarity judgments in Phase 2.

Results and Discussion

A number of interesting observations can be drawn from the evaluation. First, we can note that the highest scores are obtained by linguistically-informed models, drawing

Model (Dimensionality)	SV-3500	SV∩SpA_SVs	SV∩SpA_SpAs	SpA-Verb	SpA-Verb-THR
SGNS-BOW2 (300)	0.275	0.197	0.136	0.179	0.158
CBOW-CC (300)	0.365	0.264	0.242	0.271	0.305
ELMo-Static (300)	0.414	0.327	0.310	0.230	0.227
NON-DISTRIBUTIONAL (172, 418)	0.606	0.543	0.479	0.295	0.310
SGNS+ATTRACT-REPEL (300)	**0.766**	**0.730**	**0.567**	**0.385**	**0.394**
BERT-BASE (768) (ISO)	0.338	0.224	0.207	0.235	0.240
CONTEXT-10	0.436	0.326	0.228	0.262	0.266
CONTEXT-100	0.438	0.326	0.231	0.265	0.271
CONTEXT-500	0.439	0.327	0.231	0.265	0.270
BERT-LARGE (1,024) (ISO)	0.319	0.240	0.188	0.224	0.215
CONTEXT-10	0.403	0.305	0.226	0.255	0.269
CONTEXT-100	0.402	0.304	0.225	0.256	0.270
CONTEXT-500	0.403	0.304	0.225	0.256	0.270
BERT-LARGE-WWM (1,024) (ISO)	0.396	0.307	0.257	0.237	0.246

Table 4.7 Evaluation of selected state-of-the-art representation learning models on the full SimVerb-3500 dataset (**SV-3500**), the subset of pairs shared by SimVerb and the SpA-Verb dataset, using both the original SimVerb scores (**SV∩SpA_SVs**) and scores obtained via the proposed arena-based method (**SV∩SpA_SpAs**), as well as the full similarity dataset (**SpA-Verb**) and the thresholded subset (**SpA-Verb-THR**) of the whole dataset (10,371 pairs from the classes with IAA ≥ 0.3). All scores are Spearman's ρ correlations. Numbers in brackets refer to vector dimensionality.

from diverse rich lexical resources to better capture a range of semantic relations and phenomena (e.g., synonymy and antonymy (Vulić and Korhonen, 2018), sentiment polarity and connotation (Faruqui and Dyer, 2015)).

The fact that these representations score the highest on SpA-Verb reveals the potential of the spatial arrangement-based method to capture fine-grained semantic properties. It also indicates that non-expert native speakers without formal linguistic training reflect on the components of word meaning and perform some form of linguistic analysis intuitively. This suggests that the spatial method may lend itself to the creation of rich lexical resources, and not only simple pairwise similarity datasets. We can observe that the performance of pretrained encoders on SpA-Verb (and SimVerb) lags behind that of the top-performing static representations; However, they consistently outperform the SGNS-BOW2 model and, at their strongest, are competitive with the CBOW-CC and ELMo-Static vectors. Within the three pretraining models, we can observe that BERT-BASE mostly outperforms the larger BERT-LARGE model, whose

performance improves, however, when using word-level masking (BERT-LARGE-WWM). We can again note a clear advantage of computing word-level representations *in context*, which better leverages the ability of the Transformer models to learn dynamic representations of meaning: there are noticeable improvements over the *in isolation* (ISO) variant across the board, even if the number of contextualised (token-level) representations aggregated into the final word-level (type-level) representation has little to no impact on performance.

SpA-Verb vs. SimVerb

Interesting observations can be drawn from an analysis of correlation between model rankings on the shared subset of pairs in SimVerb-3500 (SV∩SpA_SVs) and scores obtained via the proposed arena-based method (SV∩SpA_SpAs). The two sets of results show very strong correlation (Spearman's $\rho = 0.86$), with near-perfect correlation ($\rho = 0.98$) between results obtained by static embeddings. This supports the hypothesis that it is indeed possible for humans to capture semantic similarity in their spatial arrangements, and SpA-Verb can be reliably used for comparing representation models, while offering a more comprehensive and challenging evaluation benchmark.

The analysis of correlation between model rankings on full SimVerb and SpA-Verb again produces a high correlation score ($\rho = 0.77$). These figures are enlightening when compared to similar analyses in previous work (Vulić et al., 2017a). While Vulić et al. (2017a) report very high correlations between model rankings on SimLex and SimVerb (>0.95), both of which measure semantic similarity, the scores are much lower between model rankings on SimLex or SimVerb and MEN (Bruni et al., 2014) (0.342 and 0.448, respectively), a dataset which captures broader conceptual relatedness. This suggests that SpA-Verb aligns with SimLex and SimVerb in its treatment of semantic similarity and relatedness, and that the spatial interface combined with instructions to arrange words based on the similarity of their meaning allows the annotators to capture word similarity as distinct from relatedness and association.

As results in Table 4.7 indicate, SpA-Verb is particularly challenging for models learning from co-occurrence information (however, incorporation of subword information helps performance, as seen in the scores obtained by the CBOW-CC model). Completely unrelated word pairings, which are easy to capture based on distributional data, are removed in the first phase, leaving only fine-grained semantic distinctions between related concepts.

4.8.4 Evaluation on Highly Associated Pairs and on High-IAA Classes

One of the main challenges for distributional models is to tease apart associated *and* similar words from those that are highly associated and frequently co-occurring but dissimilar (e.g., *cook – bake* vs. *cook – eat*). As the Phase 2 focuses on similarity of meaning disregarding association, it is possible to subsample all the highly associated pairs – both similar and dissimilar – to create an evaluation sample specifically testing the models' capacity to recognise this distinction. Following the evaluation on the highly associated set, I examine whether, in turn, the semantic classes which proved easier for annotators to reason about are also less challenging for models by focusing the evaluation on classes with the strongest annotator consensus.

Highly Associated Pairs

I evaluate the selection of models on the top most associated quartile of pairs (according to their USF association score, as in Hill et al. (2015)), in comparison with their performance on the entirety of SpA-Verb. This also allows for further investigating the nature of the semantic relations captured by the arena-based judgments and the proposed design's capacity to produce similarity ratings unbiased by association, despite not using an explicitly defined rating scale. As has been observed for datasets capturing similarity as distinct from association (e.g., SimLex), one could expect the performance of models learning solely from distributional information to be negatively affected, as strong co-occurrence evidence for the highly associated pairs has been shown to cause systems to overestimate word similarity (Hill et al., 2015). Table 4.8 presents the results of this analysis. As predicted, we can see a performance drop for the SGNS-BOW2 model, the subword-informed CBOW-CC vectors, as well as the three BERT models, both the *in isolation* and *in context* variants (the latter again proving more robust than the former). The remaining systems improve, with linguistically-informed Non-distributional and SGNS+Attract-Repel models performing noticeably better on these difficult cases than on the entire dataset. Notably, the consistently strong performance of representations drawing on lexicon information shows that human judgments collected in the arena-based task correlate with the expert knowledge coded in manually crafted linguistic resources.

Model	**SpA-Verb**	**Top Q**	**#3**	**#1**	**#13**
SGNS-BOW2	0.179	0.069	0.082	0.045	0.108
CBOW-CC	0.271	0.216	0.084	0.322	0.378
ELMo-Static	0.230	0.237	0.098	0.114	0.238
Non-distributional	0.295	0.450	0.386	**0.359**	0.480
SGNS+Attract-Repel	**0.385**	**0.526**	**0.718**	0.338	**0.534**
BOW2-VN	0.176	0.205	-0.037	0.198	0.212
BOW2-FN	0.210	0.287	0.095	0.242	0.156
BERT-BASE (iso)	0.235	0.181	0.123	0.201	0.330
context-10	0.262	0.181	0.154	0.177	0.366
context-100	0.265	0.183	0.149	0.193	0.380
context-500	0.265	0.185	0.148	0.190	0.379
BERT-LARGE (iso)	0.224	0.152	0.164	0.245	0.246
context-10	0.255	0.163	0.180	0.242	0.318
context-100	0.256	0.167	0.172	0.245	0.326
context-500	0.256	0.167	0.171	0.245	0.326
BERT-LARGE-WWM (iso)	0.237	0.195	0.170	0.215	0.276

Table 4.8 Evaluation on the top quartile of most associated pairs in SpA-Verb (**Top Q**), compared against Spearman's correlation scores on the whole dataset (**SpA-Verb**), and on the top 3 classes with the highest IAA ($\rho > 0.50$) (**#3,#1,#13**).

High-IAA Classes

Having evaluated how well the different representation models cope with the difficult subset of highly associated pairs, it is interesting to examine whether high human agreement on certain verb classes correlates with model performance, that is, to what extent the classes which were easier for human annotators to judge prove to be an easier benchmark for distributional models. Table 4.8 presents the results of the evaluation on the three Phase 1 classes with highest IAA ($\rho > 0.50$).

While most models improve on class #13 (verbs of sound) compared to the entire dataset, the results do not show consistent performance gains for most models. Only the Non-distributional embeddings improve across the three classes, whereas the SGNS+Attract-Repel model records the largest gain scoring a $\rho = 0.718$ on the highest IAA class #3. Notably, all models except these two record a performance drop on the same class. This is interesting considering the nature of the class, which, as illustrated in Figure 4.7, contains many verbs with opposite polarity (i.e., negative and positive rate of change), forming pairs of synonyms and antonyms. Vulić and Korhonen (2018)'s word vector space specialisation model is designed precisely to allow fine-tuning of distributional vector spaces to distinguish between synonymy and antonymy, making use of linguistic constraints derived from external resources that specify the exact lexical

relation between a pair of words. This also explains the very low correlation scores achieved by FrameNet and VerbNet-specialised models: Both of these resources group antonymous rate of change verbs together, due to their shared syntactic behaviour and high semantic overlap along all meaning dimensions but one, that of polarity or direction of change. Making this distinction is perennially difficult for statistical models learning purely from distributional information, since antonyms and synonyms have similar co-occurrence patterns in corpora; BERT embeddings prove the strongest in this category, outperforming the BOW models and static ELMo vectors on this difficult class. Crucially, the high correlation between the SGNS+Attract-Repel model and the human judgments suggests that the proposed approach enables capturing these important, cognitively salient semantic relations between the otherwise related items, and holds promise for more fine-grained linguistic analyses.

4.8.5 Evaluation on Semantically Focused Subsets

Typological study of the regularities in the way conceptual components are encoded in lexical items, that is, lexicalisation patterns, groups languages into types based on the lexicalisation strategies they permit. As far as verbs are concerned, cross-linguistic differences regard, for instance, the elements which are encoded in or outside the verb. The strategy characteristic for English directed motion verbs is to conflate the semantic elements of "Motion" and "Manner" inside the verb, and express "Path" outside (e.g., *The tennis ball rolled down the slide*) (as opposed to, for example, Italian, where "Motion" and "Path" are encoded together in the verbal root, and "Manner" may be expressed as a gerundive adjunct; see Chapter 5, Section 5.4.2) (Folli and Ramchand, 2005; Talmy, 1985). The preference for a certain lexicalisation pattern impacts the verb inventory of a given language: English and other languages where the first pattern is typical tend to have large repertoires of verbs expressing motion occurring in various manners. This is reflected in the most numerous class in the SpA-Verb dataset, which includes 100 motion verbs, many of which make subtle distinctions regarding the way in which an action is performed. However, this phenomenon is not restricted to verbs of motion. Languages which display a preference for lexicalising manner of motion often possess an extensive inventory of verbs expressing manner in general, for example, manner of speaking, manner of looking (Majewska et al., 2018b). The vast coverage of the proposed resource allows for zooming into these densely populated meaning domains to examine the capacity of representation models to capture the subtle meaning distinctions in the manner in which an action described by a verb is performed.

Model	**Motion**	**Heat**	**Sound**	**Emotion**	**Pain**
SGNS-BOW2	0.247	0.078	0.186	0.160	0.023
CBOW-CC	0.275	0.534	0.416	0.265	0.277
ELMo-Static	0.300	0.232	0.301	0.317	0.003
NON-DISTRIBUTIONAL	0.341	0.631	0.549	0.232	0.224
SGNS+ATTRACT-REPEL	**0.410**	0.374	**0.588**	**0.359**	0.445
BOW2-VN	0.327	**0.787**	0.201	0.294	0.075
BOW2-FN	0.368	0.393	0.419	0.264	**0.465**
BERT-BASE (ISO)	0.262	0.138	0.230	0.201	0.185
CONTEXT-10	0.239	-.125	0.283	0.336	0.243
CONTEXT-100	0.239	-.081	0.290	0.330	0.253
CONTEXT-500	0.238	-.081	0.289	0.331	0.255
BERT-LARGE (ISO)	0.254	0.011	0.176	0.248	0.280
CONTEXT-10	0.258	0.179	0.272	0.333	0.337
CONTEXT-100	0.256	0.188	0.279	0.329	0.347
CONTEXT-500	0.256	0.188	0.280	0.332	0.348
BERT-LARGE-WWM (ISO)	0.283	0.339	0.197	0.242	0.260

Table 4.9 Evaluation of representation models on subsets of SpA-Verb verb pairs focused on particular semantic domains. All scores are Spearman's ρ.

I evaluate the proposed selection of models on five verb pair sets, comprising verbs belonging to specific meaning domains. Included are motion verbs (Class 10, 4950 pairs), as well as four subsets of pairs defined in terms of participation in FrameNet frames: verbs related to heat (*Absorb_heat*, *Apply_heat*; 46 pairs), experiencing emotions (*Experiencer_obj*, *Experiencer_focus*, *Cause_to_experience*, *Feeling*; 237 pairs), producing sound (*Cause_to_make_noise*, *Communication_noise*, *Make_noise*, *Motion_noise*, *Sound_movement*; 235 pairs), and causing or experiencing pain (*Cause_harm*, *Experience_bodily_harm*, *Perception_body*, *Cause_bodily_experience*; 219 pairs). Table 4.9 summarises the results. Across the domains, we can see best performance from the linguistically-informed representations and consistently strong performance from SGNS+ATTRACT-REPEL. Moreover, the patterns of correlation scores obtained by the two sets of vectors specialised for VerbNet (BOW2-VN) and FrameNet (BOW2-FN) provide some more evidence regarding where these two lexical resources and the SpA-Verb dataset align and diverge in terms of the organisation of the same concepts. The BOW2-VN model achieves by far the highest result on the Heat subset (0.787), but performs very poorly on the Pain subset (0.075), where, in turn, the FrameNet-specialised model leads (0.465). In the meaning domain related to causing and experiencing pain, the annotators' judgments align more closely with the distinctions captured by the different FrameNet frames (i.e., causing harm vs. experiencing

harm vs. experiencing a (non-harmful) bodily sensation vs. causing a non-harmful bodily experience) than they do with the semantic-syntactic classes containing the same verbs in VerbNet. However, where FrameNet frames are broader, as it is the case in the Heat domain, we can see the VerbNet-specialised vectors achieving better performance: For example, while the FrameNet *Apply_heat* frame groups verbs such as *cook*, *boil*, and *melt* together, VerbNet divides them into two classes, 'cooking-45.3' and 'other_cos (change of state)-45.4', which aligns more closely with the SpA-Verb annotators' judgments (*cook* – *boil* have a small score 0.238, while *melt* – *cook* 0.797 and *melt* – *boil* 0.885, reflecting the greater distances between these concepts).

Examination of the scores recorded for the three BERT model variants *in isolation* reveals the large model with whole word masking (BERT-LARGE-WWM) to be the most robust across the different semantic areas, and this advantage is particularly visible on the smallest set of Heat verbs. Notwithstanding the subpar performance of contextualised BERT-BASE on this semantic domain (where fluctuations in scores are more likely given the very small size of this set), the benefits of computing BERT embeddings *in context* are again clear: For instance, contextualised BERT-BASE and BERT-LARGE embeddings are especially competitive on Emotion verbs, coming a close second behind SGNS+Attract-Repel.

The scale of the SpA-Verb dataset allows for zooming in on word subsets with desired characteristics and creating smaller datasets controlling for some specific feature, showing potential for more focused analyses of representation models. Similar analyses could shed light on particular strengths and weaknesses of representation architectures and help identify meaning domains and semantic properties requiring systems to be more specialised, or different modeling strategies altogether.

4.8.6 Further Discussion

In the preceding sections, I examined how well the lexical representations derived from the selected models correlate with human judgments collected through spatial arrangements. The experiments revealed that the semantic distinctions which are easier for humans to make often elude representation models, and that discriminating between similar and highly associated but dissimilar words remains a challenge for most systems. Moreover, the results showed that model performance varies across different semantic classes, revealing inconsistencies in representation quality for verbs belonging to different domains. The results have also revealed interesting patterns which open up further questions concerning the implications of evaluating representation architectures

on SpA-Verb (and lexical semantic similarity datasets in general), which I address below.

SpA-Verb vs. SimVerb

In Sections 4.6 and 4.7, I examined how the differences in two data collection methodologies, pairwise ratings and spatial arrangements, affect the characteristics of the produced datasets. In turn, the evaluation results in Table 4.7 showed that these differences translate to the datasets' respective difficulty: The consistently lower performance on SpA-Verb with respect to SimVerb suggests the former is a more challenging benchmark. In light of these differences, an important question which arises is: Are the insights into the intrinsic quality of lexical representations which each of these datasets provides fundamentally different or rather complementary? And if SpA-Verb offers larger, more comprehensive coverage and higher granularity with respect to SimVerb, is there some important signal that SimVerb provides that is missing from SpA-Verb?

As discussed in Section 4.6, an important difference between the two datasets concerns unrelated verb pairs (e.g., *broil – respect, bounce – prohibit*). The proposed annotation design filters out the completely unrelated pairs in Phase 1, where verbs are grouped based on shared semantics and relatedness. The main motivation behind this choice was to elicit similarity judgments on comparable concepts (the importance of which has been emphasised in psychology (Turner et al., 1987)), while avoiding the conflation of scores for unrelated words and antonyms, which are related but dissimilar (a phenomenon characteristic for SimVerb). However, a potential disadvantage of such a solution is the complete exclusion of unrelated pairs from SpA-Verb, which precludes direct evaluation of the capacity of a model to tackle such examples. In other words, a system could be overestimating the similarity of unrelated words and still score highly on SpA-Verb.

First of all, it is important to note that such a hypothetical scenario seems rather unlikely. Since distributional models learn about word meaning from co-occurrence patterns, the notion of (associative) relatedness, characterising words frequently appearing together in text, is the knowledge easiest to glean from raw data. SpA-Verb requires models to exhibit an understanding of lexical semantics that is much more advanced: Systems need to be sensitive to fine-grained meaning distinctions and degrees of similarity between related words. Crucially, it includes antonymous pairs, which pose a significant challenge to models learning from distributional information (Mrkšić et al., 2016; Vulić and Korhonen, 2018). Further, the high number of close similarity scores makes the task difficult, compared to the sparse, discrete ratings in SimVerb.

Viewed this way, the exclusion of unrelated pairs makes the dataset more challenging: The unrelated pairs are easiest to judge both for humans (i.e., pairs of words which have nothing in common are given 0 scores in SimVerb) and for distributional semantic models, based on the negative signal (i.e., no co-occurrence in text) that is easily derived from corpora. Since SpA-Verb does not reward models on these easy cases, the absolute scores are unsurprisingly lower than on SimVerb. However, they may provide a more realistic measure of representation quality and the capacity of models to reason about lexical semantics. Conversely, the fact that the unrelated pairs are most numerous in SimVerb (i.e., the 0-1 score interval in Figure 4.8) may result in an undesirable inflation of the estimate of representation quality. To sum up, while both SimVerb and SpA-Verb reward models capable of distinguishing between similar and related-but-dissimilar concepts, SpA-Verb requires models to make many nuanced, fine-grained distinctions between the *degrees* of similarity and dissimilarity of related, semantically proximate words, in a densely populated semantic space.

Notwithstanding the potential benefits of not including unrelated pairs in the dataset, researchers may be explicitly interested in examining the models' effectiveness in dealing with such cases. While using the subset of unrelated pairs from SimVerb is the obvious choice, unrelated examples are easily obtainable from SpA-Verb as well, based on the output of the semantic clustering in Phase 1. Since annotators group verbs into theme classes based on their semantics, the unrelated words are separated into different clusters. Unrelated pairs can therefore be easily generated by randomly sampling pairs from different Phase 1 classes.

Static Word Embeddings vs. Pretrained Encoders

Another noteworthy pattern which emerges from the evaluation experiments is the relatively weaker performance of the BERT models compared to the stronger static representations, despite the proven superiority of Transformer-based architectures across diverse NLP tasks. Since intrinsic tasks such as word similarity prediction serve as a proxy for estimating model performance in downstream applications, the under-performing BERT embeddings raise questions: What makes the contextualised representations less successful on SpA-Verb, and what is it telling us about the lexical semantic signal they encode?

The first important factor responsible for this phenomenon lies in the fundamental difference between static word embeddings and the representations produced by BERT. While the former assign a single, fixed vector to a given word, the latter encode meaning dynamically, in context. In order to derive comparable lexical representations from

BERT, it is necessary to abstract away from individual word occurrences and aggregate token embeddings into a single, static type-level representation. A number of solutions have been proposed to this end (Cao et al., 2020; Conneau et al., 2020b; Liu et al., 2019b, *inter alia*). In this chapter, I experimented with two different approaches, one where the target word is fed into the model in isolation, and one where it appears in N full sentences and the type-level embedding is derived by averaging over each such token-level representation. Each of these variants, however, requires careful tuning of the configuration of the following parameters: (i) the choice of hidden representations to average over; (ii) the selection of special tokens (i.e., `[SEP]` and `[CLS]` tokens in BERT) to include in the subword representation averaging step. While the results presented in this chapter are achieved by BERT representations yielded by the strongest such parameter configuration across the reported tasks and word pair sets, further investigations into alternative approaches to deriving lexical representations from pretrained models are needed to maximise their competitiveness.

What is worth noting, intrinsic word similarity datasets do not directly reward the capacity of Transformer-based models to capture word meaning in context. Indeed, the results in Table 4.7 show analogous patterns of model scores on SimVerb and on SpA-Verb, and for both the static embeddings drawing on external linguistic information outperform the embeddings derived from BERT by a significant margin. Notably, the scores achieved by all models on the shared pairs from both resources show strong correlation ($\rho = 0.86$), which indicates that the observed phenomenon is not an idiosyncrasy of the spatial similarity dataset, but is common to both resources. However, this characteristic does not preclude SpA-Verb's applicability as a discriminator of the quality of lexical representations yielded by pretrained encoders such as BERT. Many recent efforts focused on investigating *why* state-of-the-art pretrained models perform as well as they do in downstream applications, probing the linguistic knowledge captured by those architectures (Hewitt and Manning, 2019; Jawahar et al., 2019; Liu et al., 2019a; Tenney et al., 2019a). Importantly, it has been noted that their success may be due to learning shortcuts in NLP tasks, rather than being a direct product of the quality and richness of the encoded linguistic knowledge (Rogers et al., 2020b). Indeed, a number of works have drawn attention to BERT's reliance on shallow heuristics in natural language inference and reading comprehension (McCoy et al., 2019; Rogers et al., 2020a; Si et al., 2019; Sugawara et al., 2020; Zellers et al., 2019).

Since BERT takes subword tokens as input in pretraining, the question of whether and how it captures the *lexical* signal merits investigation. SpA-Verb meets this need as a lexical semantic probing tool, enabling direct evaluation of the quality of

the lexical knowledge stored in the parameters of BERT. The experiments on SpA-Verb and specific subsets of the dataset have already provided some insights. They revealed that the word-level embeddings obtained by averaging over N occurrences in context encode richer lexical semantic knowledge than those derived by feeding words into the pretrained model in isolation. Further, the experimentation with different configurations of lexical representation extraction parameters, such as the choice of hidden layers from which to derive the ultimate representation or the inclusion of special tokens, revealed it is advantageous to draw type-level verbal lexical knowledge from the first 6 layers, while the inclusion of special tokens degrades representation quality. Future experiments using the spatial similarity data may help scrutinise the nature and location of the lexical semantic signal encoded in these representations, and its contribution to downstream performance. Moreover, probing analyses using subsets of SpA-Verb targeting particular semantic domains may help uncover the areas where BERT's lexical representation quality is still insufficient and aid development of systems with a stronger grasp of verbs' lexical-semantic properties.

4.9 Conclusion

In this chapter, I presented and thoroughly analysed a new method for large-scale collection of semantic similarity data based on clustering and spatial arrangements of lexical items. The two phases of the proposed design produce semantic clusters and word pair scores within an integrated framework, and can be readily applied to other parts of speech and types of stimuli. The study yielded SpA-Verb, a dataset of fine-grained similarity scores for 29,721 unique verb pairs, together with 17 relatedness-based verb classes. The comparative analyses against FrameNet, VerbNet, and WordNet showed that the two-phase design allows humans to differentiate between a range of semantic relations and intuitively capture fine-grained linguistic distinctions pertaining to verb semantics through subtle relative judgments. The automatic clustering experiments run on top of the distance matrices from Phase 2 also demonstrated the potential of the presented design to yield semantic clusters within each broad class, which can be employed in future work to evaluate the capacity of models to create semantic classifications and taxonomies automatically. What is more, by yielding complete distance matrices for each class, the proposed design allows in-depth exploration of the dimensions underlying the organisation of the semantic space in question, holding promise to support cognitive linguistics research. Employing the methodology as a tool for probing the representation of verbal concepts in the mental lexicon and using these

insights to inform computational lexicon creation is an especially promising avenue for future research.

From the methodological standpoint, there are several challenges which future work building on the present approach could focus on. First, while the chosen verb sample provides broad coverage of diverse verb meanings, the ultimate goal would be to extend it further, to encompass an even greater proportion of the lexicon. To achieve this without increasing the cognitive load on the annotators, (semi-)automatic pre-processing methods (e.g., automatic word clustering) should be considered, which would permit partitioning large word samples into manageable sets prior to initiating manual clustering in Phase 1. Further, the analysis of Phase 2 data revealed that large word samples (80+ verbs) are still difficult to reliably arrange within the space provided in the current interface, especially during the first trial with the entire set. This, in turn, leads to lower inter-annotator agreement. One possible solution to this problem is modifying the current interface to allow zooming in on dense areas of the circular arena and refining the relative placements of items before moving on to the next trial. Future endeavours could also explore whether creating hierarchical clusters in Phase 1 and collecting Phase 2 judgments both on the broader higher-level classes and the narrower fine-grained subclasses would allow improving the accuracy of the recorded similarity scores for large semantic domains.

From the perspective of model evaluation, the extensive experiments using the data from both phases demonstrated the potential of the dataset to support probing analyses crucial for further developments in representation learning. The fact that SpA-Verb contains nuanced similarity judgments between semantically close verbs means that the resource provides a challenging benchmark for state-of-the-art systems, which will be useful in research aimed at improving the capacity of NLP models to represent the complex meaning of verbs and events they describe. Moreover, the large size of the dataset offers vast possibilities for robust analyses on different word subsets and semantically related classes, allowing for better informed tuning and comparison of the adequacy and potential of various representation learning architectures to capture fine-grained semantic distinctions present in the mental lexicon, while helping achieve greater model interpretability. In the next chapter, I extend these analyses beyond English and investigate the method's potential and portability to a diverse selection of languages.

Chapter 5

Verb Knowledge Acquisition for Multilingual Evaluation

5.1 Introduction

Many recent efforts in semantic modeling have focused on unsupervised pretraining to extend the benefits offered by recently proposed text encoders (Devlin et al., 2019) to new languages and domains. In these approaches, general language representations are learned from large volumes of unlabelled text, and subsequently leveraged in downstream systems by means of fine-tuning on a given supervised task. The release of large multilingual pretrained encoders (Conneau and Lample, 2019; Devlin et al., 2019) boosted the state of the art on a range of multilingual tasks (Artetxe et al., 2020; Hu et al., 2020; Kondratyuk and Straka, 2019; Mueller et al., 2020; Pires et al., 2019; Qiu et al., 2020; Wang et al., 2019b; Wu and Dredze, 2019). In parallel, the number of language-specific pretrained architectures available has also been steadily growing, with the advantage of being more attuned to the properties of the language in question (Nozza et al., 2020; Virtanen et al., 2019). The ease of incorporating these powerful encoders into downstream task pipelines has made them widely popular. However, there is a disproportionate shortage of resources allowing for probing of the learned representations in most languages. In this chapter, I extend the methodology introduced in the preceding chapter to a typologically diverse selection of languages in order to address this deficit. I first discuss the steps involved in multilingual data collection, highlighting the main language-specific challenges. Next, using cross-lingual mappings, I carry out analyses of cross-lingual overlap in the semantic classes created in Phase 1, as well as quantitative and qualitative comparisons of the semantic distance matrices from Phase 2. These investigations shed light on one of the questions explored

in this research, i.e., to what extent are the meaning components underlying the organisation of verbs in the lexicon cross-lingually shared? Further, I directly evaluate the extent to which semantically driven classes and clusters created in the two phases of the data collection process align with those semantically *and* syntactically informed, which allows me to assess the potential of the presented method to aid the construction of verb lexicons in languages lacking such resources. Subsequently, I perform evaluation of static and contextualised representation models on the tasks of lexical similarity and semantic clustering using the data from both phases. This allows for identification of models' strengths and shortcomings, as well as specific challenges posed by the languages' properties and different domains of verb meaning. The collected data, comprising semantic classes and fine-grained pairwise similarity scores for Chinese, Japanese, Finnish, Italian, and Polish, are released together with the English dataset presented in the previous chapter as a multilingual resource targeting verb semantics, Multi-SpA-Verb.[1]

5.2 Background and Design Motivation

Word similarity has been widely used as a go-to intrinsic evaluation task, in which rankings of similarity scores computed between word embeddings produced by representation models are compared against ranked human similarity judgments. The dataset design involving sets of word pairs and their associated rating on a discrete scale has been particularly common, due to its reliance on non-expert native-speaker judgments, quicker and cheaper to obtain than the large expert-curated lexical-semantic or semantic-syntactic resources such as WordNet (Fellbaum, 1998) or VerbNet (Kipper et al., 2006; Kipper Schuler, 2005). In English, examples include WordSim-353 (Agirre et al., 2009; Finkelstein et al., 2002), MEN (Bruni et al., 2014) and SimLex-999 (Hill et al., 2015). Analogous datasets have been created in other languages, either through translation from an existing English dataset (e.g., from SimLex: German, Italian, and Russian (Leviant and Reichart, 2015b), Hebrew and Croatian (Mrkšić et al., 2017) and Polish (Mykowiecka et al., 2018)), or from a new set of concept pairs (e.g., Turkish (Ercan and Yıldız, 2018), Mandarin Chinese (Huang et al., 2019), Japanese (Sakaizawa and Komachi, 2018)). While these datasets are dominated by nouns (e.g., SimLex includes 222 verb pairs), verb-oriented datasets are harder to come by. In English, these include datasets of Yang and Powers (2006) (130 verb pairs), Baker et al. (2014)

[1]The multilingual dataset has been introduced in Majewska et al. (2020b) and is available at https://github.com/om304/Multi-SpA-Verb.

Language	ID	N verbs	N classes	N pairs	THR pairs
Mandarin Chinese	ZH	771	17	23,990	1,898
Finnish	FI	761	16	28,641	10,065
Italian	IT	817	17	24,747	6,436
Japanese	JA	704	17	22,915	7,916
Polish	PL	850	18	28,895	6,735

Table 5.1 Data statistics including the number of unique verbs in each sample (translated from English) (**N verbs**), the number of Phase 1 classes (**N classes**), the total number of pairwise scores in the final dataset (**N pairs**) and the thresholded subset of each dataset (**THR pairs**) (See §5.5.2).

(143 verb pairs), and Gerz et al. (2016) (3,500 verb pairs). A recent multilingual word similarity dataset, Multi-SimLex (Vulić et al., 2020a), extends coverage of verb semantic similarity to 469 verb pairs in 12 languages, including Mandarin Chinese, Finnish, and Polish. The large-scale English verb resource presented in Chapter 4 (SpA-Verb) comprises verb classes and unmatched coverage of nearly 30k verb similarity scores. In what follows, I demonstrate that the same dataset creation methodology based on spatial arrangement (SpAM) can be extended to other and typologically diverse languages such as Mandarin Chinese, Japanese, Finnish, Polish, and Italian. For each language, I create a dataset comprising 16-18 verb classes with similarity scores between all class members, resulting in over 20k such scores in each language (Table 5.1).

I start from the English SpA-Verb sample translated into five target languages and apply the two-phase annotation method combining semantic clustering and spatial arrangements based on semantic similarity outlined in Chapter 4. In Phase 1, a large word sample is divided into a number of broad categories of similar and related items. Each of these classes is then used as input in Phase 2, where the related class members are arranged in a 2D space based on their semantic similarity. Each item placement simultaneously communicates its semantic distance to all other items present and the inter-stimulus Euclidean distances represent pairwise dissimilarities between words in the sample. The final representational dissimilarity matrix (RDM) estimate is produced by statistically combining the evidence from multiple subsequent 2D arrangements and contains a dissimilarity estimate for each pairing of words in the set (see Chapter 4 and Kriegeskorte and Mur (2012) for the details). The dissimilarities collected for each Phase 1 class are then normalised to ensure inter-class consistency in the final dataset.

The precursor work on English (Chapter 4) showed that inducing similarity judgments through spatial arrangements contributes to producing nuanced similarity scores

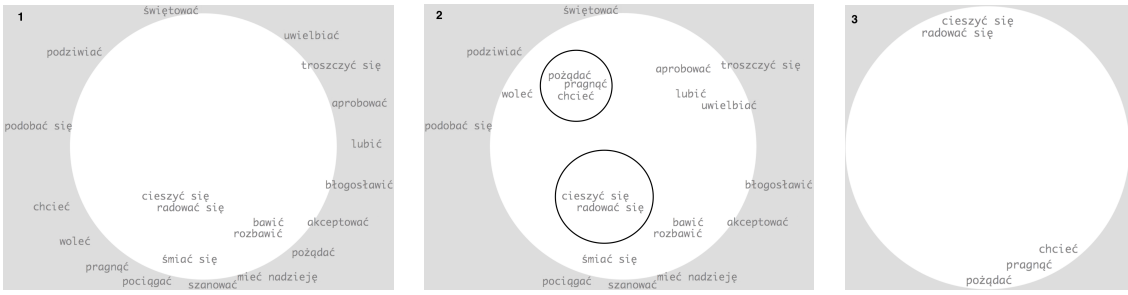


Fig. 5.1 Consecutive Phase 2 trials on a class of Polish emotion verbs. In the first trial (1-2), the whole class is displayed around the arena and word labels are placed one by one based on the similarity of their meaning. Words put closer together in the first trial (2) are subsampled for the subsequent trial (3), and arranged again in a less crowded space (annotators are asked to use the entire space available in each trial and the relative inter-item distances represent the dissimilarities).

that capture a range of fine-grained semantic relations (e.g., antonymy, troponymy or hypernymy), encoded in the pairwise distance matrices available for all items in a sample. What is more, while discrete scales require the annotator to quantify the degree of similarity between two words at a time, the repeated spatial arrangements on different configurations of the same items elicit simultaneous multi-way judgments, which is both efficient and avoids the problem of judging word pair similarity in isolation.

Importantly, the two-phase design offers a practical advantage for porting the method to other languages. The approach starts from a verb sample, rather than a set of word pairs, which allows easy translation into the target language, avoiding many of the complications encountered in translation of pairs. These include cases where both words in the source language pair translate into the same word (e.g., *cup* – *mug* → Italian *tazza* – *tazza*), or several pairs in the source language translate into identical target pairs (e.g., *easy* – *hard*, *easy* – *difficult* → Polish *łatwy* – *trudny*). By translating on a word-by-word basis, problems related purely to the pairwise design are completely avoided. Each unique source word receives its best target translation, unless no equivalent exists. Conversely, if a source word translates into several equally adequate target words, all candidates are included, and thus shortages in one lexical area are compensated in another avoiding major reduction in dataset size.

5.3 Data Collection

The languages selected for this study were sampled from 5 different language families to ensure typological diversity: Sino-Tibetan (Mandarin Chinese ZH), Japonic (Japanese JA), Uralic (Finnish FI), Slavic (Polish PL) and Romance (Italian IT). Following

translation from English (EN), the two data collection phases were set up on an online platform (`meadows-research.com`) as two separate studies for each language. Recruitment was carried out on a crowdsourcing website, `prolific.co`. Participants were native speakers of the target language with at least undergraduate education level. To ensure high quality of the data obtained through crowdsourcing, I also restricted the eligible participant pool to users with at least a 90% approval rating (a screening option offered by the website). Each phase featured a short qualification task simulating the main experiment, which tested the participants' understanding of the guidelines (Appendix C.1), embedded in the task interface and accessible throughout the study.

5.3.1 Word Sample Translation

Translation was carried out by one native speaker translator per language. The translators were all fluent in English, but could use external resources (e.g., dictionaries) as an additional aid. In case several equally suitable candidates were identified for one source word, all of them were kept. This was especially true for polysemous English verbs which translated to more than one target verb, each expressing a distinct sense of the source word (e.g., *bear* → Finnish (1) *kantaa*, 'carry', (2) *sietää*, 'endure'). On the other hand, if two English words had only one adequate translation equivalent, the two-to-one mapping was kept where unavoidable (e.g., *restrict, limit* → Mandarin Chinese 限制, *xiànzhì*). This flexible approach helped avoid unduly adjusting the translations in the target language to English semantics and consequently prevented from perpetuating the source language's biases into the resultant data. Table 5.1 shows the number of unique verbs in the final target sample for each language. While the translation stage produced sizeable target samples in all languages (with the biggest decrease in number of unique items observed in Japanese, from 825 to 704), there were several areas which required particular design choices, which I discuss below.

Multi-word Expressions

Multi-word expressions have often been avoided in word similarity dataset creation (e.g., Camacho-Collados et al., 2015b; Ercan and Yıldız, 2018; Leviant and Reichart, 2015b). for the sake of evaluation simplicity. They have long been considered challenging for NLP applications due to their unpredictable semantics, which often eludes compositional interpretation (Sailer and Markantonatou, 2018). However, given how widespread the phenomenon is cross-linguistically, I chose not to exclude them. In the present language sample, cases of multi-word translations concern light verb constructions

(where the light verb contributes little to no semantics), phrasal verbs and idiomatic expressions (e.g., Finnish *olla varuillaan*, 'beware', *tehdä yhteistyötä*, 'cooperate', Polish *mieć nadzieję*, 'hope', Italian *dare un'occhiata*, 'glance'). These were most common in Finnish (63 out of 762 entries in the final sample). A particular case of a multi-word unit prominent in Polish are reflexive verbs (see discussion below), composed of the main verb and a reflexive pronoun *się* (99 reflexives in a 850-word sample). In Section 5.5, I describe how I handled multi-word expressions in semantic model evaluation.

Intransitive and Transitive Variants

A particular challenge was posed by English verbs which can be used either transitively or intransitively, e.g., *bow*, *shrink*. In Finnish, Italian, and Polish, these cases often require providing two translation variants: e.g., English *decrease* → Finnish *vähentää* (TR) and *vähentyä* (INTR), Polish *zmniejszyć* (TR) and *zmniejszyć się* (INTR), Italian *ridurre* (TR), *ridursi* (INTR). While these pairs are usually single dictionary entries, sometimes one of the variants has senses not captured by its counterpart. In order to include these additional senses without excessively increasing the target sample, I kept the two variants wherever they captured the source verb's ambiguity and important meaning distinctions. Otherwise, the translators chose the variant best corresponding to the meaning of the source verb: e.g., in Italian, the verb *hurry* was translated to the reflexive *sbrigarsi* (the non-reflexive *sbrigare* translates to 'attend to, deal with (e.g., a task)'). In Finnish, for the verb *decrease* only the transitive variant *vähentää* was kept; whereas in Polish, for *impose* both the reflexive and non-reflexive variants were kept, since, similarly to English, each captures a distinct sense: *narzucać* – 'to force someone to accept something (e.g., a belief)', or *narzucać się* – 'to cause inconvenience to someone by demanding their attention'. It is worth mentioning that the dominant variant is not the same across languages. For instance, while in Italian and Polish verbs which can participate in the causative-inchoative alternation (e.g., *open*: *He opened the door. – The door opened.*) have the basic transitive form (with the inchoative verb bearing reflexive marking – *otworzyć się, aprirsi*), in Japanese, the corresponding verbs are predominantly lexicalised in the non-causative type (Talmy, 1985), which is reflected in the translations (e.g., 溶ける *tokeru* 'melt' (INTR), 減少する *genshō suru* 'decrease' (INTR), 縮む *chidjimu* 'shrink' (INTR); with an exception for 'increase', 増す *masu* (INTR), 増やす *fuyasu* (TR)). A special case involves the so-called double inchoatives in Polish, i.e., equipollent pairs of reflexive and non-reflexive inchoative verbs (e.g. 'drown', *utopić się – utonąć*). In this work, the transitive and intransitive sense of the verb 'drown' were translated into the transitive reflexivising variant (*utopić*, without

the pronoun) and the non-reflexive intransitive counterpart (*utonąć*), respectively, which allowed for reducing the number of multi-word entries (*utopić się*).

Aspect

Another phenomenon potentially complicating the translation process is verbal aspect. In Polish, all verbs have two variants, perfective and imperfective, expressing a completed or ongoing action (e.g., *pozdrowić* PFV – *pozdrawiać* IPFV, 'greet'). The aspectual distinction is either lexicalised (*brać* – *wziąć*, 'take') or expressed through affixes (*malować* – *namalować*, 'paint', *zdać* – *zdawać*, 'pass (an exam)'). Due to its relatively low regularity (with respect to other verbal morphological categories), the status of aspect has eluded clear categorisation: it has been treated either as inflectional, lexical (word-formational), or a borderline phenomenon between both (Smereczniak, 2018). In the present work, I took on a moderate approach similar to the above discussed case of reflexives. In order to avoid doubling the sample in size to account for both aspectual variants for each verb, both variants were listed only if they captured the polysemy of the source verb, otherwise, the unmarked variant was kept (e.g., the root form, rather than the affixed form), or else, the translator arbitrarily decided on the suitable translation (e.g., *uprowadzić* PFV, 'abduct', capturing the temporal boundedness of the action, but *nadużywać* IPFV, 'abuse', emphasising the habitualness of the action). Aspectual considerations also surfaced in Finnish, where repeated actions are expressed by frequentative aspect: e.g., 'jump' can be translated to *hypätä* (jump once) or *hyppiä* (jump repeatedly). The availability of such variants helped capture the proliferation of English manner of motion verbs: these variants were used as translations for synonymous English verbs, e.g. *bounce* – *hyppiä*, *jump* – *hypätä*. Overall, minimal aspectual oppositions (e.g., pairings of aspectual variants) are very few in the final datasets (e.g., Polish *biec* 'run' and *biegać* 'jog'). To explicitly study the impact of aspect on human judgments of lexical semantic similarity, in future work it would be interesting to apply the semantic multi-arrangement approach to an aspect-focused item sample, including all aspectual variants of a set of verbs.

5.3.2 Phase 1: Semantic Clustering

Five native speakers per language independently performed a rough clustering of the initial verb sample into broad semantic classes. Users dragged words one by one from a queue and placed them in circles representing broad semantic groupings (Figure 5.2). The participants were instructed to create groupings of similar and related verbs, each

Fig. 5.2 Finnish Phase 1 task interface (zoomed in; the label font is enlarged).

	ZH	JP	PL	FI	IT
B-cubed F-score	0.251	0.267	0.291	0.234	0.319

Table 5.2 Average pairwise B-cubed F-score calculated between individual clusterings across annotators within each language.

containing roughly 30-50 words. This rule of thumb, applied in the English precursor study, ensures similar granularity across languages.[2] To ensure annotation quality, the produced classifications were manually reviewed to identify rogue annotators and low-effort responses (e.g., multiple consecutive words in the queue were placed in the same class indiscriminately or large numbers of words were placed in the Trash circle and missing from the final classification), which were subsequently discarded (8% of submissions). Table 5.2 reports average pairwise B-Cubed overlap scores across all pairings of clusterings within each language. The final sets of classes for Phase 2 were produced in each language by first identifying the overlap in Phase 1 classifications, which determined the class structure and broad semantics of each class (e.g., movement, emotion, communication), and then populating the classes based on majority decisions (see Chapter 4, Section 4.4). Finally, for each language, the cross-subject classes were reviewed manually by a native-speaker adjudicator; In the process, the verbs missing from the intersection of individual clusterings were added to valid classes of related verbs (based on the criterion of semantic similarity and relatedness, ensuring semantic coherence of the resultant classes). Phase 1 produced 16-18 classes in each language (Table 5.1) and took between approx. 2.5 (Finnish) and 3.5 hours (Mandarin Chinese) to complete (Table 5.3).

[2]Note that the English sample's clusterability into broad semantic classes is a consequence of the original sampling criteria from across VerbNet classes (Chapter 4).

Language	Phase 1	Phase 2
Mandarin Chinese	199	815
Finnish	139	830
Italian	143	786
Japanese	168	780
Polish	147	870
Average	**158**	**816**

Table 5.3 Average total time (mins) spent on the completion of each phase (e.g., arrangement of all Phase 2 classes).

5.3.3 Phase 2: Similarity Multi-Arrangement

The classes from Phase 1 were fed into Phase 2, divided into 5-6 batches of 3-4 classes each. Verbs from one class were annotated independently from all others. Annotators were instructed to arrange presented verbs in a circular arena based on similarity of their meaning, disregarding similarity of sound, letters or simple association. For each batch, the aim was to obtain at least 5 valid sets of annotations and recruitment continued until this condition was satisfied. I employed the same post-processing quality assurance protocol as outlined in Chapter 4. First, I filtered annotators who performed the first arrangement too quickly (i.e., averaging less than 1 second per word placement upon first seeing the sample; 4% of submissions); next, for each class, I filtered out annotators for whom the average pairwise Spearman's correlation of arena dissimilarities with those of all other annotators was more than one standard deviation below the mean of all such average correlations (as done by Hill et al. (2015); 17% of submissions). To produce the final dataset and ensure consistency between differently sized classes, I calculated the average of the Euclidean distances from all accepted annotators for each verb pair and then normalised them, using the approach from previous work (Kriegeskorte and Mur, 2012) where each dissimilarity matrix is scaled to have a root mean square (RMS) of 1 (Eq. 4.9). Table 5.3 summarises the total time spent on the completion of the task for all classes in a given language.

5.4 Data Analysis

5.4.1 Phase 1: Cross-lingual Comparison

Inspection of the classifications emerging from the manual clustering task provides the first set of insights regarding the categorisation of verbal meaning in each of the languages, as well as the different clustering strategies and class membership criteria

ID	Class	Chinese		Japanese		Polish		Finnish		Italian	
#	Label	size	ρ	size	ρ	size	ρ	size	ρ	size	ρ
1	emotion	26	0.38	54	0.66	21	0.36	37	0.57	41	0.63
2	cooking	30	0.39	22	0.53	54	0.30	48	0.42	34	0.39
3	possession	30	0.39	49	0.13	46	0.35	36	0.14	38	0.28
4	motion (S)	74	0.13	76	0.18	92	0.13	87	0.18	86	0.14
5	motion (A/P)	66	0.09	39	0.46	88	0.16	82	0.13	82	0.13
6	sensory perception	32	0.28	-	-	32	0.36	↓	↓	38	0.40
7	physiology	52	0.24	53	0.25	55	0.20	64	0.23	49	0.39
8	state of being	↑	↑	24	0.43	↑	↑	↓	↓	↑	↑
9	change	38	0.44	45	0.35	29	0.47	47	0.26	23	0.58
10	cognition	44	0.24	58	0.29	79	0.17	61	0.18	62	0.24
11	physical contact	47	0.24	↓	↓	55	0.34	37	0.31	44	0.45
12	violence	↑	↑	84	0.31	↑	↑	36	0.40	↑	↑
13	law/crime	75	0.04	↑	↑	68	0.20	69	0.23	73	0.23
14	negative interaction	73	0.07	-	-	50	0.20	50	0.36	↑	↑
15	interaction	69	0.18	60	0.21	49	0.30	79	0.35	69	0.28
16	work/organisation	74	0.07	64	0.24	67	0.19	71	0.21	58	0.38
17	handicraft	51	0.11	63	0.23	60	0.14	↑	↑	71	0.14
18	destruction	39	0.24	46	0.22	39	0.20	52	0.21	26	0.39
19	sound	48	0.13	28	0.32	32	0.34	74	0.33	27	0.25
20	communication	↑	↑	52	0.26	55	0.23	↑	↑	44	0.29
21	combining	-	-	29	0.50	-	-	-	-	-	-

Table 5.4 Semantic classes produced in Phase 1, aligned cross-lingually based on member overlap (*size* = number of verbs in class, ρ = Spearman's IAA); English labels serve to identify broad semantic categories. ↑/↓ indicate a category is subsumed by the one above or below. S/A/P labels signal arguments typically selected by class members (agent-like (A), patient-like (P), or sole argument of an intransitive verb (S)).

adopted. Table 5.4 summarises Phase 1 output. Given the similar granularity of classifications, I aligned the monolingual classes with most overlap (via English mappings) for an easier comparison and assigned descriptive English labels corresponding to the shared broad semantics of the aligned classes.

Although there is cross-lingual variation in class size, we can observe a lot of high-level category overlap. Most broad semantic themes are recognised in all languages, either as independent classes or parts of larger categories. Only one category is specific to a single language, i.e., 'combining' in Japanese. In turn, two of the themes found in other languages ('sensory perception' and 'negative interaction') were disregarded as criteria for establishing separate groupings in the Japanese study. Verbs describing emotional states, movement, and cognitive processes consistently form separate groupings, as well as verbs belonging to the 'cooking' domain and verbs of possession.

Interesting patterns emerge from the inspection of merged classes and category boundaries, shedding light on the salient components of meaning taken into consideration in the categorisation process. For instance, in most languages physical contact verbs of varying degree of intensity and opposing sentiment form one big grouping (e.g.,

	EN	ZH	JP	PL	FI	IT
EN		0.56	0.28	0.63	0.61	0.75
ZH	0.43		0.56	0.53	0.54	0.49
JP	0.36	0.30		0.37	0.37	0.25
PL	0.65	0.40	0.33		0.68	0.74
FI	0.42	0.34	0.33	0.40		0.61
IT	0.60	0.40	0.34	0.56	0.40	

Fig. 5.3 Pairwise overlap in Phase 1 for all language pairs (B-Cubed F-scores) (**lower triangle**), with respect to the proportion of shared WALS typological features (**upper triangle**).

touch, cuddle, suffocate, hit). Whereas in Finnish, two separate clusters are formed, one with verbs of neutral and positive physical contact (e.g., *kutittaa* 'tickle', *hieroa* 'rub', *halata* 'hug'), and one with verbs of violent physical contact (e.g., *lyödä* 'hit', *hakata* 'beat', *lakkoilla* 'strike'), suggesting violence or intensity of the action described by the verb served as discriminators of verb meaning. Positive and negative sentiment are predominantly taken into account as class membership criteria in the domain of human interaction, where verbs describing negative and positive behaviours consistently form separate classes. Another interesting pattern concerns verbs describing states of being (e.g., *be, exist*), which are thought of as physiological process in Chinese, Polish, and Italian, while in Finnish they are grouped together with verbs such as *begin, originate, finish*, viewed as describing intermediary states between dynamic processes of origination and termination of existence. Furthermore, Japanese is the only language in the sample where attention is paid to the aspect of joining and adding, resulting in a separate class (e.g., 加わる *kuwawaru* 'join', 加える *kuwaeru* 'add', 関連づける *kanrendzukeru* 'connect, associate', 結合する *ketsugō suru* 'combine').

In order to quantitatively measure the degree of cross-lingual alignment, I calculate pairwise item-level overlap using the B-Cubed metric (Amigó et al., 2009; Jurgens and Klapaftis, 2013) between all language pairs, also including the English classes from the study presented in Chapter 4. I confront class overlap with the degree of typological affinity between language pairings, quantified as overlap in syntactic, morphological and lexical typological features from the WALS database (Dryer and Haspelmath, 2013) (Figure 5.3).[3] The strongest pairwise class alignment is found between the three Indo-European languages in the selection: English, Polish, and Italian. Further,

[3]Feature overlap is a proportion of shared feature values (see Appendix D.3 for a full list of typological features considered).

Italian also shares the highest proportion of typological features out of all languages considered with English (0.75) and Polish (0.74). As seen in Table 5.4, Polish and Italian display a similar distribution of verbs across semantic categories. Semantic areas such as 'motion', 'cognition', 'law/crime' are especially populated, in contrast to the smaller classes of sensory perception verbs or verbs of change. An analogous pattern was observed in the English classes in Chapter 4 (Table 4.2). Japanese, the only SOV language in the selection, has the lowest average pairwise overlap with other languages both in terms of features (0.36) and Phase 1 classes (0.33). In section 5.4.2, I zoom into specific semantic classes to analyse patterns of similarity and variation in depth.

5.4.2 Phase 2

Inter-annotator Agreement

Inter-annotator agreement (IAA) in Phase 2 is computed using Spearman's rank correlation coefficient (ρ): for each language, for each class, I calculate the average correlation of an individual annotator with the average of all other annotators (Gerz et al., 2016; Hill et al., 2015). Table 5.4 shows class size (left-hand side column) and IAA. Comparing the scores cross-lingually for classes with the highest overlap (and shared broad semantics) reveals that certain classes proved consistently easier (emotion, change, cooking), and some more challenging across languages (motion, handicraft, law/crime). In the previous study on English, class size was the the main factor affecting IAA, and an analogous correspondence (strong negative Spearman's correlation) can be observed between class size and IAA in Chinese ($\rho = -0.89$), Polish (-0.86), and Italian (-0.71). However, class size is less of a factor in Japanese ($\rho = -0.51$) and Finnish ($\rho = -0.39$).

The easier, higher-IAA classes tend to include verbs whose meanings are more concrete (*boil, bake, grate*) or are organised along clear dimensions of meaning, for instance based on increasing or decreasing intensity (negative and positive emotions, negative and positive rate of change). In more populated classes, which were naturally more heterogeneous, there was much more room for variation in item placements. Overall, the task proved especially difficult in Chinese. The difficulty arranging many characters within a crowded space was also reported in annotator feedback. This reveals important discrepancies in what class sizes can be comfortably accommodated by languages using different writing systems. While the <100 threshold proved suitable for languages using Latin script, Chinese characters may require a larger font to ensure

	EN	ZH	JP	PL	FI	IT
EN		0.30	0.29	0.36	0.34	0.40
ZH	12,423		0.23	0.26	0.21	0.25
JP	10,600	9,449		0.22	0.26	0.34
PL	20,444	11,331	9,916		0.23	0.30
FI	12,816	12,193	12,091	13,740		0.29
IT	15,652	10,436	9,087	14,744	12,749	

Fig. 5.4 Spearman's ρ correlations (above the main diagonal) on N shared pairs (below the main diagonal) for all language pairings from the sample, as well as English SpA-Verb dataset.

legibility, and consequently more space for fewer items. Although Japanese IAA scores are higher (and the impact of class size less strong), based on these observations, in future work setting a lower class size limit in Phase 1 clustering guidelines should help obtain samples that are easier for users to handle. In this study, I refrained from imposing language-specific class size restrictions to ensure direct cross-lingual comparability at all stages of data collection.

Cross-lingual Comparisons

To examine cross-lingual similarities, I compute pairwise Spearman's correlations on shared pairs sampled from the entire datasets based on the English translations for each language pair in the sample, as well as against the English SpA-Verb (Figure 5.4). The number of shared pairs is a consequence of overlap in Phase 1 (i.e., Phase 2 pairwise scores correspond to all possible pairings of items within a single class, for each class). While Polish and English share most equivalent pairs, Italian and English have the highest correlation in scores ($\rho = 0.40$), followed by English and Polish ($\rho = 0.36$). Overall, 3,052 unique pairs are shared by all languages; these pairs represent concepts considered semantically similar or related in all languages (clustered together in Phase 1) and the correlation (mean Spearman's ρ) between their scores is $\rho = 0.45$.

The availability of complete dissimilarity matrices from Phase 2 enables reconstruction of the multi-dimensional semantic space encoded in annotators' judgments in each language and exploration of cross-lingual similarities in how concepts pertaining to a given domain are organised. I zoom into two semantic areas, verbs of motion (#4) and change (#9) (IDs from Table 5.4), and compute the correlation between distance matrices (on overlap verb pairs) for all pairings of languages (including English SpA-Verb data) using the non-parametric Mantel test (Mantel, 1967), based on matrix

	EN	ZH	JP	PL	FI	IT
EN		0.33	0.36	0.25	0.39	0.43
ZH	0.85		0.37	0.28	0.32	0.26
JP	0.62	0.80		0.31	0.40	0.42
PL	0.73	0.82	0.79		0.23	0.25
FI	0.73	0.73	0.65	0.78		0.43
IT	0.87	0.81	0.68	0.88	0.66	

Fig. 5.5 Mantel test results (Pearson's correlation) between Phase 2 distance matrices for two classes, 'motion' and 'change', for pairs of languages in the multilingual dataset, as well as English SpA-Verb data (all correlations with ($p < 0.05$)).

permutations, with Pearson's product-moment correlation coefficient as test statistic. I find statistically significant correlations between all pairings of languages ($p < 0.05$) but the results in Figure 5.5 show cross-lingual and cross-domain differences. Overall, correlations on verbs of change are higher than on verbs of motion, one reason being the difference in class size: the sets of motion verbs include 74-92 members, while verbs of change 23-47 (Table 5.4). While there is more room for cross-lingual variation in pairwise distances in a more populated class of movement verbs, the cross-lingual alignment on verbs of change is also due to the nature of the class, dominated by antonymous verb pairs of opposite polarity (e.g., *increase – decrease, grow – shrink*), which are consistently spread out in the arena. Figure 5.6 includes example visualisations of the distribution of shared concepts in this class in Italian and Polish (using Principal Coordinates Analysis and English translation mappings), showing a near identical distribution along the polarity dimension of verbs of rate of change and a clear separation between verbs of inception and termination. We can observe an interesting difference in the treatment of the latter group. In Polish, verbs *start, begin* and *finish* are treated as having the positive or negative meaning component, while in Italian they are located in the neutral central area of the polarity dimension, distributed at the opposite poles of the second axis.

The moderate to strong correlations recorded indicate that the dimensions which underlie the organisation of concepts in this class – especially the polarity dimension – are cross-lingually shared. The alignment is the strongest between Polish and Italian, followed by Italian and English, while Japanese is the least aligned with other languages. To compare, positive and negative polarity play an equally important role in the organisation of verbs of emotion. In Japanese and Finnish (Figure 5.7) verbs describing positive emotions (e..g, *enjoy, like, rejoice*) are clustered at one end of the first axis

(PC1) and negative emotion verbs (e.g., *hate, despise, dislike*) occupy the opposite end, while the neutral verb *feel* is located in the centre. Further, in both languages there is a separate cluster of positive causative verbs (e.g., *please, amuse*). In Japanese, there is also a separate grouping of verbs describing causing fear, astonishment, or distress (*frighten, scare, amaze, dismay*), which cluster near the centre but on the negative polarity side, and another dense grouping of verbs causing anger, annoyance and revulsion (*enrage, frustrate, irritate, disgust*).

Results of the Mantel test on the 'motion' class (Figure 5.5) illustrate there is variation in patterns of cross-lingual affinities across different semantic domains. While Italian correlates the most with English, the correlation with Polish motion verbs is weak. Running agglomerative clustering on top of distance matrices revealed that in all three there emerge subclusters corresponding to the different medium of movement ([*dive, swim, flow*], [*run, walk, crawl*]) and a separation between static and dynamic verbs ([*lounge, poise, remain*], [*chase, dance, dash*]). However, Polish makes some additional fine-grained distinctions based on manner and speed of movement (e.g., jumping, fast vs. slow movement, motion with a change of direction). Whereas in Italian and English, verbs describing motion towards the speaker/listener form a distinct cluster. Interestingly, the distribution of verbs of motion in Japanese most closely correlates with Italian, another verb-framed language (Talmy, 1985). As mentioned in Chapter 2 (Section 2.2.2), these languages pattern together with regard to lexicalisation patterns, and encode the elements of 'Motion' and 'Path' in the verbal root, while the manner of motion is expressed outside the verb.

These preliminary analyses suggest that the collected semantic multi-arrangement data may support many other, fine-grained and in-depth lexical-typological analyses in future work, e.g., focusing on cross-lingual comparisons of the organisation of different semantic fields and examination of the most salient meaning dimensions underlying a given conceptual space.

5.4.3 Semantic vs. Semantic-Syntactic Classes

The comparisons above reveal that similar meaning components are taken into consideration when categorising the same starting sample of verbs across diverse languages. While both tasks explicitly involved semantic judgments only, given the interrelatedness of verbs' meaning and behaviour, it is interesting to examine the degree of alignment between the semantic classes – and narrower clusters emerging from the spatial semantic similarity data – and semantically-syntactically informed partitions.

In the previous section, I discussed the distribution of Japanese emotion verbs in the Phase 2 judgment space. Defined in the linguistics literature as psychological verbs of state or *psych verbs* (Belletti and Rizzi, 1988; Levin, 1993; Pesetsky, 1987), verbs in this category express a psychological state and assign the role of Experiencer to one of its arguments (Bachrach et al., 2014). In English (and other languages), psych verbs display a split pattern where some verbs map the Experiencer onto the subject and the Stimulus onto the object (*fear*-type verbs, e.g., *I fear spiders. I like dogs.*) and some show opposite behaviour (*frighten*-type verbs, e.g., *Spiders frighten me. Dogs please me.*) (e.g., Belletti and Rizzi, 1988; Hartshorne et al., 2016; Levin, 1993). Figure 5.7 shows that this distinction is reflected in semantic judgments in Japanese, where *frighten*-type verbs form distinct clusters from positive and negative *fear*-type verbs (e.g., *despise, hate, enjoy, like*). In order to quantify to what extent the semantic judgments collected in this study correspond to Levin-style classes in the different languages included in this work, I compute the overlap between the multilingual data and the manually constructed VerbNet-style classes in the same languages in previous work (Majewska et al., 2018b). These classes were created starting from a sample of 17 English VerbNet classes with 12 members each, which underwent translation and subsequent refinement by native speakers with linguistics training based on semantic and syntactic criteria (i.e., participation in the same diathesis alternations and syntactic frames) into Polish, Finnish, Mandarin, Japanese, Italian, and Croatian.

I compute B-cubed F-score between the classes of Majewska et al. (2018b) and Phase 1 classes from the present study (and Chapter 4), taking into account the subset of verbs present in both. The results in Table 5.5 show that there is substantial alignment between the two types of classifications in all languages, with most overlap found in Italian and English. Many VerbNet classes included in the sample of Majewska et al. (2018b) have their analogs in Phase 1 data. For example, the class of possession verbs (Class 3 in Table 5.4) mirrors class GET-13.5.1, while Class 20 of communication verbs corresponds to the class SAY-37.7. Interestingly, as noted in Section 5.4.1, Japanese is the only language where verbs describing 'combining' or 'uniting' (e.g., 結合する *ketsugō su* 'combine', 統合する *tōgō suru* 'integrate'), which fall within the class AMALGAMATE-22.2, form a separate semantic class in Phase 1. In the remaining languages, these verbs belong to broader Phase 1 classes (e.g., Class 5 with verbs describing moving objects in Finnish, Chinese, and Polish) and are clustered together in Phase 2.

Cluster analysis using the DBSCAN algorithm (Ester et al., 1996) on the Phase 2 similarity matrices reveals further similarities. For instance, in all languages there

	EN	ZH	JP	PL	FI	IT
B-cubed F-score	0.63	0.58	0.50	0.58	0.46	0.64

Table 5.5 B-cubed F-score calculated between Phase 1 classes and the semantic-syntactic classes of Majewska et al. (2018b).

are clusters of verbs describing violent physical contact (e.g., *beat, slap, batter, smack*) corresponding to the VerbNet class HIT-18.1: e.g., in Italian: *battere, colpire, frustare, percuotere, picchiare, urtare* (all of which are listed in the Italian HIT class in Majewska et al. (2018b)). As seen above, within the class of emotion verbs, *frighten*-type verbs, which belong to the AMUSE-31.1 class in VerbNet, form separate clusters (of positive and negative polarity) in Japanese (喜ばせる *yorokobaseru* 'please' and 楽しませる *tanoshima seru* 'amuse' vs. 驚かせる *odoroka seru* 'frighten', 怖がらせる *kowagara seru* 'scare', 狼狽させる *rōbai sa seru* 'dismay'). Whereas within perception verbs, light emission verbs (*flash, glow, shine*) form a separate cluster, e.g., in Polish *błyskać, promienieć, świecić*, or Chinese 发光 *fāguāng*, 发亮 *fā liàng*, 闪现 *shǎnxiàn* (LIGHT EMISSION-43.1), and sensory perception verbs (*stare, look, glance*) form another cluster: e.g., in Polish, *gapić się, patrzeć, spojrzeć*, and in Finnish, *tuijottaa, katsella, silmäillä* (PEER-30.3).

Since both phases of the data collection protocol presented in this work target verb meaning, rather than syntactic properties, groupings which emerge from the spatial data at the highest granularity levels are mostly narrow clusters of synonymous and near-synonymous verbs. However, the Phase 2 data lend themselves to flexible tuning of cluster size, which offers practical advantages for building verb classifications, as discussed in Chapter 4. Although the classifications of Majewska et al. (2018b) and the present dataset differ in granularity and verb coverage, the encouraging degree of item-level overlap suggests that human semantic judgments could serve as a basis for building VerbNet-style classes, further refined based on syntactic criteria. Rather than starting from translations of English classes, as in the work of Majewska et al. (2018b), the process could leverage semantic classes obtained from native speakers of the target language, thus avoiding any source language bias.

5.5 Evaluation

Evaluation is focused on two types of representation architectures: static word embeddings (Bojanowski et al., 2017) and more recently proposed large pretrained encoders (Devlin et al., 2019). I compare their ability to capture word-level semantics across

languages and domains of verb meaning. I also contrast the performance of language-specific BERT models with their massively multilingual counterpart (Devlin et al., 2019), and examine the impact of computing word-level representations *in context*, rather than by feeding items to a pretrained model *in isolation*.

Representation Models

I evaluate `fastText` (FT) as a representative non-contextualised word embedding model with proven representation capabilities on diverse NLP tasks (Mikolov et al., 2018) and coverage of 157 languages. For multi-word expressions, I compute their representations by averaging the vectors of their constituent words. I contrast the performance of FT vectors with the omnipresent state-of-the-art BERT model (Devlin et al., 2019). I derive word-level BERT representations of words and multi-word expressions in two different ways: (a) *in isolation* and (b) *in context*. In method (a), I follow the steps of Liu et al. (2019b) by (1) feeding each item to the pretrained model in isolation, (2) averaging the H hidden representations for each of the subword tokens constituting the item, and finally (3) taking the average of these subword representations to obtain the final d-dimensional representation ($d = 768$ in BERT-BASE). Again, following Liu et al. (2019b), I average over all layers (12 with BERT-BASE). In (b), I encode word meaning in context of other words using external corpora[4] in the following way. First, I randomly sample N sentences containing each item in the corpus, then, I compute the item's representation in each of N sentential contexts (averaging over constituent subword representations and hidden layers as in steps (2)-(3) above), and finally average over the N sentential representations to obtain the final representation for each item.[5] I evaluate the uncased multilingual BERT model (M-BERT) (Devlin et al., 2019), pretrained on monolingual Wikipedia corpora of 102 languages, as well as language-specific pretrained BERT encoders released for Chinese, Japanese (BERT-BASE with and without whole word masking (+WWM)), Polish, Finnish, and Italian (BERT-BASE and BERT-BASE-XXL trained on a larger Italian corpus), available in the Transformers repository (Wolf et al., 2019).[4]

[4]Details and URLs for the models and corpora used in this study are provided in Appendices D.1 and D.2.

[5]I tested different values of N (10, 100, 500) and due to negligible differences in scores only report results for $N = 100$.

Models		**Chinese**		**Japanese**		**Polish**		**Finnish**		**Italian**	
	$k =$	*gold*	*optimal*	*gold*	*optimal*	*gold*	*optimal*	*gold*	*optimal*	*gold*	*optimal*
FT		.314	.333	.250	.259	**.358**	**.377**	**.326**	**.386**	**.341**	**.389**
BERT											
(1) ISO		.246	.250	.221	.249	.190	.227	.249	.267	.205	.231
+WWM		-	-	.215	.244	-	-	-	-	-	-
+XXL		-	-	-	-	-	-	-	-	.205	.220
(2) CTX		**.333**	**.352**	.251	.253	.238	.265	.269	.306	.269	.270
+WWM		-	-	.237	.253	-	-	-	-	-	-
+XXL		-	-	-	-	-	-	-	-	.268	.300
M-BERT											
(1) ISO		.260	.284	.247	.264	.169	.216	.171	.211	.185	.196
(2) CTX		.264	.303	**.257**	**.271**	.216	.277	.200	.254	.227	.255

Table 5.6 Clustering results (F1 score) on Phase 1 classes, for the *optimal* clustering solution (highest F1 score) and with k clusters equal to the number of *gold* classes in each language (see Table 5.1). I report F1 scores for (M-)BERT embeddings computed *in isolation* (ISO) and *in context* (CTX) (see §5.5).

5.5.1 Semantic Verb Clustering

First, I evaluate the models on semantic clustering, where the task is to group the starting verb sample (Table 5.1, **N verbs**) into clusters based on semantic similarity. For each vector collection, I apply the spectral clustering algorithm (Meila and Shi, 2001; Yu and Shi, 2003), shown to produce strong results in previous work on verb clustering (Scarton et al., 2014; Sun et al., 2010; Vulić et al., 2017b), and evaluate the produced groupings against the Phase 1 classes in each language using standard clustering evaluation metrics, modified purity (MPUR) (i.e., mean precision of induced verb clusters) and weighted class accuracy (WACC), calculated as follows:

$$\text{MPUR} = \frac{\sum_{C \in \textbf{Clust}, n_{prev(C)} > 1} n_{prev(C)}}{n_{test\_verbs}} \quad (5.1) \qquad \text{WACC} = \frac{\sum_{C \in \textbf{Gold}} n_{dom(C)}}{n_{test\_verbs}} \quad (5.2)$$

where (1) each cluster C from the set of all K_{Clust} automatically induced clusters **Clust** is associated with its prevalent Phase 1 class, and $n_{prev(C)}$ is the number of verbs in an induced cluster C appearing in that class (all other verbs are considered errors). n_{test_verbs} is the total number of test verbs, and singleton clusters ($n_{prev(C)} = 1$) are not counted. In Eq. 5.2, for each C from the set of Phase 1 classes **Gold** I identify the dominant cluster from the set of induced clusters which has most verbs in common with C ($n_{dom(C)}$). The metrics are combined into an F1 score, the balanced harmonic mean of MPUR and WACC.

Results

Table 5.6 includes the results for the optimal number of clusters (highest F1), and for a fixed k equal to the number of gold truth classes. Several interesting patterns emerge in the F1 scores. First, I note that FT vectors clearly outperform the BERT models in languages using the Latin script, Italian, Finnish, Polish, achieving the top three F1 scores overall (0.389, 0.386, 0.377).[6] In Chinese and Japanese, FT vectors surpass BERT embeddings *in isolation*, but are outperformed by BERT vectors computed *in context* (in Chinese) and by the multilingual BERT in Japanese. The stronger performance of the massively multilingual model in Japanese and Chinese contrasts with the results in Italian, Finnish, and Polish, where it mostly lags behind the language-specific counterparts. In terms of relative scores, BERT and M-BERT embeddings computed over a number of sentential contexts consistently outperform their *in isolation* counterparts across all languages. On the other hand, whole word masking does not reliably improve clustering performance in Japanese, nor does using a larger training corpus in Italian (BERT-XXL).

Error Analysis

Manual inspection of the induced clusters reveals some common pitfalls and areas of difficulty. First, the evaluated models are largely oblivious to idiomatic meaning. In Polish and Italian, the FT model produces a cluster of 'possession' verbs (EN *have, give, lend, buy*), including the verbs *mieć* (PL,'to have'), *dać* and *dare* (PL/IT,'to give'). However, it also incorporates all phrasal verbs and multi-word expressions featuring these words, with meanings unrelated to the rest of the class: PL *mieć coś przeciw* ('to mind/object to'), *mieć nadzieję* ('to hope'), *mieć wpływ* ('to influence'), *dać klapsa* ('to spank'); IT *dare un'occhiata* ('to glance'). This is even more evident in Finnish, where a separate cluster of phrasal verbs with *olla* ('to be/have') emerges (e.g., *olla varuillaan* 'beware', *olla peräisin* 'originate', *olla samaa mieltä* 'agree'). Similarly, all Polish models produce a cluster of just reflexive verbs (e.g., *ślizgać się* ('to slide'), *cieszyć się* ('to rejoice'), *zdarzyć się* ('to happen')), regardless of discrepancies in meaning. In Italian, BERT models fall into the same trap, clustering reflexives regardless of their meaning (*informarsi* 'to inquire', *precipitarsi* 'to rush', *abbronzarsi* 'to tan'). However, FT vectors are more robust: there emerges a separate cluster of movement verbs, with both reflexives and non-reflexives (*saltare* 'to jump', *precipitarsi* 'to rush', *andare* 'to go'), and of knowledge-related verbs (*informarsi* 'to inquire', *studiare* 'to study',

[6]Note that lower absolute scores are also due to the overlap in Phase 1 classes, while models perform hard clustering.

Models	**Chinese**			**Japanese**			**Polish**			**Finnish**			**Italian**		
	#1	#9	#2	#1	#9	#2	#1	#9	#2	#1	#9	#2	#1	#9	#2
FT	.277	.111	**.425**	.038	−.03	.022	**.316**	.138	**.239**	**.286**	**.307**	**.414**	**.243**	.030	**.288**
BERT															
(1) ISO	.219	.041	.307	.086	−.01	.230	.205	−.02	.165	.157	.110	.108	.090	.008	.030
+WWM	-	-	-	.164	−.06	.211	-	-	-	-	-	-	-	-	-
+XXL	-	-	-	-	-	-	-	-	-	-	-	-	.083	.073	.016
(2) CTX	**.344**	.231	.330	.128	.064	.201	.237	.130	.073	.042	.245	.108	.117	.085	.038
+WWM	-	-	-	**.165**	.063	**.269**	-	-	-	-	-	-	-	-	-
+XXL	-	-	-	-	-	-	-	-	-	-	-	-	.128	**.123**	.079
M-BERT															
(1) ISO	.079	.166	.326	.118	.060	.034	.201	.039	.024	.028	.001	.137	.043	−.22	.062
(2) CTX	.305	**.236**	.213	.116	**.098**	.128	.093	**.166**	−.01	.075	.174	−.07	.146	.021	.004

Table 5.7 Word similarity evaluation results (Spearman's ρ) on three semantic domains, 'emotion' (#1), 'change' (#9), and 'cooking' (#2), in each language. BERT word-level embeddings are computed *in isolation* (ISO) or *in context* (CTX). See §5.5 for details of model configurations. Class IDs (#) correspond to aligned Phase 1 classes (Table 5.4).

istruire 'to instruct'). The attention to subword signal is apparent in clusters produced by BERT models. In languages using logographic scripts, this yields valid groupings, e.g., Japanese 再現する *saigen suru* 'to reproduce/reappear', 再生する *saisei suru* 'to reproduce', 再生利用する *saisei riyō suru* 'to recycle'. In Polish, however, *narzucać się* ('to impose') and *podrzucać* ('to toss'), and *polować* ('to hunt') and *malować* ('to paint') end up clustered together. While it could be argued that a weak semantic link (apart from the etymological one) exists between the first pair, the second pair has only coincidental orthographic overlap. Similarly, a semantically heterogeneous cluster of Italian verbs ending in *-lare* is produced (*coccolare* 'cuddle', *capitolare* 'capitulate', *scongelare* 'defrost'). Whether computed *in context* or *in isolation*, BERT word-level representations capture a lot of subword- and surface-level information without fully capturing the higher-level semantic signal, which negatively affects cluster quality.

5.5.2 Word Similarity

I compute Spearman's ρ correlation between the ranks of models' similarity scores and those of human judgments from Phase 2. To ensure reliability of the results, I perform evaluation on a thresholded subset of each dataset focusing on classes with IAA above $\rho = 0.3$ (THR) (Table 5.1). I also compare the models' capacity to discriminate between

related concepts within a narrow semantic domain and report scores on three semantic classes: 'emotion' (#1), 'change' (#9), 'cooking' (#2).[7]

Main Results

Figure 5.8 shows the results on the thresholded datasets for all the model configurations. The primacy of FT vectors in Polish, Finnish, and Italian is again conspicuous, while in Chinese and Japanese the pretrained encoders are in the lead, with noticeably lower FT performance recorded for Japanese than in the other languages. Results achieved on the THR sets repeat the patterns seen in the clustering task: contextualised variants of BERT embeddings (CTX) outperform those computed *in isolation* (ISO), and the language-specific encoders prove to capture richer semantics than the massively multilingual model – with the exception of Japanese, where contextualised M-BERT again achieves the top result (albeit noticeably lower than top THR scores in other languages). The relatively stronger M-BERT results on Japanese, as well as Chinese, illustrate the known unfavourable characteristic of multilingual pretraining with a subword vocabulary shared by 102 languages. The languages with scripts distinct from those of the majority of languages covered by M-BERT do not share their subwords with a large number of other languages, and their language-specific subwords constitute a large proportion of the total subword vocabulary; in consequence, the model can capitalise on this proliferation to produce higher-quality representations. Conversely, the representation quality is degraded for languages with very rich and productive morphology like Finnish or Polish, despite the availability of training data. This also applies to language-specific BERT models: given the same vocabulary capacity, morphologically rich languages have many words split into subwords and fewer full words represented in the vocabulary than analytic languages like Chinese or English.

Impact of Lexical Representation Extraction Parameters

To explore the potential of generating stronger word-level embeddings from BERT models, I investigated the impact of two parameters on lexical representation extraction: the number of hidden layers to average over (all 12 or first 8) and the inclusion of the special classification token ([CLS]) in the subword averaging step. Table 5.8 summarises the results for Polish and Finnish. The experiments reveal that including the [CLS] token yields better lexical embeddings in both languages. However, whether averaging over 12 or 8 layers produces strongest results is language-specific. Notably, the top-performing

[7] I carried out evaluation on all classes but report selected results (on highest IAA classes cross-lingually) for brevity.

BERT		**Polish**				**Finnish**			
L	CLS	THR	#1	#9	#2	THR	#1	#9	#2
12	−	.097	.205	−.019	.165	.112	.157	.110	.108
	+	**.172**	**.228**	**.083**	**.202**	.227	.186	.151	.190
8	−	.086	.183	−.011	.141	.151	.223	.141	.160
	+	.142	.185	.081	.175	**.250**	**.234**	**.205**	**.280**

Table 5.8 Results for Polish and Finnish BERT *in isolation* vectors, averaged over all 12 or first 8 layers (L), with the CLS token (+) or without it (−).

Finnish BERT configuration ($\rho = 0.250$) outperforms FT embeddings ($\rho = 0.248$) on the THR set. These findings suggest that careful language-specific tuning of the extraction configuration is crucial to achieve optimal performance.

Results on Semantic Domains

Although more variation in terms of primacy of one model variant over the other is expected on the semantic classes due to smaller dataset size, the general pattern whereby computing BERT embeddings by averaging N contextualised representations boosts performance applies in 72% of cases. Additional observations can be drawn from results on individual domains.

Class #9, including verbs describing change in size or speed (e.g., *accelerate, increase, shrink*), is especially rich in synonyms and antonyms. Due to antonyms' high semantic overlap they are often confused with synonyms by distributional models learning purely from patterns of occurrences in raw text. This effect also emerges in the present results, where performance on this class is the lowest for most model configurations and languages. For example, Polish BERT assigns a nearly identical similarity score to both synonym pairs *wzrastać – rosnąć* ('to rise/grow'), *zapoczątkować – rozpoczynać* ('originate – begin') and antonym pairs *rosnąć – maleć* ('grow – decrease'), *kończyć – zaczynać* ('finish – start'), mistaking the strong relatedness of the latter pairs for similarity. The overall stronger `fastText` is prone to the same kind of error: it assigns one of the highest similarity scores in this set ($sim_{cos} = 0.89$) to the pair of antonyms *zmniejszać – zwiększać* ('diminish – increase'). Finnish results are the exception, with relatively stronger model performance on this set. This is partly due to the slightly broader coverage of this class, which also incorporates verbs of being and existence (Table 5.4), with smaller proportion of challenging antonymous pairs. Among these, some receive a lower score (e.g., *vähentää – kasvaa*, 'decrease – increase' 0.44 FT), but the similarity of many is still overestimated (e.g., *aloittaa – lopettaa*, 0.79 FT). Interestingly, this class is where the multilingual model outperforms the language-specific counterparts in Chinese, Japanese, and Polish. In Italian, where

class #9 has only 23 members, most of which stand in antonymous relations (e.g., *crescere* – *decrescere* 'increase – decrease', *aumentare* – *diminuire* 'rise – drop', *iniziare* – *finire* 'begin – finish'), the BERT model trained on the larger corpus is the most robust. Results on this class illustrate that semantic areas which are easier for humans to reason about are not necessarily less challenging for models.

An area where greater ease of human judgment is reflected in relatively higher model performance is the domain of cooking verbs in Chinese, Finnish, and Japanese, where the highest overall scores are recorded (>0.4 for `fastText` in Chinese and Finnish), and the top model scores in Japanese (BERT). While I do not report all class-specific correlations for brevity, they reveal further interesting patterns as to the semantic domains which prove easiest for models to accurately capture. In Italian, highest model correlations are achieved on verbs of communication and destruction (top scores >0.4 (FT)), while verbs of physical violence are the domain with highest correlations in Polish (>0.3 FT), Finnish (>0.4 FT) and Chinese (>0.4 BERT). In Japanese, the best result overall is achieved on verbs of cognition (0.276 BERT), followed by verbs of physiological processes with >0.2 correlations scored by the contextualised BERT models. Similar analyses on specific semantic domains can help identify strengths and deficiencies of different types of embeddings and highlight the areas of meaning which pose challenges across languages, guiding further developments in representation learning.

5.5.3 Main Observations

The presented evaluation revealed the dataset to be a challenging benchmark, and provided a number of insights into the potential of the evaluated models to capture verbal lexical semantics across languages.

- Overall, model performance across tasks shows a split pattern: the pretrained encoders surpass static word embeddings in Chinese and Japanese, but are outperformed by `fastText` by a significant margin in languages using the Latin script, Polish, Finnish, and Italian. There is potential to derive higher-quality word-level BERT embeddings in those languages through careful selection of language-specific lexical representation extraction parameters.

- BERT word-level embeddings derived by averaging over N occurrences in context prove predominantly stronger than those obtained by feeding words into the pretrained model in isolation, with more variation observed in the case of the smaller semantic sets.

- The results achieved on the thresholded datasets show a clear advantage of monolingual pretraining over the massively multilingual pretraining – with the exception of Japanese, where M-BERT achieves the top results, as is the case in semantic clustering.
- Error analysis revealed that clustering performance of the pretrained encoders suffers due to the primacy given to low-level subword signal over the high-level semantic information, while an important area of difficulty for all models in the lexical similarity task is the problem of teasing apart antonymous and synonymous word pairs.

5.6 Conclusion and Future Work

In this chapter, I examined the applicability of the two-phase data collection protocol to a diverse sample of languages, which yielded the first large-scale multilingual evaluation resource constructed via semantic clustering and spatial arrangement, targeting verb semantics in Chinese, Japanese, Polish, Finnish, and Italian. The resource includes 16-18 semantic classes and over 20 thousand fine-grained pairwise lexical similarity scores in each language, made available online. I employed the aligned semantic classes to examine cross-lingual variation in the organisation of particular domains of verb meaning, as well as to identify areas of overlap (e.g., positive vs. negative verbs of rate of change) and cross-lingual affinities in human similarity judgments (e.g., the same classes are easier and harder to judge). Further, I compared the semantically driven classes created in this study with Levin-style semantic-syntactic partitions, revealing substantial overlap between the two types of classifications. The dual nature and vast coverage of the dataset enable evaluation of representation models on two tasks, semantic clustering and word similarity, and focused probing analyses on specific semantic domains, revealing aspects of verbal meaning which elude models' representation capacity. The low overall model performance indicates that estimating similarity between a large number of semantically proximate concepts linked by fine-grained relations is a challenging task.

While the multilingual adaptation of the two-phase approach proved largely successful, demonstrating the flexibility of the design, there are several areas which would benefit from further research efforts and improvements. First, the data collection pipeline could be extended to include a more developed training phase for annotators, to allow them not only to understand the purpose of the task, but also to practice more and get instant feedback on their first attempts. The current qualification tasks serve

the former role, but given their small scale (7 verbs to cluster or arrange), they do not directly prepare the participants for the more challenging samples. Dedicating more time to annotator training would increase the duration of the process, but would likely pay off in terms of annotation quality. Furthermore, as discussed earlier in this chapter, the issue of legibility which arose with logographic scripts in crowded arenas shows that smaller samples and larger labels should be favoured going forward to ensure annotation quality. This, in turn, would require imposing rigid constraints on the cluster sizes permitted in Phase 1. As hinted at in the previous chapter, hierarchical clustering emerges as a promising avenue for future endeavours: Instead of enforcing small clusters regardless of the composition of the sample, the Phase 1 task could explicitly encourage creating narrower clusters within the broad groupings, and these smaller sets would then be used as input to Phase 2. This would, however, result in more limited coverage of each semantic domain in terms of pairwise distance scores. Future work should therefore explore whether relative similarities between different subsets belonging to the same broader category could be inferred, for example, by first identifying the prototypical members of each cluster (e.g., based on the closeness centrality measure), and then collecting Phase 2 judgments on cluster prototypes, rather than the entire broad categories.

A promising direction for future work is the study of typological variation in the way that different semantic domains are represented in the mental lexicon using the spatial arrangement data. As demonstrated in the analyses presented in this chapter, the multilingual Phase 1 and Phase 2 data encode a wealth of information about how the speakers of different languages reason about verb similarity and conceptualise verb meaning. Future work could expand these efforts into a large-scale psycholinguistic study with larger groups of participants and a fully randomised experiment design. As far as model evaluation is concerned, the fine-grained clusters derived from the Phase 2 data can be readily used to explore the potential of NLP models to (semi-)automatically create verb classes and semantic resources in languages where those are still lacking. Moreover, mapped cross-lingual verb similarity datasets derived for all language pairs using English translation mappings can be employed for general and semantic domain-specific evaluation of cross-lingual representation learning algorithms.

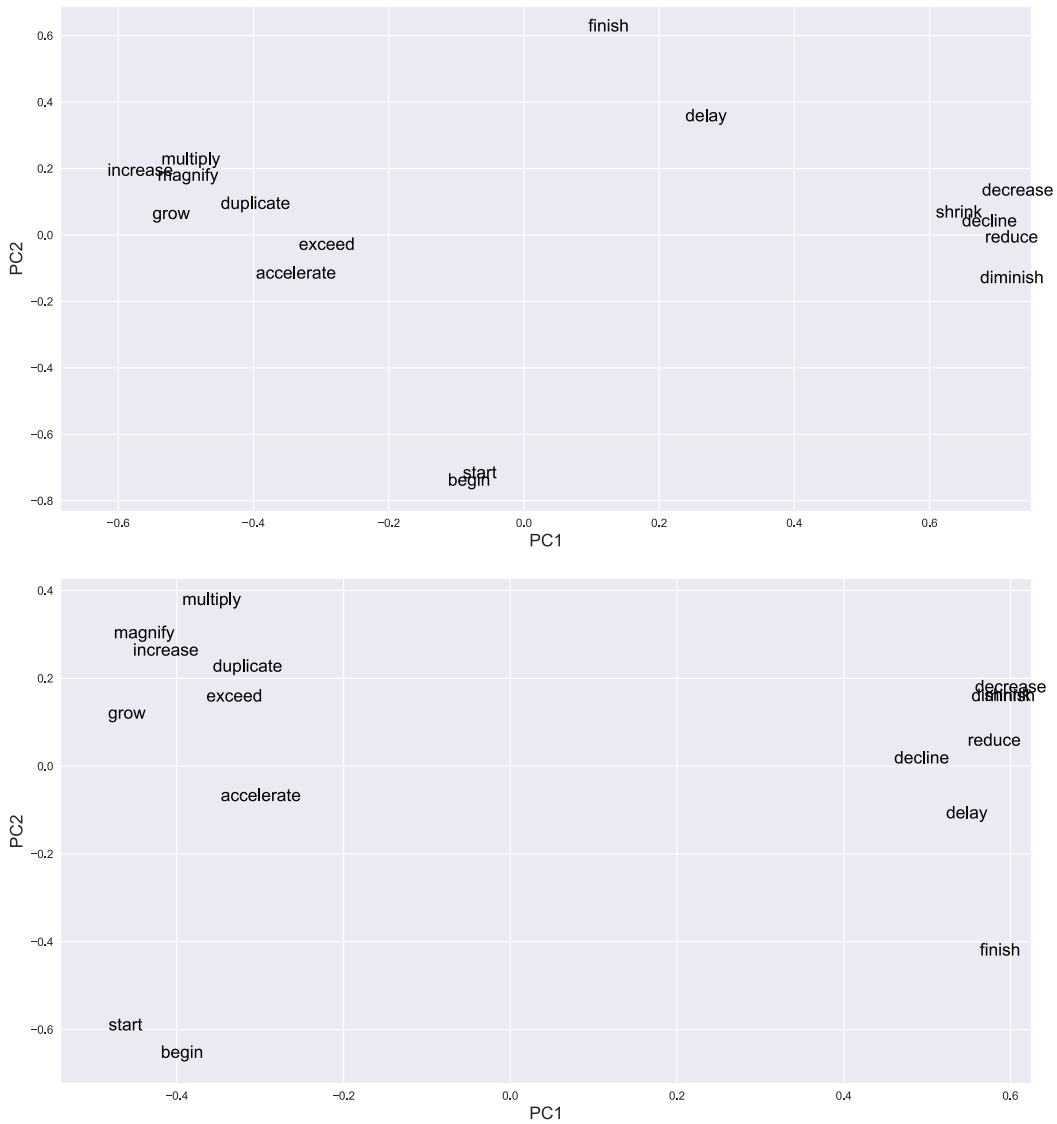


Fig. 5.6 Visualisation of PCoA on the 'change' class in Italian (above) and Polish (below) using English translation labels.

Fig. 5.7 Visualisation of PCoA on the ‘emotion’ class in Japanese (above) and Finnish (below) using English translation labels.

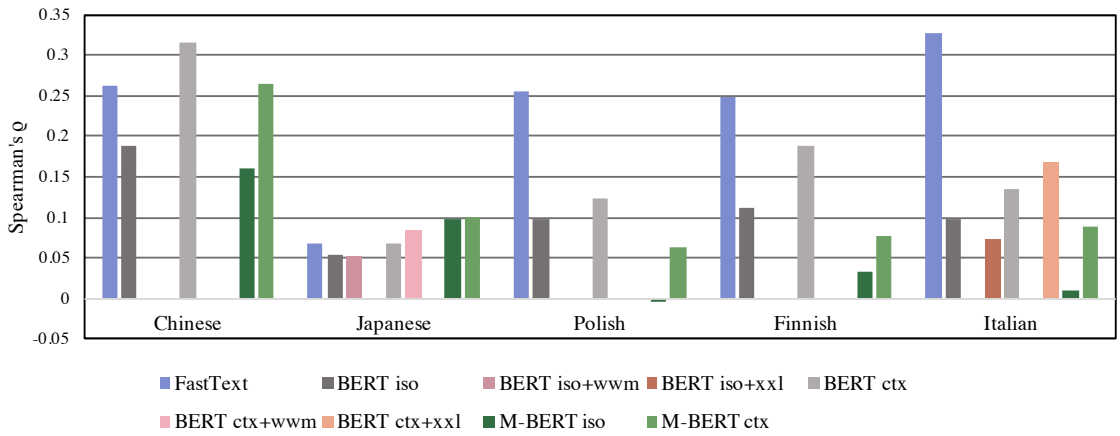


Fig. 5.8 Word similarity evaluation results (Spearman's ρ) on the thresholded sets (THR). Included are `fastText` word embeddings trained on Common Crawl and Wikipedia corpora (CC+Wiki) and language-specific and multilingual BERT models. BERT word-level embeddings are computed *in isolation* (iso) or *in context* (ctx). See §5.5 for details of model configurations.

Chapter 6

Verb Knowledge Injection for Multilingual Event Processing

6.1 Introduction

Large Transformer-based encoders, pretrained with self-supervised language modelling (LM) objectives, form the backbone of state-of-the-art models for most Natural Language Processing (NLP) tasks (Devlin et al., 2019; Liu et al., 2019d; Yang et al., 2019b). Recent probing experiments showed that they implicitly extract a non-negligible amount of linguistic knowledge from text corpora in an unsupervised fashion (Hewitt and Manning, 2019; Rogers et al., 2020b; Vulić et al., 2020b, *inter alia*). In downstream tasks, however, they often rely on spurious correlations and superficial clues (Niven and Kao, 2019) rather than a deep understanding of language meaning (Bender and Koller, 2020), which is detrimental for both generalisation and interpretability (McCoy et al., 2019). In this chapter, I focus on a specific facet of linguistic knowledge, namely reasoning about events. For instance, in the sentence "*Stately, plump Buck Mulligan came from the stairhead, bearing a bowl of lather*", an event of COMING occurs in the past, with BUCK MULLIGAN as a participant, simultaneously to an event of BEARING with an additional participant, a BOWL. Identifying tokens in the text that mention events and classifying the temporal and causal relations among them (Ponti and Korhonen, 2017) is crucial to understand the structure of a story or dialogue (Carlson et al., 2002; Eisenberg and Finlayson, 2017; Miltsakaki et al., 2004) and to ground a text in real-world facts (Doddington et al., 2004).

Verbs (with their arguments) are prominently used for expressing events (with their participants). Thus, fine-grained knowledge about verbs, such as the syntactic patterns in which they partake and the semantic frames they evoke, may help pretrained

encoders to achieve a deeper understanding of text and improve their performance in event-oriented downstream tasks. In English, there already exist expert-curated computational resources that organise verbs into classes based on their syntactic-semantic properties (Jackendoff, 1990; Levin, 1993). In particular, here I consider VerbNet (Kipper Schuler, 2005) and FrameNet (Baker et al., 1998) as rich sources of verb knowledge. Expanding a line of research on injecting external linguistic knowledge into (LM-)pretrained models (Lauscher et al., 2020b; Levine et al., 2020; Peters et al., 2019), I integrate verb knowledge into contextualised representations for the first time. I devise a new method to distill verb knowledge into dedicated *adapter* modules (Houlsby et al., 2019; Pfeiffer et al., 2020b), which reduce the risk of (catastrophic) forgetting of distributional knowledge and allow for seamless integration with other types of knowledge. I hypothesise that complementing pretrained encoders through verb knowledge in such a modular fashion should benefit model performance in downstream tasks that involve event extraction and processing. I first put this hypothesis to the test in English monolingual event identification and classification tasks from the TempEval (UzZaman et al., 2013) and ACE (Doddington et al., 2004) datasets. Foreshadowing, I report modest but consistent improvements in the former, and significant performance boosts in the latter, thus verifying that verb knowledge is indeed paramount for deeper understanding of events and their structure.

However, as discussed in the preceding chapters, expert-curated resources are not available for most of the languages spoken worldwide. One possibility lies in the automatic transfer of verb knowledge across languages. In what follows, I investigate its effectiveness using English as the source and transferring from it the information pertaining to verbs' properties to three diverse languages, Spanish, Arabic and Mandarin Chinese. Concretely, I compare (1) zero-shot model transfer based on massively multilingual encoders and English constraints with (2) automatic translation of English constraints into the target language. The results demonstrate that both techniques are successful and there is a lot of useful signal that can be carried through to the target language from the resource-rich source. Second, given that fine-grained verb class membership information is language-specific and automatic transfer inevitably introduces some noise into the lexical constraints, I examine the potential of leveraging non-expert verb knowledge in the target language instead. I directly evaluate the usefulness of the semantic verb classes and similarity data acquired from non-expert native speakers in the studies presented in Chapters 4 and 5. I investigate, for the first time, whether human judgments of verb semantics can be harnessed as a source of non-distributional knowledge to enrich the linguistic capacity of large pretrained models

and benefit their downstream task performance. The results of the evaluation on event processing tasks reveal that the non-expert data provide useful information beyond what is already encoded in the underlying model's parameters. Crucially, they offer a valuable alternative to cross-lingual transfer, resulting in performance boosts superior to those offered by automatically transferred English expert knowledge in Chinese. The comparison of the relative benefits of cross-lingual transfer and injection of in-target lexical knowledge sheds further light on an important question guiding the research presented in this thesis: To what extent can verb classes (and predicate–argument structures) be considered universal, rather than varying across languages (Hartmann et al., 2013)?

Overall, the main contributions of the work presented in this chapter consist in 1) mitigating the limitations of pretrained encoders regarding event understanding by supplying verb knowledge from external resources; 2) proposing a new method to do so in a modular way through adapter layers; 3) exploring techniques to transfer verb knowledge to resource-poor languages, and 4) evaluating the potential of infusing models with semantic information derived directly from non-expert language users. The gains in performance observed across four diverse languages and several event processing tasks and datasets warrant the conclusion that complementing distributional knowledge with human-generated verb knowledge is both beneficial and cost-effective.[1]

6.2 Verb Knowledge for Event Processing

Figure 6.1 illustrates the proposed framework for injecting verb knowledge from a lexical resource and leveraging it in downstream event extraction tasks. First, I inject the external verb knowledge, formulated as the so-called *lexical constraints* (Mrkšić et al., 2017; Ponti et al., 2019) (in the case of this study – verb pairs, see §6.2.1), into a (small) additional set of *adapter parameters* (§6.2.2) (Houlsby et al., 2019). In the second step (§6.2.3), I combine the language knowledge encoded by the Transformer's original parameters and the verb knowledge from *verb adapters* to solve a particular event extraction task. To this end, I either *a)* fine-tune both sets of parameters (1. pretrained LM; 2. verb adapters) or *b)* freeze both sets of parameters and insert an additional set of *task-specific adapter parameters*. In both cases, the task-specific training is informed both by the general language knowledge captured in the pretrained LM, and the specialised verb knowledge, captured in the verb adapters.

[1]The research presented in this chapter has been published in Majewska et al. (2021b).

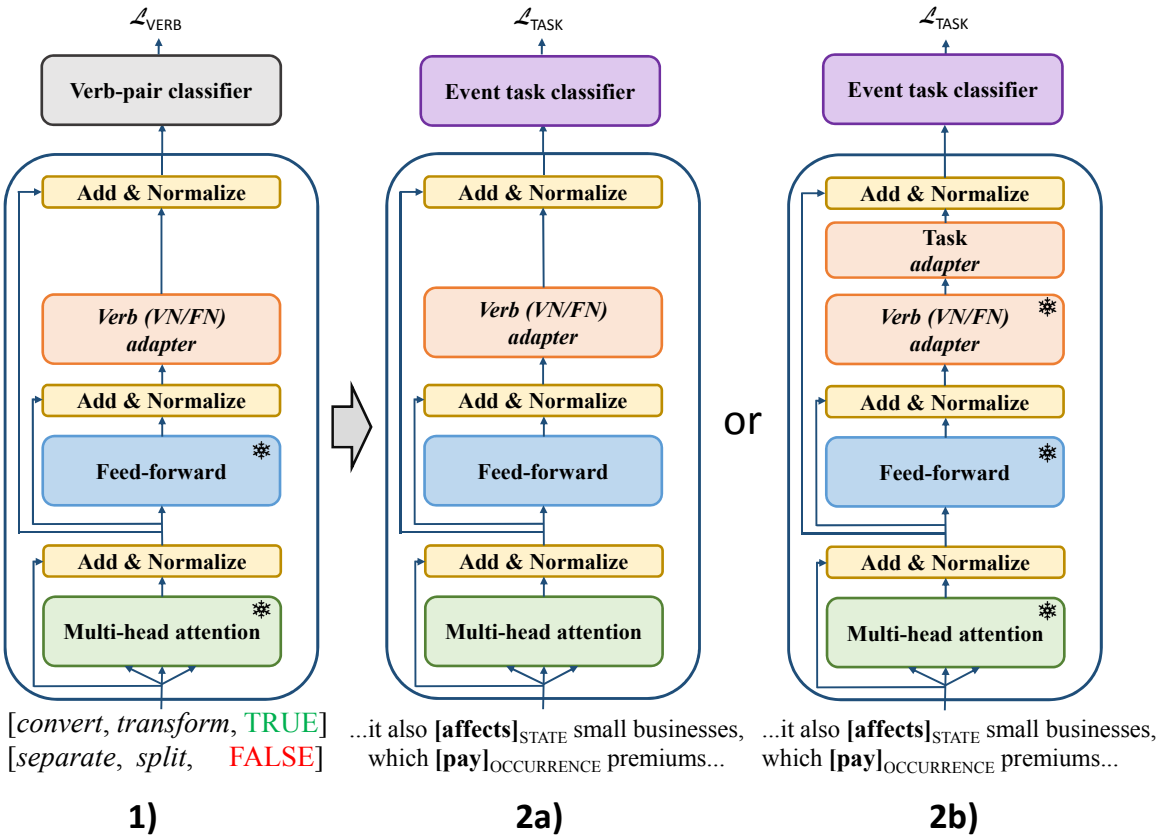


Fig. 6.1 Framework for injecting verb knowledge into a pretrained Transformer encoder for event processing tasks. **1)** Dedicated *verb adapter* parameters trained to recognise pairs of verbs from the same VerbNet (VN) class or FrameNet (FN) frame; **2)** Fine-tuning for an event extraction task (e.g., event trigger identification and classification (UzZaman et al., 2013)): **a)** *full fine-tuning* – Transformer's original parameters and verb adapters both fine-tuned for the task; **b)** *task adapter (TA) fine-tuning* – additional *task adapter* is mounted on top of *verb adapter* and tuned for the task. For simplicity, I show only a single transformer layer; verb- and task-adapters are used in all Transformer layers. Snowflakes denote frozen parameters in the respective training step.

6.2.1 Sources of Lexical Verb Knowledge

Given the inter-connectedness between verbs' meaning and syntactic behaviour (Kipper Schuler, 2005; Levin, 1993), I assume that refining latent representation spaces with semantic content and predicate-argument properties of verbs should have a positive effect on event extraction tasks that strongly revolve around verbs. As discussed in Chapter 2, lexical classes defined in terms of shared semantic-syntactic properties provide a mapping between the verbs' senses and the morpho-syntactic realisation of their arguments (Jackendoff, 1990; Levin, 1993); for any given verb, a set of rich

semantic-syntactic properties can be inferred based on its class membership. In this work, I explicitly harness this rich linguistic knowledge to help LM-pretrained Transformers in capturing regularities in the properties of verbs and their arguments, and consequently improve their ability to reason about events. I select two major English lexical databases – VerbNet (Kipper Schuler, 2005) and FrameNet (Fillmore, 1982) – as sources of verb knowledge at the semantic-syntactic interface, each representing a different lexical framework. Despite the different theories underpinning the two resources, their organisational units – verb classes and semantic frames – both capture regularities in verbs' semantic-syntactic properties.[2]

As discussed in detail in Chapter 2, VerbNet (VN) (Kipper et al., 2006; Kipper Schuler, 2005) organises verbs into classes based on the overlap in their semantic properties and syntactic behaviour, building on the premise that a verb's predicate-argument structure informs its meaning (Levin, 1993). The resource's reliance on semantic-syntactic coherence as a class membership criterion means that semantically related verbs may end up in different classes because of differences in their combinatorial properties. For instance, verbs *split* and *separate* are members of two different classes with identical sets of arguments' thematic roles, but with discrepancies in their syntactic realisations (e.g., the syntactic frame NP V NP APART is only permissible for the 'split-23.2' verbs: *The book's front and back cover split apart*, but not **The book's front and back cover separated apart*). Although the sets of syntactic descriptions and corresponding semantic roles defining each VerbNet class are English-specific, the underlying notion of a semantically-syntactically defined verb class is thought to apply cross-lingually (Jackendoff, 1990; Levin, 1993), and its translatability has been demonstrated in previous work (Majewska et al., 2018b; Vulić et al., 2017b). In contrast, FrameNet (FN) (Baker et al., 1998) is more semantically-oriented: grounded in the theory of frame semantics (Fillmore, 1976, 1977, 1982), it organises concepts according to semantic frames, i.e., schematic representations of situations and events, which they evoke, each characterised by a set of typical roles assumed by its participants. The word senses associated with each frame (FrameNet's lexical units) are similar in terms of their semantic content, as well as their typical argument structures. Further, as emphasised in Chapter 2, semantically defined FrameNet frames are thought to be

[2]Another rich lexical resource not considered in this study is WordNet. While it provides records of verbs' senses and (some) semantic relations between them, WordNet lacks comprehensive information about the (semantic-)syntactic frames in which verbs participate. I therefore believe that verb knowledge from WordNet would be less effective in downstream event extraction tasks than that from VerbNet and FrameNet.

widely cross-lingually shared (e.g., descriptions of transactions will include the same frame elements *Buyer, Seller, Goods, Money* in most languages).

Moreover, I investigate the potential of using non-expert lexical knowledge as an alternative source of information about verb meaning. I employ the semantic verb classes created for English and Chinese (Chapters 4 and 5; hereafter *SpA-Verb* classes) to derive lexical constraints to fine-tune a pretrained Transformer to better capture verb semantics. What is more, I investigate the impact of the granularity of verb classes on the downstream performance of the model by additionally leveraging narrow semantic clusters emerging from the spatial similarity data.

6.2.2 Training Verb Adapters

Training Task and Data Generation

In order to encode information about verbs' membership in VerbNet classes or FrameNet frames into a pretrained Transformer, I devise an intermediary training task in which I train a dedicated VN/FN-knowledge adapter (hereafter *VN-Adapter* and *FN-Adapter*). I frame the task as binary word-pair classification: I predict if two verbs belong to the same VN class or FN frame. I extract training instances from FN and VN independently. This allows for a separate analysis of the impact of verb knowledge from each resource.

I generate positive training instances by extracting all unique verb pairings from the set of members of each main VN class/FN frame (e.g., *walk – march*), resulting with 181,882 positive instances created from VN and 57,335 from FN. I then generate $k = 3$ negative examples for each positive example in a training batch by combining controlled and random sampling. In controlled sampling, I follow prior work on semantic specialisation (Glavaš and Vulić, 2018b; Lauscher et al., 2020b; Ponti et al., 2019; Wieting et al., 2015). For each positive example $p = (w_1, w_2)$ in the training batch B, I create two negatives $\hat{p}_1 = (\hat{w}_1, w_2)$ and $\hat{p}_2 = (w_1, \hat{w}_2)$; $\hat{w}_1$ is the verb from batch B other than w_1 that is closest to w_2 in terms of their cosine similarity in an auxiliary static word embedding space $X_{aux} \in \mathbb{R}_d$; conversely, $\hat{w}_2$ is the verb from B other than w_2 closest to w_1. I additionally create one negative instance $\hat{p}_3 = (\hat{w}_1, \hat{w}_2)$ by randomly sampling $\hat{w}_1$ and $\hat{w}_2$ from batch B, not considering w_1 and w_2. I make sure that negative examples are not present in the global set of all positive verb pairs from the resource.

Similar to Lauscher et al. (2020b), I tokenise each (positive *and* negative) training instance into WordPiece tokens, prepended with sequence start token `[CLS]`, and with `[SEP]` tokens in between the verbs and at the end of the input sequence. I consider

the representation of the [CLS] token, $\mathbf{x}_{CLS} \in \mathbb{R}^h$ (with h as the hidden state size of the Transformer), output by the last Transformer layer to be the latent representation of the verb pair, and feed it to a simple binary classifier:[3]

$$\hat{\mathbf{y}} = \text{softmax}(\mathbf{x}_{\text{CLS}}\mathbf{W}_{cl} + \mathbf{b}_{cl}) \tag{6.1}$$

with $\mathbf{W}_{cl} \in \mathbb{R}^{h \times 2}$ and $\mathbf{b}_{cl} \in \mathbb{R}^2$ as classifier's trainable parameters. I train by minimising the standard cross-entropy loss ($\mathcal{L}_{VERB}$ in Figure 6.1).

I then apply the same method to generate training data from SpA-Verb classes for English and Chinese, which I subsequently use to train a dedicated *SpA-Adapter* in each language, or fine-tune the entire Transformer model (see §6.3). The process results in 29,721 positive instances created for English and 23,990 for Chinese (i.e., all unique pairings of verbs appearing in the same Phase 1 class). Additionally, I carry out a follow-up experiment to investigate how different classification granularities affect the utility of the derived lexical constraints. For each English SpA-Verb class, I generate k clusters of semantically similar verbs based on the averaged Euclidean distances from Phase 2 using the agglomerative clustering algorithm with average linkage. To determine the optimal value of k clusters for each class, I compute the gap statistic (Tibshirani et al., 2001) for each clustering solution which compares the change in within-cluster dispersion to what would be expected under a null distribution (e.g., no clusters). I then take the optimal k clusters from each Phase 1 class and generate all unique pairings of verbs sharing a cluster, following the above outlined approach. This results in 5,102 positive training instances.

Adapter Architecture

Instead of directly fine-tuning all parameters of the pretrained Transformer, I opt for storing verb knowledge in a separate set of adapter parameters, keeping the verb knowledge separate from the general language knowledge acquired in pretraining. This (1) allows downstream training to flexibly combine the two sources of knowledge, and (2) bypasses the issues with catastrophic forgetting and interference (Hashimoto et al., 2017; de Masson d'Autume et al., 2019). I adopt the adapter architecture of Pfeiffer et al. (2020a,c) which exhibits comparable performance to the more commonly used

[3] I also experimented with sentence-level tasks for injecting verb knowledge, with target verbs presented in sentential contexts drawn from example sentences from VN/FN: I fed (a) pairs of sentences in a binary classification setup (e.g., *Jackie leads Rose to the store. – Jackie escorts Rose.*); and (b) individual sentences in a multi-class classification setup (predicting the correct VN class/FN frame). Both these variants with sentence-level input, however, led to weaker downstream performance.

Houlsby et al. (2019) architecture, while being computationally more efficient. In each Transformer layer l, I insert a single adapter module ($Adapter_l$) after the feed-forward sub-layer. The adapter module itself is a two-layer feed-forward neural network with a residual connection, consisting of a down-projection $\mathbf{D} \in \mathbb{R}^{h \times m}$, a GeLU activation (Hendrycks and Gimpel, 2016), and an up-projection $\mathbf{U} \in \mathbb{R}^{m \times h}$, where h is the hidden size of the Transformer model and m is the dimension of the adapter:

$$Adapter_l(\mathbf{h}_l, \mathbf{r}_l) = \mathbf{U}_l(\text{GeLU}(\mathbf{D}_l(\mathbf{h}_l))) + \mathbf{r}_l \tag{6.2}$$

where $\mathbf{r}_l$ is the residual connection, output of the Transformer's feed-forward layer, and $\mathbf{h}_l$ is the Transformer hidden state, output of the subsequent layer normalisation.

6.2.3 Downstream Fine-Tuning for Event Tasks

With verb knowledge from VN/FN injected into the parameters of VN-/FN-Adapters, I proceed to the downstream fine-tuning for a concrete event extraction task. Tasks that I experiment with (see §6.3) are (1) token-level event trigger identification and classification and (2) span extraction for event triggers and arguments (a sequence labelling task). For the former, I mount a classification head – a simple single-layer feed-forward softmax regression classifier – on top of the Transformer augmented with VN-/FN-Adapters. For the latter, I follow the architecture from prior work (M'hamdi et al., 2019; Wang et al., 2019a) and add a CRF layer (Lafferty et al., 2001) on top of the sequence of Transformer's outputs (for subword tokens), in order to learn inter-dependencies between output tags and determine the optimal tag sequence.

For both types of downstream tasks, I propose and evaluate two different fine-tuning regimes: (1) *full downstream fine-tuning*, in which I update both the original Transformer's parameters and VN-/FN-Adapters (see *2a* in Figure 6.1); and (2) *task-adapter* (**TA**) *fine-tuning*, where I keep both Transformer's original parameters and VN-/FN-Adapters frozen, while stacking a new trainable *task adapter* on top of the VN-/FN-Adapter in each Transformer layer (see *2b* in Figure 6.1).

6.2.4 Cross-Lingual Transfer

Creation of curated resources like VN or FN take years of expert linguistic labour. Consequently, such resources do not exist for a vast majority of languages. Given the inherent cross-lingual nature of verb classes and semantic frames (see §6.2.1), I investigate the potential for verb knowledge transfer from English to target languages, without any manual target-language adjustments. Massively multilingual Transformers, such as multilingual BERT (mBERT) (Devlin et al., 2019) or XLM-R (Conneau et al.,

	VerbNet	FrameNet	SpA-Verb
English (EN)	181,882	57,335	29,721
Spanish (ES)	96,300	36,623	
Chinese (ZH)	60,365	21,815	23,990
Arabic (AR)	70,278	24,551	

Table 6.1 Number of positive training VN and FN verb pairs in English and in each target language obtained via the VTRANS method (§6.2.4), and in-target verb pairs obtained from the SpA-Verb data for English and Chinese.

2020a) have become the *de facto* standard mechanisms for zero-shot (ZS) cross-lingual transfer. I adopt mBERT in the first language transfer approach: I fine-tune mBERT first on the English verb knowledge and then on English task data and then simply make task predictions for the target language input.

The second transfer approach, dubbed VTRANS, is inspired by the work on cross-lingual transfer of semantic specialisation for static word embedding spaces (Glavaš et al., 2019; Ponti et al., 2019; Wang et al., 2020b). Starting from a set of positive pairs P from English VN/FN, VTRANS involves three steps: (1) automatic translation of verbs in each pair into the target language, (2) filtering of the noisy target language pairs by means of a relation prediction model trained on the English examples, and (3) training the verb adapters injected into mBERT with target language verb pairs. For (1), I translate the verbs by retrieving their nearest neighbour in the target language from the shared cross-lingual embedding space, aligned using the Relaxed Cross-domain Similarity Local Scaling (RCSLS) model of Joulin et al. (2018). Such a translation procedure is liable to error due to an imperfect cross-lingual embedding space as well as polysemy and out-of-context word translation. I mitigate these issues in step (2), where the set of noisily translated target language verb pairs is purified by means of a neural lexical-semantic relation prediction model, the Specialisation Tensor Model (STM) (Glavaš and Vulić, 2018a), here adjusted for binary classification. I train the STM for the same task as verb adapters during verb knowledge injection (§6.2.2): to distinguish (positive) verb pairs from the same English VN class/FN frame from those from different VN classes/FN frames. In training, the input to STM are static word embeddings of English verbs taken from a shared cross-lingual word embedding space. I then make predictions in the target language by feeding vectors of target language verbs (from noisily translated verb pairs), taken from the same cross-lingual word embedding space, as input for STM (see Appendix E.2 for more details on STM training). Finally, in step (3), I retain only the target language verb pairs identified by STM as positive pairs and perform *direct* monolingual FN-/VN-Adapter training

in the target language, following the same protocol used for English, as described in §6.2.2.

6.3 Experimental Setup

Event Processing Tasks and Data

In light of the pivotal role of verbs in encoding the unfolding of actions and occurrences in time, as well as the nature of the relations between their participants, sensitivity to the cues they provide is especially important in event processing tasks. There, systems are tasked with detecting that *something happened*, identifying *what type* of occurrence took place, as well as what *entities* were involved. Verbs typically act as the organisational core of each such event schema,[4] carrying a lot of semantic and structural weight. Therefore, a model's grasp of verbs' properties should have a bearing on ultimate task performance. Based on this assumption, I select event extraction and classification as suitable evaluation tasks to profile the methods from §6.2.

These tasks and the corresponding data are based on the two prominent frameworks for annotating event expressions: TimeML (Pustejovsky et al., 2003, 2005) and the Automatic Content Extraction (ACE) (Doddington et al., 2004). TimeML was developed as a rich markup language for annotating event and temporal expressions, addressing the problems of identifying event predicates and anchoring them in time, determining their relative ordering and temporal persistence (i.e., how long the consequences of an event last), as well as tackling contextually underspecified temporal expressions (e.g., *last month, two days ago*). Currently available English corpora annotated based on the TimeML scheme include the TimeBank corpus (Pustejovsky et al., 2003), a human annotated collection of 183 newswire texts (including 7,935 annotated EVENTS, comprising both punctual *occurrences* and *states* which extend over time) and the AQUAINT corpus, with 80 newswire documents grouped by their covered stories, which allows tracing progress of events through time (Derczynski, 2017). Both corpora, supplemented with a large, automatically TimeML-annotated training corpus are used in the TempEval-3 task (UzZaman et al., 2013; Verhagen and Pustejovsky, 2008), which targets automatic identification of temporal expressions, events, and temporal relations. Whereas the ACE dataset provides annotations for entities, the relations between them, and for events in which they participate in newspaper and newswire text. For

[4]Event expressions are not, however, restricted to verbs: adjectives, nominalisations or prepositional phrases can also act as event triggers (consider, e.g., *Two weeks after the murder took place...*, *Following the recent acquisition of the company's assets...*).

		Train	Test
TempEval	English	830,005	7,174
	Spanish	51,511	5,466
	Chinese	23,180	5,313
ACE	English	529	40
	Chinese	573	43
	Arabic	356	27

Table 6.2 Number of tokens (TempEval) and documents (ACE) in the training and test sets.

each event, it identifies its lexical instantiation, i.e., the *trigger*, and its participants, i.e., the *arguments*, and the roles they play in the event. For example, an event type "Conflict:Attack" ("It could swell to as much as $500 billion if we go to war in Iraq."), triggered by the noun "war", involves two arguments, the "Attacker" ("we") and the "Place" ("Iraq"), each of which is annotated with an entity label ("GPE:Nation").

In this study, I rely on the TimeML-annotated corpus from *TempEval* tasks (UzZaman et al., 2013; Verhagen et al., 2010), which targets automatic identification of temporal expressions, events, and temporal relations. Then, I use the ACE dataset which provides annotations for entities, the relations between them, and for events in which they participate in newspaper and newswire text.

Task 1: Trigger Identification and Classification (TempEval)

I frame the first event extraction task as a token-level classification problem, predicting whether a token triggers an event and assigning it to one of the following event types: OCCURRENCE (e.g., *died, attacks*), STATE (e.g., *share, assigned*), Reporting (*e.g., announced, said*), I-ACTION (e.g., *agreed, trying*), I-STATE (e.g., *understands, wants, consider*), ASPECTUAL (e.g., *ending, began*), and PERCEPTION (e.g., *watched, spotted*).[5] I use the TempEval-3 data for English and Spanish (UzZaman et al., 2013), and the TempEval-2 data for Chinese (Verhagen et al., 2010) (see Table 6.2 for dataset sizes).

Task 2: Trigger and Argument Identification and Classification (ACE)

In this sequence-labelling task, I detect and label event triggers and their arguments, with four individually scored subtasks: (i) trigger identification, where I identify the key

[5]E.g., in the sentence: *"The rules can also affect small businesses, which sometimes pay premiums tied to employees' health status and claims history."*, *affect* and *pay* are event triggers of type STATE and OCCURRENCE, respectively.

word conveying the nature of the event, and (ii) trigger classification, where I classify the trigger word into one of the predefined categories; (iii) argument identification, where I predict whether an entity mention is an argument of the event identified in (i), and (iv) argument classification, where the correct role needs to be assigned to the identified event arguments. I use the ACE data available for English, Chinese, and Arabic.[6]

Event extraction as specified in these two frameworks is a challenging, highly context-sensitive problem, where different words (most often verbs) may trigger the same type of event, and conversely, the same word (verb) can evoke different types of event schemata depending on the context. Adopting these tasks for the present experimental setup thus tests whether leveraging fine-grained knowledge of verbs' semantic-syntactic behaviour can improve models' reasoning about event-triggering predicates and their arguments.

Model Configurations

For each task, I compare the performance of the underlying "vanilla" BERT-based model (see §6.2.3) against its variant with an added VN-Adapter or FN-Adapter[7] (see §6.2.2) in two regimes: (a) full fine-tuning, and (b) task adapter (TA) fine-tuning (see Figure 6.1 again). To ensure that any performance gains are not merely due to increased parameter capacity offered by the adapter module, I evaluate an additional setup where I replace the knowledge adapter with a randomly initialised adapter module of the same size (*+Random*). Additionally, I examine the impact of increasing the capacity of the trainable task adapter by replacing it with a *'Double Task Adapter'* (2TA), i.e., a task adapter with double the number of trainable parameters compared to the base architecture described in §6.2.2. Further, I compare the VN-/FN-Adapter approach with a computationally more expensive alternative method for injecting external verb knowledge, *sequential fine-tuning*, where the full BERT is first fine-tuned on the FN/VN data (as in 6.2.2) and then on the task (see below for fine-tuning details).

[6]The ACE annotations distinguish 34 trigger types (e.g., *Business:Merge-Org*, *Justice:Trial-Hearing*, *Conflict:Attack*) and 35 argument roles. Following previous work (Hsi et al., 2016), I conflate eight time-related argument roles – e.g., 'Time-At-End', 'Time-Before', 'Time-At-Beginning' – into a single 'Time' role in order to alleviate training data sparsity.

[7]I also experimented with inserting both verb adapters simultaneously; however, this resulted in weaker downstream performance than adding each separately, a likely product of the partly redundant, partly conflicting information encoded in these adapters (see §6.2.1 for comparison of VN and FN).

Training Details

Verb Adapters. I experimented with $k \in \{2, 3, 4\}$ negative examples and the following combinations of controlled (c) and randomly (r) sampled negatives (see §6.2.2): $k = 2$ [cc], $k = 3$ [ccr], $k = 4$ [$ccrr$]. In preliminary experiments I found the $k = 3$ [ccr] configuration to yield best-performing adapter modules. The downstream evaluation and analysis presented in §6.4 is therefore based on this setup.

The VN- and FN-Adapters are injected into the cased variant of the BERT Base model. Following Pfeiffer et al. (2020a), I train the adapters for 30 epochs using the Adam algorithm (Kingma and Ba, 2015), a learning rate of $1e-4$ and the adapter reduction factor of 16 (Pfeiffer et al., 2020a), i.e., $d = 48$. The training batch size is 64, comprising 16 positive examples and $3 \times 16 = 48$ negative examples (since $k = 3$). I provide more details on hyperparameter search in Appendix E.1.

Sequential Fine-Tuning. In the sequential fine-tuning setup, I first train the full cased variant of the BERT-based model on the VN/FN data. I generate negative examples using the strongest performing configuration of sampling parameters: $k = 3$ [ccr]. I train the model for 4 epochs using the Adam algorithm (Kingma and Ba, 2015), a learning rate of $2e-5$ with 1000 warmup steps and a batch size of 64. Next, I fine-tune the VN-/FN-pretrained model on the two downstream tasks, as outlined below.

Downstream Task Fine-Tuning

In downstream fine-tuning on Task 1 (TempEval), I train for 10 epochs in batches of size 32, with a learning rate $1e-4$ and maximum input sequence length of $T = 128$ WordPiece tokens. In Task 2 (ACE), in light of a greater data sparsity,[8] I search for an optimal hyperparameter configuration for each language and evaluation setup from the following grid: learning rate $l \in \{1e-5, 1e-6\}$ and epochs $n \in \{3, 5, 10, 25, 50\}$ (with maximum input sequence length of $T = 128$).

Transfer Experiments

For zero-shot (ZS) transfer experiments, I leverage mBERT, to which I add the VN- or FN-Adapter trained on the English VN/FN data. I train the model on English training data available for each task and evaluate it on the test set in the target language.

[8] Most event types in ACE ($\approx 70\%$) have fewer than 100 labelled instances, and three have fewer than 10 (Liu et al., 2018).

		FFT	+RAND	+FN	+VN	+FN$_{seq}$	+VN$_{seq}$	**TA**	+RAND	+FN	+VN
TE	T-ID&CL	73.6	73.5	73.6	73.6	**74.2**	73.9	74.5	74.4	**75.0**	**75.2**
ACE	T-ID	69.3	69.6	**70.8**	**70.3**	70.0	69.8	65.1	65.0	**65.7**	**66.4**
	T-CL	65.3	65.5	**66.7**	**66.2**	65.4	65.4	58.0	58.5	**59.5**	**60.2**
	ARG-ID	33.8	33.5	34.2	**34.6**	**36.3**	**36.2**	2.1	1.9	2.3	2.5
	ARG-CL	31.6	31.6	**32.2**	**32.8**	**34.3**	**33.9**	0.6	0.6	0.8	0.8

Table 6.3 Results for full fine-tuning (**FFT**) and the task adapter (**TA**) setup on English TempEval (TE; trigger identification and classification (T-ID&CL)) and ACE test sets (four subtasks: trigger (T) and argument (ARG) identification (ID) and classification (CL)). Provided are average F1 scores over 10 runs. Statistically significant (paired t-test; $p < 0.05$) improvements over both baselines marked in bold; the same labelling is also used in all subsequent tables.

For the VTRANS approach (see §6.2.4), I use language-specific BERT models readily available for the selected target languages, and leverage target-language adapters trained on translated and automatically refined verb pairs. The model, with or without the target-language VN-/FN-Adapter, is trained and evaluated on the training and test data available in the language. I carry out the procedure for three target languages (see Table 6.1). I use the same negative sampling parameter configuration proven strongest in the English experiments ($k = 3$ [*ccr*]). The same procedure is followed in the sequential fine-tuning setup, but instead of using adapter modules, the full mBERT or language-specific BERT models are first trained on the English or target language VN/FN data, respectively (as described in §6.3), and then evaluated on the task.

6.4 Results and Discussion

6.4.1 Main Results

English Event Processing

Table 6.3 shows the performance on English Task 1 (TempEval) and Task 2 (ACE). First, I note that the computationally more efficient setup with a dedicated task adapter (TA) yields higher absolute scores compared to full fine-tuning (FFT) on TempEval. When the underlying BERT is frozen along with the added FN-/VN-Adapter, the TA is enforced to encode additional task-specific knowledge into its parameters, beyond what's provided in the verb adapter, which results in two strongest results overall from the +FN/VN setups. In Task 2, the primacy of TA-based training is overturned in favour of full fine-tuning. Encouragingly, boosts provided by verb adapters are visible regardless of the chosen task fine-tuning regime, that is, regardless

		FFT	+RAND	+FN	+VN	+FN$_{seq}$	+VN$_{seq}$	**TA**	+RAND	+FN	+VN
ES	MBERT-ZS	37.2	37.2	37.0	36.6	37.3	37.4	38.0	38.0	**38.6**	36.5
	ES-BERT	77.7	77.1	77.6	77.4	77.8	77.6	70.0	70.0	**70.7**	**70.6**
	ES-MBERT	73.5	73.6	**74.4**	**74.1**	73.3	73.2	65.3	65.4	65.8	**66.2**
ZH	MBERT-ZS	49.9	49.9	**50.5**	47.9	**51.4**	50.0	49.2	49.5	**50.1**	48.2
	ZH-BERT	82.0	81.6	81.8	81.8	82.3	82.2	76.2	76.3	75.9	**76.9**
	ZH-MBERT	80.2	80.1	79.9	80.0	80.1	79.1	71.8	71.8	72.1	71.9

Table 6.4 Results on Spanish and Chinese TempEval test sets for full fine-tuning (**FFT**) and the task adapter (**TA**) setup, for zero-shot (**zs**) transfer with **mBERT** and monolingual target language evaluation with language-specific BERT (**ES-BERT / ZH-BERT**) or mBERT (**ES-mBERT / ZH-mBERT**), with FN/VN adapters trained on VTRANS-translated verb pairs (see §6.2.4). F1 scores are averaged over 10 runs, with statistically significant (paired t-test; $p < 0.05$) improvements over both baselines marked in bold.

of whether the underlying BERT's parameters remain fixed or not. I notice consistent statistically significant[9] improvements in the +VN setup, although the performance of the TA-based setups clearly suffers in argument (ARG) tasks due to decreased trainable parameter capacity. Lack of visible improvements from the Random Adapter supports the interpretation that performance gains indeed stem from the added useful 'non-random' signal in the verb adapters. In addition, I verify how the setup with added adapter modules compares to an alternative established approach, sequential fine-tuning (+FN$_{seq}$/VN$_{seq}$). In TempEval, fine-tuning all model parameters on VN/FN data allows retrieving more additional verb knowledge beneficial for task performance than adding smaller pretrained adapters on top of the underlying model. However, FN$_{seq}$/VN$_{seq}$ scores are still inferior to the results achieved in the TA-based +FN/VN setup. In ACE, the FN$_{seq}$/VN$_{seq}$ results in trigger tasks are weaker than those achieved through the addition of self-contained knowledge adapters, however, they offer additional boosts in argument tasks.

Multilingual Event Processing

Table 6.4 compares the performance of zero-shot (ZS) transfer and monolingual target-language training (via the VTRANS approach) on TempEval in Spanish and Chinese. For both the addition of the FN-Adapter in the TA-based setup boosts zero-shot transfer. Benefits of injecting verb knowledge from FrameNet extend to the full fine-tuning setup in Chinese, where the +FN setups achieve the top scores overall.

[9] I test significance with the Student's t-test with a significance value set at $\alpha = 0.05$ for sets of model F1 scores.

			FFT	+Rand	+FN	+VN	**TA**	+Rand	+FN	+VN
AR	mBERT-zs	T-id	15.8	13.5	**17.2**	16.3	29.4	30.3	**32.9**	**32.4**
		T-cl	14.2	12.2	**16.1**	**15.6**	25.6	26.3	**27.8**	**28.4**
		ARG-id	1.2	0.6	**2.1**	**2.7**	2.0	3.3	3.3	3.6
		ARG-cl	0.9	0.4	1.5	**1.9**	1.2	1.6	1.6	1.3
	AR-BERT	T-id	68.8	68.9	**70.2**	68.6	24.0	21.3	**24.6**	23.5
		T-cl	63.6	62.8	**64.4**	62.8	22.0	19.5	**23.1**	22.3
		ARG-id	31.7	29.3	**34.0**	**33.4**	–	–	–	–
		ARG-cl	28.4	26.7	**30.3**	**29.7**	–	–	–	–
	AR-mBERT	T-id	64.9	65.2	65.1	65.6	20.7	18.0	**23.2**	19.5
		T-cl	56.2	56.5	**57.4**	56.2	14.4	14.0	**16.5**	14.5
		ARG-id	25.4	25.4	**27.2**	24.6	–	–	–	–
		ARG-cl	21.3	21.9	**23.0**	19.9	–	–	–	–
ZH	mBERT-zs	T-id	36.9	36.7	**42.1**	36.8	47.8	49.4	**55.0**	**55.4**
		T-cl	27.9	25.2	**30.9**	**29.8**	38.6	40.1	**43.5**	**44.9**
		ARG-id	4.3	3.1	**5.5**	**6.1**	5.1	6.0	**7.6**	**8.4**
		ARG-cl	3.9	2.7	**4.9**	**5.2**	3.5	4.7	**5.7**	**7.1**
	ZH-BERT	T-id	75.5	74.9	74.5	74.9	69.8	69.3	70.0	70.2
		T-cl	67.9	68.2	68.0	68.6	58.4	57.5	**59.9**	**60.0**
		ARG-id	27.3	26.1	**29.8**	**28.8**	–	–	–	–
		ARG-cl	25.8	25.2	**28.2**	**27.2**	–	–	–	–
	ZH-mBERT	T-id	74.1	74.4	74.0	73.3	62.2	62.6	**63.8**	62.5
		T-cl	62.9	62.9	**64.3**	**63.6**	52.4	52.2	**54.3**	**54.0**
		ARG-id	26.2	26.3	**27.2**	**28.0**	–	–	–	–
		ARG-cl	24.8	25.3	**26.2**	**26.4**	–	–	–	–

Table 6.5 Results on Arabic and Chinese ACE test sets for full fine-tuning (**FFT**) setup and task adapter (**TA**) setup, for zero-shot (zs) transfer with **mBERT** and vtrans transfer approach with language-specific BERT (**AR-BERT / ZH-BERT**) or mBERT (**AR-mBERT / ZH-mBERT**) and FN/VN adapters trained on noisily translated verb pairs (§6.2.4). F1 scores averaged over 5 runs; significant improvements (paired t-test; $p < 0.05$) over both baselines marked in bold.

In monolingual evaluation, I observe consistent gains from the added transferred knowledge (i.e., the vtrans approach) in the TA-based setup in Spanish, while in Chinese performance boosts come from the transferred VerbNet-style class membership information (+VN). These results suggest that even the noisily translated verb pairs carry enough useful signal through to the target language. To tease apart the contribution of the language-specific encoders and transferred verb knowledge to task performance, I carry out an additional monolingual evaluation substituting the monolingual target language BERT with the massively multilingual encoder, trained on the (noisy) target language verb signal (ES-mBERT/ZH-mBERT). Notably, although the performance of the massively multilingual model is lower than the language-specific

BERTs in absolute terms, the addition of the transferred verb knowledge helps reduce the gap between the two encoders, with tangible gains achieved over the baselines in Spanish (see discussion in §6.4.2).

In ACE, the top performance scores are achieved in the monolingual full fine-tuning setting. As seen in English, keeping the full capacity of BERT parameters unfrozen noticeably helps performance.[10] In Arabic, FN knowledge provides performance boosts across the four tasks and with both the zero-shot (ZS) and monolingual (VTRANS) transfer approaches, whereas the addition of the VN adapter boosts scores mainly in ARG tasks. The usefulness of FN knowledge extends to zero-shot transfer in Chinese, and both adapters benefit the ARG tasks in the monolingual (VTRANS) transfer setup. Monolingual fine-tuning using the multilingual BERT shows an analogous pattern to the results on TempEval, where it lags behind the setups leveraging the language-specific models; still, it visibly benefits from the added transferred verb knowledge, especially from FrameNet, which consistently supports both Arabic and Chinese tasks. Notably, in zero-shot transfer, I observe that the highest scores are achieved in the task adapter (TA) fine-tuning, where the inclusion of the knowledge adapters offers additional performance gains. Overall, however, the argument tasks elude the restricted capacity of the TA-based setup, with very low scores.

Comparing the adapter-based setups with knowledge injection by means of sequential fine-tuning, one can observe advantages of using the full capacity of BERT parameters to encode verb knowledge in most setups in TempEval, whereas in ACE the adapter-based approach proves stronger overall (see Appendix E.3 for the sequential fine-tuning results on ACE). While sequential fine-tuning is a strong verb knowledge injection variant, it is computationally more expensive and less portable. The modular and efficient adapter-based approach therefore presents an attractive alternative, while offering competitive task performance. Crucially, the strong results from the sequential setup further corroborate the core finding of this chapter that external lexical verb information is indeed beneficial for event processing tasks.

6.4.2 Further Discussion

Zero-shot Transfer vs. Monolingual Training

The results reveal a considerable gap between the performance of zero-shot transfer and monolingual fine-tuning. The event extraction tasks pose a significant challenge

[10] This is especially the case in ARG tasks, where the TA-based setup fails to achieve meaningful improvements over zero, even with extended training up to 100 epochs. Due to the computational burden of such long training, the results in this setup are limited to trigger tasks (after 50 epochs).

		2TA	+FN	+VN
English	EN-BERT	74.5	74.8	74.8
Spanish	mBERT-zs	37.7	**38.3**	37.1
	ES-BERT	73.1	**73.6**	**73.6**
Chinese	mBERT-zs	49.1	**50.1**	48.8
	ZH-BERT	78.1	78.1	**78.6**

Table 6.6 Results on TempEval for the Double Task Adapter-based approaches (**2TA**). Significant improvements (paired t-test; $p < 0.05$) in bold.

to the zero-shot transfer via mBERT, where downstream event extraction training data is in English. However, mBERT exhibits much more robust performance in the monolingual setup, when presented with training data for event extraction tasks in the target language – here it trails language-specific BERT models by less than 5 points (see Table 6.4). This is an encouraging result, given that LM-pretrained language-specific Transformers currently exist only for a narrow set of well-resourced languages: for all other languages – should there be language-specific event extraction data – one needs to resort to massively multilingual Transformers. What is more, mBERT's performance is further improved by the inclusion of transferred verb knowledge (the vtrans approach, see §6.2.4): in Spanish, where the greater typological vicinity to English (compared to Chinese) renders direct transfer of semantic-syntactic information more viable, the addition of verb adapters (trained on noisy Spanish constraints) yields significant improvements both in the FFT and the TA setup. These results confirm the effectiveness of lexical knowledge transfer (i.e., the vtrans approach) observed in previous work (Ponti et al., 2019; Wang et al., 2020b) in the context of semantic specialisation of static word embedding spaces.

Double Task Adapter

The addition of a verb adapter increases the parameter capacity of the underlying pretrained model. To verify whether increasing the number of trainable parameters in TA cancels out the benefits from the frozen verb adapter, I run additional evaluation in the TA-based setup, but with trainable task adapters double the size of the standard TA (**2TA**). Promisingly, Tables 6.6 and 6.7 reveal that the relative performance gains from FN/VN adapters are preserved regardless of the added trainable parameter capacity. As expected, the increased task adapter size helps argument tasks in ACE, where verb adapters produce additional gains. Overall, this suggests that verb adapters indeed

			2TA	+FN	+VN
EN	EN-BERT	T-ID	67.5	**68.1**	**68.9**
		T-CL	61.6	**62.6**	**62.7**
		ARG-ID	6.2	**8.9**	**7.1**
		ARG-CL	3.9	**6.7**	**5.0**
AR	MBERT-ZS	T-ID	31.2	**32.6**	31.7
		T-CL	26.3	**27.1**	**29.3**
		ARG-ID	5.9	6.0	**6.9**
		ARG-CL	3.9	4.1	4.3
	AR-BERT	T-ID	40.6	**42.3**	**43.0**
		T-CL	36.9	**38.1**	**39.5**
		ARG-ID	–	–	–
		ARG-CL	–	–	–
ZH	MBERT-ZS	T-ID	54.6	**56.3**	**58.1**
		T-CL	45.6	**46.2**	**46.9**
		ARG-ID	9.2	**10.8**	**11.3**
		ARG-CL	8.0	8.5	**9.9**
	ZH-BERT	T-ID	72.3	**73.1**	72.0
		T-CL	59.6	**63.0**	**61.3**
		ARG-ID	2.6	2.8	3.3
		ARG-CL	2.3	2.6	2.9

Table 6.7 Results on ACE for the Double Task Adapter-based approaches (**2TA**). Significant improvements (paired t-test; $p < 0.05$) in bold.

encode additional, non-redundant information beyond what's offered by the pretrained model alone, and boost the dedicated task adapter in solving the problem at hand.

Cleanliness of Verb Knowledge

Gains from verb adapters suggest that there is potential to find supplementary information within structured lexical resources that can support distributional models in tackling tasks where nuanced knowledge of verb behaviour is important. The fact that the best transfer performance is obtained through noisy translation of English verb knowledge suggests that these benefits transcend language boundaries.

There are, however, two main limitations to the translation-based (VTRANS) approach used to train the target-language verb adapters (especially in the context of VerbNet constraints): (1) noisy translation based on cross-lingual semantic similarity may already break the VerbNet class membership alignment (i.e., words close in meaning may belong to different VerbNet classes due to differences in syntactic behaviour); and (2) the language-specificity of verb classes, which cannot be directly ported to another language without adjustments due to the delicate language-specific

	FFT+FN$_{ES}$	**FFT**+FN$_{ESseq}$	**TA**+FN$_{ES}$	**2TA**+FN$_{ES}$
ES-BERT	78.0 (+0.4)	77.9 (+0.1)	70.9 (+0.1)	73.8 (+0.2)

Table 6.8 Results (F1 scores) on Spanish TempEval for different configurations of Spanish BERT with added Spanish FN-Adapter (**FN**$_{ES}$), trained on clean Spanish FN constraints. Numbers in brackets indicate relative performance w.r.t. the corresponding setup with FN-Adapter trained on (a larger set of) noisy Spanish constraints obtained through automatic translation of verb pairs from English FN (VTRANS approach).

interplay of semantic and syntactic information. This is in contrast to the proven cross-lingual portability of synonymy and antonymy relations shown in previous work on semantic specialisation transfer (Mrkšić et al., 2017; Ponti et al., 2019), which rely on semantics alone. In the case of VerbNet, despite the cross-lingual applicability of the semantically-syntactically defined verb class as a lexical organisational unit, the fine-grained class divisions and exact class membership may be too English-specific to allow direct automatic translation. On the contrary, semantically driven FrameNet lends itself better to cross-lingual transfer, given that it focuses on function and roles played by event participants, rather than their surface realisations (see §6.2.1). Indeed, although FN and VN adapters both offer performance gains in the presented evaluation, the higher average improvements in cross-lingual setups with the FN-Adapter may be symptomatic of the resource's greater cross-lingual portability.

To quickly verify if noisy translation and direct transfer from English curb the usefulness of injected verb knowledge, I additionally evaluate the injection of *clean* verb knowledge obtained from a small lexical resource available in one of the target languages – Spanish FrameNet (Subirats and Sato, 2004). Using the procedure described in §6.2.2, I derive 2,886 positive verb pairs from Spanish FN and train a Spanish FN-Adapter (on top of the Spanish BERT) with this (much smaller but clean) set of Spanish FN constraints.

The results in Table 6.8 show that, despite having 12 times fewer positive examples for training the verb adapter compared to the translation-based approach, the 'native' Spanish verb adapter outperforms its VTRANS-based counterpart (Table 6.4), compensating the limited coverage with gold standard accuracy. Nonetheless, the challenge for using native resources in other languages lies in their very limited availability and expensive, time-consuming manual construction process. The results presented in this chapter reaffirm the usefulness of language-specific manually created resources and their ability to enrich state-of-the-art NLP models. This, in turn, suggests that work

on optimising resource creation methodologies merits future research efforts on a par with modelling work.

Expert vs. Non-expert Verb Knowledge

Given the demonstrated benefits of injecting in-target lexical information and the simultaneous limited availability of expert-curated resources in most languages worldwide, in the final evaluation of this chapter I investigate the potential of harnessing non-expert semantic knowledge as an alternative. Table 6.9 summarises the results for English and Chinese BERT infused with verb class information derived from the datasets introduced in this thesis. In English, we can observe small but consistent improvements over the baselines and competitive, if slightly weaker, performance compared to that offered by expert datasets (Table 6.3). Notably, in TempEval, harnessing the information about verb membership in narrow semantic similarity-based clusters derived from the spatial arrangements (Phase 2) in a dedicated adapter proves most beneficial, offering a significant improvement over the baseline on a par with that achieved by the FN adapter (TA +FN in Table 6.3). In ACE, the benefits of the knowledge derived from SpA-Verb classes are even more evident, especially in argument tasks (+1.7 over the FFT baseline in Table 6.3), where the SpA-Adapter substantially boosts the TA-based setup (+5.4 on argument identification and +4.3 on argument classification). Notably, even bigger boosts ($\geq$ 2.0 for FFT +SpA$_{seq}$, +5.7 on argument identification for TA +SpA) in argument tasks are offered by the added knowledge of semantic similarity between the most proximate verbs in Phase 2, despite the more limited volume of the training data (5,102 pairs of cluster-sharing verbs). This is a promising finding, given that the granularity of clusters obtained from the spatial data can be flexibly modified, depending on the nature of the task. Thus, in tasks where distinguishing between synonymy and antonymy is key (e.g., question answering; see discussion in §2.2.2), narrower clusters of synonymous and near-synonymous verbs arranged close together in the arena will be appropriate, yielding positive training instances of only semantically similar words. In turn, in applications where it is beneficial for the model to have a notion of broader relatedness of verbs (e.g., text summarisation), deriving training constraints based on shared membership in bigger semantic clusters would likely be preferable.

As could be expected, the top absolute scores in English are achieved by the model boosted with expert-curated knowledge from FrameNet and VerbNet (Table 6.3), rather than the semantic information derived from non-experts (Table 6.9). However, the significant improvements offered by the injection of the latter type of

				FFT+SpA	**FFT**+SpA$_{seq}$	**TA**+SpA
TempEval	EN-BERT	*Baseline*	T-ID&CL	*73.6*	*73.6*	*74.5*
		Phase 1	T-ID&CL	74.0	73.9	74.7
		Phase 2	T-ID&CL	74.0	73.9	**75.0**
	ZH-BERT	*Baseline*	T-ID&CL	*82.0*	*82.0*	*76.2*
		Phase 1	T-ID&CL	82.1	**82.5**	76.3
ACE	EN-BERT	*Baseline*	T-ID	*69.3*	*69.3*	*65.1*
			T-CL	*65.3*	*65.3*	*58.0*
			ARG-ID	*33.8*	*33.8*	*2.1*
			ARG-CL	*31.6*	*31.6*	*0.6*
		Phase 1	T-ID	**70.2**	**70.3**	**66.2**
			T-CL	65.9	65.8	**61.0**
			ARG-ID	**35.3**	**35.5**	**7.5**
			ARG-CL	**32.9**	**33.3**	**4.9**
		Phase 2	T-ID	69.9	69.4	**66.4**
			T-CL	65.2	64.9	**60.9**
			ARG-ID	**34.9**	**35.9**	**7.8**
			ARG-CL	**32.6**	**33.6**	**4.9**
	ZH-BERT	*Baseline*	T-ID	*75.5*	*75.5*	*69.8*
			T-CL	*67.9*	*67.9*	*58.4*
			ARG-ID	*27.3*	*27.3*	–
			ARG-CL	*25.8*	*25.8*	–
		Phase 1	T-ID	75.5	**76.1**	**70.8**
			T-CL	68.7	**69.3**	**60.2**
			ARG-ID	**28.0**	26.7	–
			ARG-CL	**26.7**	25.5	–

Table 6.9 Results on English and Chinese TempEval and ACE (F1 scores, averaged over 10 runs) for the three configurations of a language specific BERT with added verb knowledge from the SpA-Verb classes (Phase 1) in each language. Additionally, included is a setup fine-tuned on the lexical constraints from narrow semantic clusters (Phase 2) derived from the spatial similarity data in English. Statistically significant (paired t-test; $p < 0.05$) improvements over the FFT/TA baselines (in italics; see also Tables 6.3, 6.4 and 6.5) marked in bold.

knowledge show that it nonetheless complements the underlying architecture with useful non-distributional information. This finding reveals that there is potential to enrich the lexical information encoded in model parameters by harnessing native-speaker intuitions about verb meaning. This could prove especially valuable in scenarios where lexical resources like FrameNet or VerbNet are unavailable. The results of evaluation on Chinese TempEval and ACE provide further corroborating evidence. Notably, Chinese BERT sequentially fine-tuned on Chinese SpA-Verb data outperforms all transfer setups on TempEval ($F_1 = 82.5 >$ FFT+FN/VN$_{seq}$ in Table 6.4) achieving a

significant improvement over the FFT baseline. This not only shows that the in-target knowledge is valuable (as observed previously in Spanish), but also that the source of this information need not necessarily be an expert-created lexicon. Further, while automatic cross-lingual transfer avoids the need for human data collection altogether, leveraging native-speaker verb knowledge gathered in the target language may allow the model to better capture the language-specific aspects of verb behaviour.

The benefits of SpA-Verb information extend to Chinese ACE, where the sequential setup again achieves the top scores on trigger tasks, however, at the expense of argument tasks, where adding a pretrained SpA-Adapter proves to be a stronger setup. Interestingly, in the case of argument identification and classification, the VTRANS-based transfer approach is superior overall (Table 6.5). The reason behind this pattern most likely lies in the nature of SpA-Verb classes, which are semantically driven and do not explicitly rely on similarity in argument structure as a membership criterion. Although the information about shared SpA-Verb class membership benefits the model's ability to make predictions about both event triggers and arguments, it provides less implicit guidance on verbs' argument-taking properties. In turn, transferring this information from English to Chinese automatically offers significant gains over the baselines (Table 6.5). This suggests that there is a non-negligible amount of knowledge about the patterns of verb behaviour with regard to their arguments that is cross-lingually applicable.

6.5 Related Work

6.5.1 Event Extraction

The cost and complexity of event annotation requires robust transfer solutions capable of making fine-grained predictions in the face of data scarcity. Traditional event extraction methods relied on hand-crafted, language-specific features (Ahn, 2006; Glavaš and Šnajder, 2015; Gupta and Ji, 2009; Hong et al., 2011; Li et al., 2013; Llorens et al., 2010) (e.g., POS tags, entity knowledge, morphological and syntactic information), which limited their generalisation ability and effectively prevented language transfer.

More recent approaches commonly resorted to word embedding input and neural text encoders such as recurrent nets (Duan et al., 2017; Nguyen et al., 2016b; Sha et al., 2018) and convolutional nets (Chen et al., 2015; Nguyen and Grishman, 2015), as well as graph neural networks (Nguyen and Grishman, 2018; Yan et al., 2019) and adversarial networks (Hong et al., 2018; Zhang and Ji, 2018). Like in most other NLP

tasks, most recent empirical advancements in event trigger and argument extraction tasks have been achieved through fine-tuning of LM-pretrained Transformer networks (Liu et al., 2020a; M'hamdi et al., 2019; Wadden et al., 2019; Wang et al., 2019a; Yang et al., 2019a).

Limited training data nonetheless remains an obstacle, especially when facing previously unseen event types. The alleviation of such data scarcity issues has been attempted through data augmentation methods – automatic data annotation (Araki and Mitamura, 2018; Chen et al., 2017b; Zheng, 2018) and bootstrapping for training data generation (Ferguson et al., 2018; Wang et al., 2019a). The recent release of the large English event detection dataset MAVEN (Wang et al., 2020c), with annotations of event triggers only, partially remedies training data scarcity. MAVEN also demonstrates that even the state-of-the-art Transformer-based models fail to yield satisfying event detection performance in the general domain. The fact that it is unlikely to expect datasets of similar size for other event extraction tasks (e.g., event argument extraction) and especially for other languages only emphasises the need for external event-related knowledge and transfer learning approaches, such as the ones introduced in this work.

Beyond event trigger (and argument)-oriented frameworks such as ACE and its light-weight variant ERE (Aguilar et al., 2014; Song et al., 2015), several other event-focused datasets exist which frame the problem either as a slot-filling task (Grishman and Sundheim, 1996) or an open-domain problem consisting in extracting unconstrained event types and schemata from text (Allan, 2002; Araki and Mitamura, 2018; Liu et al., 2019c; Minard et al., 2016). Small domain-specific datasets have also been constructed for event detection in biomedicine (Buyko et al., 2010; Kim et al., 2008; Nédellec et al., 2013; Thompson et al., 2009), as well as literary texts (Sims et al., 2019) and Twitter (Guo et al., 2013; Ritter et al., 2012).

6.5.2 Semantic Specialisation

Representation spaces induced through self-supervised objectives from large corpora, be it the word embedding spaces (Bojanowski et al., 2017; Mikolov et al., 2013b) or those spanned by LM-pretrained Transformers (Devlin et al., 2019; Liu et al., 2019d), encode only distributional knowledge, i.e., knowledge obtainable from large corpora. A large body of work focused on *semantic specialisation* (i.e., refinement) of such distributional spaces by means of injecting lexical-semantic knowledge from external resources such as WordNet (Fellbaum, 1998), BabelNet (Navigli and Ponzetto, 2010) or ConceptNet (Liu and Singh, 2004) expressed in the form of lexical constraints (Faruqui et al., 2015;

Glavaš and Vulić, 2018b; Kamath et al., 2019; Lauscher et al., 2020b; Mrkšić et al., 2017, *inter alia*).

Joint specialisation models (Lauscher et al., 2020b; Levine et al., 2020; Nguyen et al., 2017; Yu and Dredze, 2014, *inter alia*) train the representation space from scratch on the large corpus, but augment the self-supervised training objective with an additional objective based on external lexical constraints. Lauscher et al. (2020b) add to the Masked LM (MLM) and next sentence prediction (NSP) pretraining objectives of BERT (Devlin et al., 2019) an objective that predicts pairs of synonyms and first-order hyponymy-hypernymy pairs, aiming to improve word-level semantic similarity in BERT's representation space. In a similar vein, Levine et al. (2020) add the objective that predicts WordNet supersenses. While joint specialisation models allow the external knowledge to shape the representation space from the very beginning of the distributional training, this also means that any change in lexical constraints implies a new, computationally expensive pretraining from scratch.

Retrofitting and post-specialisation methods (Faruqui et al., 2015; Glavaš and Vulić, 2019; Lauscher et al., 2020a; Mrkšić et al., 2017; Ponti et al., 2018; Vulić et al., 2018; Wang et al., 2020a, *inter alia*), in contrast, start from a pretrained representation space (word embedding space or a pretrained encoder) and fine-tune it using external lexical-semantic knowledge. Wang et al. (2020a) fine-tune the pretrained RoBERTa (Liu et al., 2019d) with lexical constraints obtained automatically via dependency parsing, whereas Lauscher et al. (2020a) use lexical constraints derived from ConceptNet to inject knowledge into BERT: both adopt adapter-based fine-tuning, storing the external knowledge in a separate set of parameters. In this work, I adopt a similar adapter-based specialisation approach. However, focusing on event-oriented downstream tasks, the lexical constraints employed in the process reflect verb class memberships and originate from VerbNet and FrameNet, or the semantic datasets generated in this work.

6.6 Conclusion

In this chapter, I investigated the potential of leveraging knowledge about the semantic and semantic-syntactic behaviour of verbs to improve the capacity of large pretrained models to reason about events in diverse languages. I proposed an auxiliary pretraining task to inject information about verb class membership and semantic frame-evoking properties into the parameters of dedicated adapter modules, which can be readily employed in other tasks where verb reasoning abilities are key. The results of the evaluation on event processing tasks demonstrated that state-of-the-art Transformer-

based models still benefit from the gold standard linguistic knowledge stored in lexical resources, even those with limited coverage. Crucially, the benefits of the information available in resource-rich languages can be extended to other, resource-leaner languages through translation-based transfer of verb class/frame membership information. Finally, I explored the potential of leveraging native-speaker semantic judgments to steer the pretrained model to better capture verb meaning in the target language. Using the semantic classes created as part of this thesis, I demonstrated that the non-expert knowledge leads to consistent performance gains and offers a stronger alternative to the automatic cross-lingual transfer of expert knowledge from English to Chinese. Moreover, I showed that the pairwise verb similarities encoded in the matrices output by the second phase of the data collection paradigm presented in Chapter 4 offer further potential for deriving training data based on shared cluster membership, with the added benefit of flexibility, as cluster granularity can be decided based on the nature and specific demands of the task at hand.

Several promising directions for future research emerge from this chapter. First, while the present study relied on training verb adapters by means of a word-level binary classification task, future work could explore alternative methods for generating the training data. Although in the present work a sentence-level classification setup proved weaker than providing the model with verb pairs, future work should verify whether enriching the input sentences with additional information, such as syntactic dependencies, would make it more successful. Further, given the success of the presented knowledge injection methods, future experiments could incorporate the proposed verb adapter modules into alternative, more sophisticated approaches to cross-lingual transfer to explore the potential for further improvements in low-resource scenarios. What is more, the present approach could be readily extended to specialised domains where small-scale but high-quality lexicons are available, to support distributional models in dealing with domain-sensitive verb-oriented problems.

Chapter 7

Conclusions

In this chapter, I summarise the main findings of the research presented in this thesis and discuss its implications for future work. I begin by recapitulating the thesis's motivations and principal guiding themes: acquisition of verbal lexical information, model evaluation, and verb knowledge injection (Section 7.1). Then, I highlight the main contributions of the present work (Section 7.2) and proceed to discuss their implications in light of the emerging directions for future research (Section 7.3).

7.1 Motivation and Synopsis

The overarching motivation for the work presented in this thesis has been to address the gap between the lexical information encoded by distributional models and the intuitive linguistic knowledge possessed by language users, focusing on one of its facets, the lexical properties of verbs. The prominent role occupied by verbs in sentence structure makes them pivotal in the organisation of natural language and crucial for NLP systems to accurately capture. The current state-of-the-art representation learning paradigm, deep neural networks, typically based on Transformer architectures, has recently accelerated progress in language technology. However, despite their rapidly growing ability to solve NLP tasks, the state-of-the-art models have not yet overcome the fundamental limitation at their core: everything they learn about language is distributional in nature. Exposure to vast volumes of text has enabled them to acquire an impressive amount of linguistic information automatically (Ettinger, 2020; Hewitt and Manning, 2019; Rogers et al., 2020b; Tenney et al., 2019a). Nonetheless, despite being heavily parameterised, spanning billions of parameters, they still cannot capture all the subtleties of verb meaning and behaviour. Further, probing analyses have revealed that their sometimes super-human task performance owes more to their ability

to capitalise on superficial hints and correlations specific to a given dataset, rather than the depth of their understanding (Bender and Koller, 2020; McCoy et al., 2019; Niven and Kao, 2019). In contrast, humans manipulate the complexities of verb meaning intuitively and with ease, performing inferences about causality, temporal orders of events, and the relations between the actors and objects that partake in them. The command of language that native speakers possess allows them not only to make nuanced judgments about the grammaticality of linguistic constructions and the types of entities licensed by the verb to appear in them, but also to infer the permissible behaviour of unknown verbs based solely on their semantics (Hale and Keyser, 1987; Levin, 1993; Zwicky, 1971).

Given the importance of attaining accurate representations of verbal properties for successful machine language understanding, computational language resources targeting verbs can play an important role in facilitating further advances. External lexicons proved very important for previous statistical approaches, ranging from sparse vectors to Bayesian learning, but the degree of their utility for current state-of-the-art methods has not been fully explored. Lexical classes which organise verbs based on shared semantic and syntactic behaviour are an example of a powerful model of verb behaviour capturing the interrelatedness of their structural properties and meaning (Levin, 1993), with demonstrated potential to support diverse NLP tasks (Brown and Palmer, 2012; Clark et al., 2018; Lignos et al., 2015; Lippincott et al., 2013; Martin et al., 2018, *inter alia*). However, while a very small number of well-resourced languages, such as English, already boast several expert-built lexicons that record rich information about verbs (Baker et al., 1998; Kipper et al., 2006; Kipper Schuler, 2005), they are still few and far between. Further, their slow and expensive construction hinders resource creation initiatives in other languages and specialised domains. As a time-efficient alternative, automatic acquisition of verb classes has attracted attention (e.g., Kawahara et al., 2014; Peterson et al., 2016, 2020; Scarton et al., 2014; Vulić et al., 2017b). However, automatic approaches inevitably introduce noise into the classification and their development relies on the availability of human-generated gold standard data. In light of the demand for evaluation benchmarks, crowdsourcing has emerged as a faster and cheaper alternative to expert lexicographic work. Representation learning models have been routinely evaluated on datasets consisting of similarity scores collected from non-expert raters on pairs of words (e.g., Baroni et al., 2014; Dhillon et al., 2015; Levy and Goldberg, 2014; Pennington et al., 2014). However, most such datasets are restricted in size and focus predominantly on nouns. Moreover, their reliance on discrete rating scales and judgments collected on word pairs in isolation makes them prone to a number of

limitations (Batchkarov et al., 2016; Faruqui et al., 2016; Hout et al., 2013). Further, the ease of implementation of the pairwise rating-based annotation paradigm is only superficially matched by the ease of the task itself: in fact, quantifying subtle differences in word meaning and transposing them onto a discrete numerical scale is very difficult to do consistently, particularly in the absence of context or points of reference.

In this thesis, I proposed to circumvent these problems and address the demand for verb-focused lexical resources to evaluate, probe, and boost state-of-the-art representation models by leveraging native-speaker intuitions about verb meaning in a novel two-phase data collection protocol. Drawing on the observations about the nuanced ability of language users to reason about the properties and behaviour of verbs, I framed the knowledge acquisition problem as two consecutive tasks, manual semantic clustering and spatial arrangement of verbs within each cluster. The stepwise structure and modularity of the data collection design allowed me to scale up the problem to a large verb sample, while ensuring that similarity judgments are elicited in the context of all other verbs in the set and on comparable concepts (Turner et al., 1987). Next, I examined the cross-lingual portability of the method by applying it to a diverse selection of languages, which produced the first such large-scale multilingual resource composed of semantic verb classes and fine-grained verb similarity scores. I then demonstrated its utility as a challenging evaluation benchmark allowing for probing the representation quality across languages and domains of verb meaning, uncovering the strengths and weaknesses of the current models. Further, I proposed a method for supplementing large pretrained architectures with the verb knowledge stored in external resources. I showed that, despite the paradigm shift in representation learning, the knowledge stored in them is not obsolete and that feeding it into neural architectures provides them with useful bias and additional information which cannot be captured solely based on the distributional signal. Finally, I demonstrated the potential of harnessing non-expert linguistic insights for the benefit of model performance in downstream tasks, which holds promise for languages and domains lacking expert-curated resources.

7.2 Contributions and Findings

In the pursuit of the goals of this research, I conducted a series of experiments and analyses whose outcomes are summarised in this section, ordered by chapter of appearance.

7.2.1 Verb Class Induction Through Bottom-up Semantic Clustering

In Chapter 3, I investigated the potential to acquire verb classes directly from non-expert native speakers. Given the difficulty of simultaneously considering semantic and syntactic criteria for annotators lacking linguistics training (Majewska et al., 2018b) and the strong interrelatedness of these two types of information in verbs (Jackendoff, 1990; Levin, 1993), I proposed to simplify the task by letting semantic judgments alone guide the classification process.

- I carried out the first cross-lingual evaluation of semantic verb clustering by non-experts. I conducted verb clustering experiments in English, Polish, and Croatian. This allowed me to examine cross-lingual affinities in the clustering patterns within and across language families and study the overlap in the resultant classes within and across individual languages. The analysis of inter-annotator agreement in comparison to SemEval (Jurgens and Klapaftis, 2013) showed that the alignment in classes within each language surpasses the performance of the baselines, despite the greater difficulty of this task where verb senses and the target number of clusters were left unspecified. The encouraging degree of overlap in the classifications indicates that there are regularities in how humans without linguistics training categorise verb meaning and there is potential to create verb classes starting from a simple, purely semantic task.

- I analysed the cross-lingual alignment in the resultant classes to examine whether and to what degree similar meaning components are taken into consideration when classifying verbs across different languages. Again, the degree of alignment exceeded the baselines by a significant margin, suggesting meaningful correlations in cross-lingual clustering patterns. Further, I manually inspected the resultant clusters to identify the areas of verb meaning distinguished in all three languages. The total of thirty clusters where the core of at least two members is shared by at least two languages could be identified, suggesting there are cross-lingual commonalities governing the organisation of verb meaning. Cross-lingual discrepancies in class membership mainly result from variation in patterns of polysemy, where languages differ in the number of senses available for a given verb.

- I carried out an in-depth analysis of the cross-lingual clustering patterns to identify the factors contributing to the ease or difficulty of the task. To this end, I identified a subset of perfect-agreement and low-agreement verbs shared by all

three languages. This revealed that verbs that are difficult to cluster share the characteristic of having abstract or vague meaning, often with a number of related senses, or display a degree of semantic vacuity. In contrast, verbs which proved easiest to classify were those with narrow and concrete meanings, belonging to a well-defined semantic field. This group included antonyms, consistently clustered together, which reflects their paradigmatic similarity.

7.2.2 Semantic Dataset Construction from Clustering and Spatial Arrangement

In Chapter 4, I introduced a novel two-phase semantic data collection methodology based on semantic clustering and subsequent spatial arrangements in a two-dimensional space. I carried out an extensive and in-depth analysis of the data output by each phase and conducted a comprehensive evaluation of representation learning models on two tasks corresponding to the two phases of the design, semantic word clustering and word similarity.

- I designed a two-phase protocol for the collection of semantic judgments on large samples of words. In particular, I adapted a spatial multi-arrangement method used previously to record judgments of visual stimuli in psychology to a sample of polysemous lexical stimuli seven times as numerous as the largest stimuli sets used hitherto in SpAM research. To tackle the challenges of scale and lexical ambiguity, I proposed a precursor clustering task which splits the large initial word sample into smaller theme classes, easily accommodated by standard computer screens, which provide disambiguating context for each member verb. Thanks to a label copying functionality, the users can duplicate ambiguous verb labels in Phase 1 and place distinct senses in separate classes, grouping each with related and similar verbs. This guarantees that in Phase 2 the sense of each verb is implied by the related verbs appearing in the same space, thus preventing large mismatches in judgments on ambiguous words. Further, the protocol avoids conflating similarity scores for antonyms and completely unrelated concepts, as the latter are separated in Phase 1.

- I performed cluster analysis on the output of Phase 1, applying a network analysis approach to the semantic clustering data in order to scrutinise the emerging semantic classes and gain insight into annotator decisions. I argued that identification of prototypical members within clusters and verbs acting as

inter-class links can be particularly useful for tracing patterns of polysemy and creation of comprehensive verb resources from clustering data. This analysis informed the selection of semantic classes used as input to Phase 2, which ensured that fine-grained semantic similarity judgments are made on comparable concepts, classified as similar or related in Phase 1.

- I presented an in-depth examination of the semantic information captured by the two phases, semantic clustering and spatial arrangements of lexical stimuli, by means of exploratory data analysis and qualitative and quantitative comparisons with three lexical resources, FrameNet, VerbNet, and WordNet. These analyses revealed that annotators are able to differentiate between a range of semantic relations by means of relative item placements. What is more, the encouraging overlap observed with semantic-syntactic VerbNet classes suggests that the method could help incorporate new verbs into the existing dataset, or support the creation of similar resources from scratch for other languages. I demonstrated this potential analysing the structure of concepts emerging from the spatial similarity data by means of hierarchical agglomerative clustering.

- I demonstrated the utility of the created resource by evaluating a selection of representation models on two tasks, semantic clustering and word similarity, and illustrated its potential to enable nuanced, focused analyses targeting specific semantic properties and meaning domains. In particular, the analyses revealed the primacy of static word embeddings incorporating external linguistic knowledge over state-of-the-art unsupervised Transformer-based architectures on both word-level semantic similarity and clustering. These findings provide additional evidence in support of the vast potential of drawing on external linguistic information to help vector representations better reflect fine-grained semantic relations present in the mental lexicon. Thanks to the dataset's large coverage, it is possible to contrast the performance of embeddings specialised for VerbNet and FrameNet classification information on focused semantic domains.

7.2.3 Verb Knowledge Acquisition for Multilingual Evaluation

In Chapter 5, I evaluated the cross-lingual applicability of the two-phase data collection method and its potential to facilitate model evaluation and dataset creation in typologically diverse languages lacking large-scale lexical resources.

- I carried out large-scale data collection for a diverse selection of languages representing five language families: Sino-Tibetan (Mandarin Chinese), Japonic (Japanese), Uralic (Finnish), Slavic (Polish) and Romance (Italian), which produced the first such multilingual dataset targeting verb semantics. The process started from word-by-word translation of the English sample used in Chapter 4 into each target language, which avoids many of the problems faced by methods which translate sets of word *pairs* due to variation in cross-lingual lexicalisation patterns. Further, by allowing one-to-many and many-to-one translations, the ultimate verb samples reflect the lexical distribution in the target language, rather than accommodating the semantic distinctions present in the source lexicon.

- I conducted cross-lingual analyses of similarities and variation in the data produced in each phase, quantifying the degree of overlap in the semantic classes and the correlation between selected semantic dissimilarity matrices for all pairings of languages. This analysis revealed that there are strong parallels in the semantic distinctions made in Phase 1 and statistically significant correlations between the pairwise distances recorded in Phase 2 in different languages, corroborating the hypothesis that many of the meaning components involved in the categorisation of verbs are cross-lingually shared. Further, in-depth inspection of the distribution of concepts pertaining to the domains of emotion and rate of change by means of Principal Coordinates Analysis allowed me to identify the most salient dimensions underlying the organisation of verb meaning across languages and language-specific differences in the treatment of particular verb types. These analyses indicate that the collected data may support fine-grained lexical-typological analyses in future work, enabling cross-lingual comparisons of the organisation of different semantic fields and the dimensions within each conceptual space.

- In order to investigate the potential of the method to support multilingual creation of verb taxonomies, I measured the degree of alignment between the semantically driven Phase 1 classes and VerbNet-style classifications from previous work. The results revealed substantial overlap between the two types of partitions in all languages, providing further evidence in support of the strong inter-connectedness of verb meaning and syntactic behaviour and the feasibility of 'bootstrapping' the classification process from semantic information alone. This is a promising finding, revealing that non-expert semantic judgments could be leveraged to aid the construction of verb classes, subsequently refined based on syntactic criteria.

- To assess the utility of the produced resource for supporting the development of NLP models, I carried out an evaluation of static and contextualised representation models on the tasks of lexical similarity and semantic clustering using the data from both phases, probing representation quality across languages and meaning domains. This revealed a split pattern in model performance, where static models proved superior to contextual encoders in languages using the Latin script, while the opposite was true for the languages relying on logographic scripts. The low absolute scores showed that capturing fine-grained distinctions between a large number of semantically proximate concepts is a very challenging task for state-of-the-art architectures. The vast coverage of the produced dataset and its dual nature promise to facilitate focused evaluation of semantic models, helping identify their shortcomings and thus enabling further advances.

7.2.4 Verb Knowledge Injection for Multilingual Event Processing

I concluded this thesis by exploring the potential of incorporating external verb knowledge into deep neural models to improve their linguistic capacity. Given the crucial role played by verbs in expressions of events and establishing the relations between their participants, in Chapter 6, I chose event processing tasks in English, Spanish, Chinese, and Arabic as a testbed for the proposed verb knowledge injection method.

- I proposed a method which enables the integration of verb class membership information into pretrained encoders by encapsulating the knowledge available in a lexical resource in a dedicated adapter module. This allows for flexible and seamless combination with other types of knowledge while reducing the risk of catastrophic forgetting of the distributional information encoded in the model's original parameters. To encode the information about verbs' semantic-syntactic properties into adapter layers, I devised an intermediary training task consisting in predicting whether two verbs share a class or evoke the same semantic frame. Using lexical constraints derived from VerbNet and FrameNet, I injected verb knowledge into contextualised representations for the first time.

- Given that rich lexical resources are only available in a small group of languages, I investigated the potential of transferring the structured knowledge available in English to other languages automatically, drawing on the hypothesis about the

cross-lingual nature of verb classes and semantic frames. To this end, I proposed two verb knowledge transfer techniques: (i) zero-shot model transfer based on a massively multilingual encoder using the original English lexical constraints for verb-oriented training, or (ii) automatic translation and refinement of English constraints in the target language, followed by fully monolingual training.

- I evaluated the proposed monolingual and cross-lingual approaches on the tasks of event trigger and argument identification and classification, which strongly revolve around verbs. Performance boosts achieved in English verified that injection of verb knowledge can indeed complement the distributional information available in model parameters and result in improved understanding of events and their structure. This was further corroborated in a follow-up monolingual experiment in Spanish, where using a small set of native FrameNet constraints yielded consistent performance gains. Moreover, the results of cross-lingual transfer experiments revealed benefits of incorporating verb-specific information in both setups, indicating that there is a substantial amount of source verb knowledge that is cross-lingually relevant.

- Finally, given that automatic transfer inevitably introduces noise into the training data, I evaluated the potential of leveraging non-expert linguistic knowledge available in the target language instead. Using the semantic classes and clusters derived from the spatial arrangement data collected as part of this thesis, I examined, for the first time, whether human semantic judgments can be harnessed for the purposes of refining latent representation spaces with verb-specific semantic information. The experiments in English and Chinese demonstrated that non-expert data indeed provide useful information complementary to what the model automatically acquires during pretraining from large text corpora. Further, they proved to be more beneficial to model performance in the target language than the automatically transferred expert knowledge. This finding has important implications, suggesting that quick bottom-up generation of verb knowledge based on non-expert judgments can mitigate the issue of resource scarcity and boost the capacity of distributional models to reason about verb meaning.

7.3 Implications and Future Directions

The above summarised findings have a number of implications for the three main research threads of this thesis: resource creation, model evaluation, and knowledge

injection. In this section, I discuss them in light of possible extensions of the presented work and contemplate future research pathways.

7.3.1 Data Collection for Model Evaluation, Resource Construction and Linguistic Analyses

The analyses presented in this thesis revealed that the proposed two-phase dataset creation protocol offers a number of benefits as a means of rapid generation of lexical knowledge for evaluation of semantic models. Considering the ongoing progress in representation learning, challenging evaluation datasets are essential to enable scrutinising the linguistic ability of state-of-the-art architectures. Crucially, they can help improve the interpretability of neural models, as representation quality can be probed across narrow semantic domains, thus aiding researchers in identifying the factors contributing to or degrading downstream task performance. The success of the presented method which generates such data at a large scale in a short time and at a low cost has important implications especially for extending the reach of recent advances in representation learning to so far under-studied languages. Leveraging non-expert intuitions about lexical semantics can facilitate and speed up the dataset creation process, thus encouraging efforts in language modelling in low-resource languages.

Lexical Ambiguity

In the proposed method polysemy is handled in Phase 1 through label copying and separation of distinct word senses into different classes, which provide the disambiguating context for similarity judgments. However, the number of senses is not specified a priori and their identification lies wholly with non-expert annotators. Future work could explore tackling polysemy explicitly by using a sample of (numbered) verb senses rather than verb forms. In this thesis, I used sense numbering in the sample translation phase of the multilingual data collection (Chapter 5) in the mappings from the source to the target language (e.g., *aggravate*$_1$ – PL *pogorszyć* 'to make worse or more serious' vs. *aggravate*$_2$ – PL *drażnić* 'exasperate'). The guidelines permitted one-to-many and many-to-one translations reflecting different patterns of polysemy, however, the identification of the distinct senses was left to the native-speaker translator. The sense labels were used only for post hoc cross-lingual analyses, rather than data collection itself (i.e., users in each language only dealt with verb forms). The main challenge of introducing sense labels into the two-phase protocol concerns providing the necessary disambiguating contexts or sense descriptions within the Phase 1 and Phase 2 interface.

From the technical point of view, a simple functionality could be introduced where by hovering the cursor over a word label the user could reveal a sense description (e.g., *shine* → 'to make bright by polishing'). More challenging is the selection of the set of senses for each word (Brown, 2008) and the contexts representing them, which would introduce a lot of subjectivity into the annotation design (Hill et al., 2015).

Verb Lexicon Creation

The comparative analyses in Chapters 4 and 5 revealed that there is substantial overlap between the semantic classes and spatial similarity data generated in this work and VerbNet-style classes, suggesting that human judgments could be leveraged to facilitate the creation of a verb lexicon in languages lacking such resources. A direction for future work would be to realise this goal. One possibility, hinted at earlier in this thesis, could involve expert review and refinement of the Phase 1 classes and hierarchical clusters emerging from Phase 2 matrices, potentially followed by manual addition of syntactic descriptions and semantic roles by trained linguists. Combining the semantic and syntactic levels of analysis would require design choices regarding the granularity of the classes and the exact membership criteria, and deciding which type of information should drive the classification depending on the ultimate application. Alternatively, future work could explore the potential of generating syntactic descriptions semi-automatically. Characteristic patterns of syntactic behaviour for each verb could be automatically derived from dependency-parsed corpora, like in the work of Chiu et al. (2019), and provided to annotators for manual selection and revision. Yet another research direction could examine the potential of deriving syntactic clusters automatically based on dependency-based word embeddings (Bansal et al., 2014; Levy and Goldberg, 2014) or word-level representations produced by contextual encoders fine-tuned on the task of syntactic parsing (Kondratyuk and Straka, 2019; Zhou and Zhao, 2019), where the ultimate classes would contain the intersection of automatic syntactic and human-generated semantic clusters.

A challenge for future verb lexicon creation projects lies in scaling up the methodology presented in this thesis to reach the vast coverage offered by such resources as WordNet or VerbNet (e.g., the Unified Verb Index currently includes 9344 verbs[1]). In this work, ∼800-word samples were sorted into broad clusters in Phase 1, with each participant working on the entire sample and spending, on average, over 2.5 hours on the task. Since the time required to complete the task grows as a function of sample size, a sample 10 times larger would be impossible to review in a single session by an

[1]https://uvi.colorado.edu [Accessed on 07/07/2021]

individual annotator. To avoid a dramatic increase in the duration of the task and the cognitive load on the annotators, the current task pipeline could be extended to include an automatic sorting phase, aimed at splitting the whole vocabulary into subsets of <1000 verbs, which can be accommodated in Phase 1, based on their distribution in corpora. As shown in Chapters 4 and 5, word embeddings pretrained on raw text can be readily used to derive clusters of related words based on their relative distances in the embedding space. The automatic clustering step would allow parallelising the work in Phase 1: Each coarse cluster of several hundred words would be individually fed into the Phase 1 interface, with separate groups of annotators simultaneously working on clustering each subset further, analogously to the current setup, to create input samples for Phase 2.

Multi-modal Studies of Typological Variation

The spatial multi-arrangement method prompts the user to consider the relation between any two words from different perspectives, as the judgment context is changing from trial to trial. The ultimate pairwise scores balance out these considerations by averaging over many relative judgments of all possible pairings of words in the set. A considerable benefit of the method is that users can express fine-grained meaning distinctions without consciously defining the semantic features or criteria based on which each pairwise judgment is made. They implicitly assign a two-dimensional embedding vector to each item within a presented set by determining their relative locations, and the dimensions can correspond to different aspects of meaning from trial to trial (e.g., intensity, rate of change, speed). The reconstruction of the semantic space encoded in the resultant average dissimilarity matrices (over many trials) (Chapter 5) showed that the main dimensions of meaning underlying the judgments mirror well-known linguistic phenomena (e.g., polarity, dynamicity, mode of motion). A natural extension of the presented analyses could involve a systematic study of cross-lingual variation across all the semantic domains, to uncover the dimensions and meaning components that are truly universal. Moreover, given that the knowledge probed by the two-phase approach is implicit in nature, in the future, a larger-scale, psycholinguistic study with a fully randomised design and more participants assigned to each task would allow investigating the patterns of intralingual, inter-subject variation in depth.

Further, the spatial arrangement data could be analysed in order to determine if some explicit semantic or syntactic features or category membership explain their distribution across languages. This could be achieved by analysing the correlation between human similarity judgments and a feature-based and categorical model,

similarly to the work of Jozwik et al. (2016) on visual perception. This could involve, for instance, analysing the verb sample with regard to subcategorisation and diathesis alternation information (e.g., 'ditransitive', 'participates in conative alternation') in order to create a feature-word matrix, each entry representing a presence or absence of a syntactic feature. Analogously, a semantic feature-word matrix could include features such as 'stativity', 'repetition', 'causativity', 'negative polarity'. Whereas the semantic categorical model would consist of hierarchically nested category labels (e.g., 'action', 'motion', 'directed motion'). The study could shed light on whether similarity judgments of verbs are strongly categorical, as it is the case with visual stimuli (Jozwik et al., 2016), or whether they are better explained by the feature-based models.

Moreover, an especially rich area which future research could explore is the study of typological variation across different modalities, text and vision. An investigation inspired by the much debated Sapir-Whorf hypothesis (Sapir, 1985; Whorf, 1956) could study whether and to what extent the speakers' native language impacts their perception of similarity and attention to discriminating features in a comparative study of spatial arrangements of visual and textual stimuli representing the same concepts. This could contribute further evidence to the discussion and body of research on linguistic relativity (e.g., Franklin et al., 2008; Simmons et al., 2008; Slobin, 1996) and provide answers to questions such as: Do speakers of cognate languages align more in their perceptions of similarities of objects than those speaking typologically distant languages? Does the modality impact the perceived similarity of concepts, and if so, are judgments of speakers of different languages more aligned when the concepts are represented visually rather than with language-specific labels?

Improved Software for User-Friendly Data Collection

One of the advantages of the spatial arrangement method is its intuitiveness: users express their perception of degree of similarity and difference in word meaning through fluid item placements, rather than having to transpose them onto a discrete numerical scale. However, as discussed in Chapter 5, languages using different scripts may impose different technical demands on the multi-arrangement interface. In order to ensure that the intuitive nature of the task is not compromised by poor legibility (e.g., in the case of logographic scripts), it would be useful to add a zooming feature to the existing interface allowing participants to close in on word labels as needed. To improve the user experience even further, the ultimate goal would be to adapt the software to enable collecting spatial judgments on touchscreen devices.

Similarly, the verb clustering task (Phase 1) interface could benefit from a search functionality, which would allow quickly locating a previously seen verb label. In the studies reported in Chapters 4 and 5, users were presented with a scrollable alphabetical queue of ∼800 verbs, from which they could drag and drop the verb labels into the circles representing clusters. In order to go back to a previously encountered word, they had to scroll the list all the way to the word's location. The alphabetical order was chosen to facilitate that process: In trial experiments, participants reported that managing a randomised sample made the task very arduous, as the whole list needed to be searched every time in order to find a previously seen word. However, to avoid any bias from the order of presentation on the resultant clusters, it would be desirable to present the sample in a randomised order each time. Adding a search function to the interface would make that feasible by allowing directly accessing a word of interest, regardless of its location in a randomised list.

7.3.2 Knowledge Injection and Cross-lingual Transfer

The experiments presented in the last chapter revealed that pretrained encoders still benefit from external human-generated lexical information. What is more, there is potential to harness the sensitivity of native speakers to subtle distinctions in verb meaning for the purpose of supplementing the purely distributional signal encoded by neural models. Combined with a time-efficient data collection protocol, this has powerful implications for their success in many applications and problems relying on accurate verb processing, such as conversational agents, human-machine communication in voice-controlled robotics, or temporal information extraction for financial forecasting and the clinical domain (e.g., Chen et al., 2018; Li et al., 2003; She and Chai, 2017; Tourille, 2018; Zi Huang et al., 2003). Further, in specialised domains the data collection process can be carried out ad hoc to focus on the specific demands of a given application, allowing for computationally efficient boosting of neural architectures with injected specialised verb knowledge modules targeting narrow, highly fine-grained areas of verb meaning. The ultimate goal is equipping machines with the ability to acquire such information automatically, and resources which can serve both as evaluation benchmarks for model development and external sources of knowledge covering specific problem areas can significantly facilitate the pursuit of this objective.

Alternative Adapter Training Regimes

In this thesis, training verb adapters by means of a binary classification task consisting in predicting whether a pair of words shared a VerbNet class or FrameNet frame proved the most successful. It yielded better performing adapters than two sentence-level tasks considered, where target verbs were presented in sentential contexts drawn from example sentences illustrating verb behaviour in both resources, with each member of a given class substituted for the target verb in each sentence to generate the training data. For each positive training instance (i.e., two sentences representing usages of two verbs belonging to the same class), negative instances were generated for the binary classification task by pairing sentences featuring verbs from different classes. The model then had to predict if the target verbs shared a class/frame or (in the multi-class setup) predict the correct VerbNet class or FrameNet frame. Sentence-level setups allow for polysemy-aware training where distinct verb senses are presented in disambiguating sentential contexts, playing to the strengths of the naturally dynamic representations of the contextual encoders. However, it is likely that the relatively poorer performance of the sentence-level adapters was due to the fact that, in training, the model learned to solve the classification task based on superficial contextual cues without developing any awareness of the semantic-syntactic relationship between the two target verbs in a sentence pair.

In order to unlock the potential of sentence-level verb adapter pretraining, future work could explore alternative, more sophisticated methods for training data generation, for instance, complementing sentence input with syntactic parses to enforce learning of subcategorisation patterns. Further, an alternative approach could be explored where the verb knowledge adapter is additionally trained on a synthetic corpus generated from the templates based on VerbNet/FrameNet example sentences via masked language modelling (e.g., *The clown's antics [MASK] the children.*), which could help sensitise the model to the typical class-specific contexts in which a given verb appears. Another so far uncharted territory is lexical knowledge injection eschewing pairwise word-level constraints completely and directly encoding the multi-directional relations within a semantic space captured by the dissimilarity matrices from Phase 2. Injecting knowledge of the relative distances between words within specific semantic fields would entail fine-tuning the local topology of semantic sub-spaces (Glavaš et al., 2019).

Alternative Methods for Cross-lingual Transfer

The cross-lingual transfer approaches evaluated in this work involved either direct model transfer using the massively multilingual BERT and source language VerbNet/FrameNet data, or annotation transfer where the class/frame-sharing verb pairs were automatically translated from English into the target language, filtered, and used as training instances for the language-specific BERT. Future work could explore yet another approach, inspired by the work on semantic specialisation of Vulić et al. (2017b) (see Section 2.3.1). This would involve, first, fine-tuning a static word vector space using a specialisation method which injects positive constraints between words to make their representations more similar (Wieting et al., 2015). These constraints would be (i) the English verb pairs sharing a VerbNet class or FrameNet frame used in this work, and (ii) cross-lingual synonymy (translation) pairs (e.g., from BabelNet (Navigli and Ponzetto, 2010); EN *cry* – IT *piangere*). Next, in a post-specialisation step, the learned specialisation function would be applied to all verbs in the target language vocabulary. Then, using an off-the-shelf clustering algorithm, target language verbs would be clustered based on their post-specialised vector space representations. Finally, pairwise constraints would be derived from these verb clusters (i.e., all unique pairings of cluster-sharing verbs), analogously to the method used with Phase 2 data in Chapter 6; these would then be used for monolingual training in the target language using an available pretrained encoder. This would allow for further examination of cross-lingual portability of verb classes and their potential to boost downstream model performance.

References

Charu C. Aggarwal and Chandan K. Reddy. 2014. *Data clustering: Algorithms and applications.* CRC Press, London.

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Pasca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and WordNet-based approaches. In *Proceedings of NAACL-HLT*, pages 19–27.

Jacqueline Aguilar, Charley Beller, Paul McNamee, Benjamin Van Durme, Stephanie Strassel, Zhiyi Song, and Joe Ellis. 2014. A comparison of the events and relations across ace, ere, tac-kbp, and framenet annotation standards. In *Proceedings of the Second Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 45–53.

David Ahn. 2006. The stages of event extraction. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, pages 1–8.

Syed Sarfaraz Akhtar, Arihant Gupta, Avijit Vajpayee, Arjit Srivastava, and Manish Shrivastava. 2017. Word similarity datasets for indian languages: Annotation and baseline systems. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 91–94.

Suriati Akmal, Li-Hsing Shih, and Rafael Batres. 2014. Ontology-based similarity for product information retrieval. *Computers in Industry*, 65(1):91–107.

Haldun Akoglu. 2018. User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine*, 18(3):91–93.

Rami Al-Rfou, Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. 2015. Polyglot-NER: Massive multilingual named entity recognition. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, pages 586–594.

Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual NLP. In *Proceedings of CoNLL*, pages 183–192.

James Allan. 2002. *Topic Detection and Tracking: Event-Based Information Organization*, volume 12. Springer Science & Business Media, New York.

Faaza A. Almarsoomi, James D. O'Shea, Zuhair Bandar, and Keeley Crockett. 2013. AWSS: An algorithm for measuring Arabic word semantic similarity. In *2013 IEEE International Conference on Systems, Man, and Cybernetics*, pages 504–509.

Gerry T. M. Altmann and Yuki Kamide. 1999. Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, 73(3):247–264.

Enrique Amigó, Julio Gonzalo, Javier Artiles, and Felisa Verdejo. 2009. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*, 12(4):461–486.

Prithviraj Ammanabrolu, Ethan Tien, Wesley Cheung, Zhaochen Luo, William Ma, Lara J. Martin, and Mark O. Riedl. 2020. Story realization: Expanding plot events into sentences. In *Proceedings of AAAI*, volume 34, pages 7375–7382.

Pranav Anand and Valentine Hacquard. 2008. Epistemics with attitude. In *Semantics and Linguistic Theory*, volume 18, pages 37–54.

Juan Aparicio, Mariona Taulé, and Maria Antònia Martí. 2008. AnCora-Verb: A lexical resource for the semantic annotation of corpora. In *Proceedings of LREC*, pages 797–802.

Jun Araki and Teruko Mitamura. 2018. Open-domain event detection using distant supervision. In *Proceedings of COLING*, pages 878–891.

Carlos Santos Armendariz, Matthew Purver, Matej Ulčar, Senja Pollak, Nikola Ljubešić, and Mark Granroth-Wilding. 2020. CoSimLex: A resource for evaluating graded word similarity in context. In *Proceedings of LREC*, pages 5878–5886.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. Translation artifacts in cross-lingual transfer learning. In *Proceedings of EMNLP*, pages 7674–7684.

Shima Asaadi, Saif Mohammad, and Svetlana Kiritchenko. 2019. Big BiRD: A large, fine-grained, bigram relatedness dataset for examining semantic composition. In *Proceedings of NAACL-HLT*, pages 505–516.

F. Gregory Ashby and Nancy A. Perrin. 1988. Toward a unified theory of similarity and recognition. *Psychological Review*, 95(1):124.

Jordi Atserias, Luis Villarejo, German Rigau, Eneko Agirre, John Carroll, Bernardo Magnini, and Piek Vossen. 2004. The meaning multilingual central repository. In *Proceedings of the 2nd International Global Wordnet Conference, January 20-23, 2004*, pages 23–30.

Oded Avraham and Yoav Goldberg. 2016. Improving reliability of word similarity evaluation by redesigning annotation task and performance measure. In *Proceedings of REPEVAL*, pages 106–110.

Asaf Bachrach, Isabelle Roy, and Linnaea Stockall. 2014. *Structuring the Argument: Multidisciplinary Research on Verb Argument Structure*, volume 10. John Benjamins Publishing Company.

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *Proceedings of LREC*, pages 563–566.

Dan Bailey, Yuliya Lierler, and Benjamin Susman. 2015. Prepositional phrase attachment problem revisited: How Verbnet can help. In *Proceedings of IWCS*, pages 12–22.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of COLING*, pages 86–90.

Collin F. Baker and Josef Ruppenhofer. 2002. FrameNet's frames vs. Levin's verb classes. In *Annual Meeting of the Berkeley Linguistics Society*, volume 28, pages 27–38.

Mark Baker and Jonathan Bobaljik. 2017. On inherent and dependent theories of ergative case. *The Oxford Handbook of Ergativity*, 111:134.

Simon Baker, Roi Reichart, and Anna Korhonen. 2014. An unsupervised model for instance level subcategorization acquisition. In *Proceedings of EMNLP*, pages 278–289.

Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2014. Tailoring continuous word representations for dependency parsing. In *Proceedings of ACL*, pages 809–815.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of ACL*, pages 238–247.

Marco Baroni and Alessandro Lenci. 2011. How we BLESSed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics, EMNLP*, pages 1–10.

Marco Baroni, Brian Murphy, Eduard Barbu, and Massimo Poesio. 2010. Strudel: A corpus-based semantic model based on properties and types. *Cognitive Science*, 34(2):222–254.

Siamak Barzegar, Brian Davis, Manel Zarrouk, Siegfried Handschuh, and André Freitas. 2018. SemR-11: A multi-lingual gold-standard for semantic similarity and relatedness for eleven languages. In *Proceedings of LREC*, pages 3912–3916.

Miroslav Batchkarov, Thomas Kober, Jeremy Reffin, Julie Weeds, and David Weir. 2016. A critique of word similarity as a method for evaluating distributional semantic models. In *Proceedings of REPEVAL*, pages 7–12.

Montserrat Batet, David Sánchez, Aida Valls, and Karina Gibert. 2013. Semantic similarity estimation from multiple ontologies. *Applied Intelligence*, 38(1):29–44.

Robert de Beaugrande. 1991. *Linguistic Theory: The Discourse of Fundamental Works*. Routledge, London.

Cosmin Bejan and Sanda Harabagiu. 2010. Unsupervised event coreference resolution with rich linguistic features. In *Proceedings of ACL*, pages 1412–1422.

Adriana Belletti and Luigi Rizzi. 1988. Psych-verbs and θ-theory. *Natural Language & Linguistic Theory*, pages 291–352.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of ACL*, pages 5185–5198.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research*, 3(Feb):1137–1155.

Jiang Bian, Bin Gao, and Tie-Yan Liu. 2014. Knowledge-powered deep learning for word embedding. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 132–148.

Eckhard Bick. 2011. A FrameNet for Danish. In *Proceedings of the 18th Nordic Conference of Computational Linguistics*, pages 34–41.

Chris Biemann. 2005. Ontology learning from text: A survey of methods. In *LDV forum*, volume 20, pages 75–93.

Magdalena Biesialska, Bardia Rafieian, and Marta R. Costa-jussà. 2020. Enhancing word embeddings with knowledge extracted from lexical resources. In *Proceedings of ACL Student Research Workshop*, pages 271–278.

William Black, Sabri Elkateb, Horacio Rodriguez, Musa Alkhalifa, Piek Vossen, Adam Pease, and Christiane Fellbaum. 2006. Introducing the Arabic WordNet project. In *Proceedings of the Third International WordNet Conference*, pages 295–300.

Philip Blair, Yuval Merhav, and Joel Barry. 2017. Automated generation of multilingual clusters for the evaluation of distributed representations. In *Proceedings of ICLR Workshop Papers*, volume abs/1611.01547.

Leonard Bloomfield. 1933. *Language*. Holt, Rinehard and Winston, New York.

BNC Consortium. 2007. British National Corpus. *Oxford Text Archive Core Collection*.

Hans C. Boas. 2002. Bilingual FrameNet dictionaries for machine translation. In *Proceedings of LREC*, pages 1364–1371.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the ACL*, 5:135–146.

Dwight Bolinger. 1968. Postposed main phrases: An English rule for the Romance subjunctive. *Canadian Journal of Linguistics/Revue canadienne de linguistique*, 14(1):3–30.

Francis Bond and Ryan Foster. 2013. Linking and Extending an Open Multilingual Wordnet. In *Proceedings of ACL*, pages 1352–1362.

Jean-Paul Boons, Alain Guillet, and Christian Leclère. 1976. *La structure des phrases simples en français: Constructions intransitives*, volume 1. Droz.

Ingwer Borg and Patrick J.F. Groenen. 2005. *Modern Multidimensional Scaling: Theory and Applications*. Springer Science & Business Media.

Peter Bosch. 1988. On representing lexical meaning. *Meaning and Lexicography*, pages 62–72.

Ulrik Brandes. 2001. A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 25(2):163–177.

Chris Brew and Sabine Schulte im Walde. 2002. Spectral clustering for German verbs. In *Proceedings of EMNLP*, pages 117–124.

Penelope Brown and Melissa Bowerman. 2008. *Crosslinguistic Perspectives on Argument Structure: Implications for Learnability.* Taylor & Francis, New York.

Richard W. Brown. 1957. Linguistic determinism and the part of speech. *Journal of Abnormal Psychology*, 55 1:1–5.

Susan Windisch Brown. 2008. Choosing sense distinctions for WSD: Psycholinguistic evidence. In *Proceedings of ACL-08: HLT, Short Papers*, pages 249–252.

Susan Windisch Brown and Martha Palmer. 2012. Semantic annotation of metaphorical verbs with VerbNet: A case study of 'climb' and 'poison'. In *Workshop on Interoperable Semantic Annotation*, pages 72–76.

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam Khanh Tran. 2012. Distributional semantics in technicolor. In *Proceedings of ACL*, pages 136–145.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47.

Alexander Budanitsky and Graeme Hirst. 2001. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In *Proceedings of the Workshop on WordNet and Other Lexical Resources*, pages 29–34.

Alexander Budanitsky and Graeme Hirst. 2006. Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47.

Ekaterina Buyko, Elena Beisswanger, and Udo Hahn. 2010. The GeneReg Corpus for gene expression regulation events – An overview of the corpus and its in-domain and out-of-domain interoperability. In *Proceedings of LREC*, pages 2662–2666.

Jose Camacho-Collados, Mohammad Taher Pilehvar, Nigel Collier, and Roberto Navigli. 2017. Semeval-2017 task 2: Multilingual and cross-lingual semantic word similarity. In *Proceedings of SemEval*, pages 15–26.

José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. 2015a. A framework for the construction of monolingual and cross-lingual word similarity datasets. In *Proceedings of ACL–IJCNLP*, pages 1–7.

José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. 2015b. A unified multilingual semantic representation of concepts. In *Proceedings of ACL–IJCNLP*, pages 741–751.

Marie Candito, Pascal Amsili, Lucie Barque, Farah Benamara, Gaël de Chalendar, Marianne Djemaa, Pauline Haas, Richard Huyghe, Yvette Yannick Mathieu, Philippe Muller, Benoît Sagot, and Laure Vieu. 2014. Developing a French FrameNet: Methodology and first results. In *Proceedings of LREC*, pages 1372–1379.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. In *Proceedings of ICLR*.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2002. RST Discourse Treebank LDC2002T07. Technical report, Philadelphia: Linguistic Data Consortium. Web Download.

Daniel Casasanto. 2008. Similarity and proximity: When does close in space mean close in mind? *Memory & Cognition*, 36(6):1047–1056.

Y. H. Chan. 2003. Biostatistics 104: Correlational analysis. *Singapore Medical Journal*, 44(12):614–9.

Ian Charest, Rogier A. Kievit, Taylor W. Schmitz, Diana Deca, and Nikolaus Kriegeskorte. 2014. Unique semantic space in the brain of each beholder predicts perceived similarity. *Proceedings of the National Academy of Sciences*, 111(40):14565–14570.

C. Chen, H. Huang, Y. Shiue, and H. Chen. 2018. Numeral understanding in financial tweets for fine-grained crowd-based forecasting. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 136–143.

Danqi Chen and Christopher Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of EMNLP*, pages 740–750.

Fuzan Chen, Chenghua Lu, Harris Wu, and Minqiang Li. 2017a. A semantic similarity measure integrating multiple conceptual relationships for web service discovery. *Expert Systems with Applications*, 67:19–31.

Yubo Chen, Shulin Liu, Xiang Zhang, Kang Liu, and Jun Zhao. 2017b. Automatically labeled data generation for large scale event extraction. In *Proceedings of ACL*, pages 409–419.

Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of ACL–IJCNLP*, pages 167–176.

Christine Chiarello, Curt Burgess, Lorie Richards, and Alma Pollock. 1990. Semantic and associative priming in the cerebral hemispheres: Some words do, some words don't. . . sometimes, some places. *Brain and Language*, 38(1):75–104.

Billy Chiu, Olga Majewska, Sampo Pyysalo, Laura Wey, Ulla Stenius, Anna Korhonen, and Martha Palmer. 2019. A neural classification method for supporting the creation of BioVerbNet. *Journal of Biomedical Semantics*, 10(1):2.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. MIT Press.

Noam Chomsky. 2014. *The Minimalist Program*. MIT press.

Radoslaw M. Cichy, Nikolaus Kriegeskorte, Kamila M. Jozwik, Jasper J. F. van den Bosch, and Ian Charest. 2019. The spatiotemporal neural dynamics underlying perceived similarity for real-world objects. *NeuroImage*, 194:12–24.

Philipp Cimiano, Andreas Hotho, and Steffen Staab. 2005. Learning concept hierarchies from text corpora using formal concept analysis. *Journal of Artificial Intelligence research*, 24:305–339.

Silvie Cinková, Eva Fučíková, Jana Šindlerová, and Jan Hajič. 2014. EngVallex – English Valency Lexicon. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Eve V. Clark. 2009. *First Language Acquisition*. Cambridge University Press.

Peter Clark, Bhavana Dalvi, and Niket Tandon. 2018. What happened? Leveraging VerbNet to predict the effects of actions in procedural text. *arXiv preprint arXiv:1804.05435*.

Ronan Collobert and Jason Weston. 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.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020a. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*, pages 8440–8451.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Proceedings of NeurIPS*, pages 7057–7067.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Emerging cross-lingual structure in pretrained language models. In *Proceedings of ACL*, pages 6022–6034.

Dick Crouch and Tracy Holloway King. 2005. Unifying lexical resources. In *Proceedings of the Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*, pages 32–37.

David A. Cruse. 1986. *Lexical Semantics*. Cambridge University Press.

Christoph Dalitz and Katrin E. Bednarek. 2016. Sentiment lexica from paired comparisons. In *Proceedings of ICDM*, pages 924–930.

Barbara Dancygier. 2017. *The Cambridge Handbook of Cognitive Linguistics*. Cambridge University Press.

Hoa Trang Dang. 2004. *Investigations into the Role of Lexical Semantics in Word Sense Disambiguation*. Ph.D. thesis, University of Pennsylvania.

Hoa Trang Dang, Karin Kipper, Martha Palmer, and Joseph Rosenzweig. 1998. Investigating regular sense extensions based on intersective Levin classes. In *Proceedings of ACL–COLING*, pages 293–299.

Donald Davidson. 1967. The logical form of action sentences. In Nicholas Rescher, editor, *The Logic of Decision and Action*, pages 81–95. University of Pittsburgh Press.

William H.E. Day and Herbert Edelsbrunner. 1984. Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of Classification*, 1(1):7–24.

Daniel P. De Oliveira, James H. Garrett Jr, and Lucio Soibelman. 2011. A density-based spatial clustering approach for defining local indicators of drinking water distribution pipe breakage. *Advanced Engineering Informatics*, 25(2):380–389.

Daniel Defays. 1977. An efficient algorithm for a complete link method. *The Computer Journal*, 20(4):364–366.

Leon R.A. Derczynski. 2017. *Automatically Ordering Events and Times in Text*. Springer, Berlin.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL–HLT*, pages 4171–4186.

Paramveer S. Dhillon, Dean P. Foster, and Lyle H. Ungar. 2015. Eigenwords: Spectral word embeddings. *Journal of Machine Learning Research*, 16:3035–3078.

Robert M.W. Dixon. 1991. *A Semantic Approach to English Grammar*. Oxford University Press.

George R. Doddington, Alexis Mitchell, Mark A. Przybocki, Lance A. Ramshaw, Stephanie M. Strassel, and Ralph M. Weischedel. 2004. The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of LREC*, pages 837–840.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Shaoyang Duan, Ruifang He, and Wenli Zhao. 2017. Exploiting document level information to improve event detection via recurrent neural networks. In *Proceedings of IJCNLP*, pages 352–361.

Jean Dubois, Françoise Dubois-Charlier, and Françoise Dubois-Charlier. 1997. *Les verbes français*. Larousse.

Joshua Eisenberg and Mark Finlayson. 2017. A simpler and more generalizable story detector using verb and character features. In *Proceedings of EMNLP*, pages 2708–2715.

Gökhan Ercan and Olcay Taner Yıldız. 2018. AnlamVer: Semantic model evaluation dataset for Turkish-word similarity and relatedness. In *Proceedings of COLING*, pages 3819–3836.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, volume 96, pages 226–231.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the ACL*, 8:34–48.

Ingrid Falk, Claire Gardent, and Jean-Charles Lamirel. 2012. Classifying French verbs using French and English lexical resources. In *Proceedings of ACL*, pages 854–863.

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of NAACL*, pages 1606–1615.

Manaal Faruqui and Chris Dyer. 2015. Non-distributional word vector representations. In *Proceedings of NAACL-HLT*, pages 464–469.

Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer. 2016. Problems with evaluation of word embeddings using word similarity tasks. In *Proceedings of REPEVAL*, pages 30–35.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

James Ferguson, Colin Lockard, Daniel Weld, and Hannaneh Hajishirzi. 2018. Semi-supervised event extraction with paraphrase clusters. In *Proceedings of NAACL*, pages 359–364.

Todd R. Ferretti, Ken McRae, and Andrea Hatherell. 2001. Integrating verbs, situation schemas, and thematic role concepts. *Journal of Memory and Language*, 44(4):516–547.

Charles J. Fillmore. 1965. *Indirect Object Constructions in English and the Ordering of Transformations*. Mouton.

Charles J. Fillmore. 1967. *The Grammar of Hitting and Breaking*. Ohio State University. Department of Linguistics.

Charles J. Fillmore. 1968. Lexical entries for verbs. *Foundations of Language*, pages 373–393.

Charles J. Fillmore. 1976. Frame semantics and the nature of language. In *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, volume 280, pages 20–32.

Charles J. Fillmore. 1977. The need for a frame semantics in linguistics. In *Statistical Methods in Linguistics*. Ed. Hans Karlgren. Scriptor.

Charles J. Fillmore. 1982. Frame semantics. In *Linguistics in the Morning Calm*, pages 111–137. Hanshin Publishing Co.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131.

Cynthia Fisher. 1996. Structural limits on verb mapping: The role of analogy in children's interpretations of sentences. *Cognitive Psychology*, 31(1):41–81.

Cynthia Fisher, Henry Gleitman, and Lila R Gleitman. 1991. On the semantic content of subcategorization frames. *Cognitive psychology*, 23(3):331–392.

Raffaella Folli and Gillian Ramchand. 2005. Prepositions and results in Italian and English: An analysis from event decomposition. In *Perspectives on Aspect*, pages 81–105. Springer.

Anna Franklin, Gilda V. Drivonikou, Laura Bevis, Ian R.L. Davies, Paul Kay, and Terry Regier. 2008. Categorical perception of color is lateralized to the right hemisphere in infants, but to the left hemisphere in adults. *Proceedings of the National Academy of Sciences*, 105(9):3221–3225.

André Freitas, Siamak Barzegar, Juliano Efson Sales, Siegfried Handschuh, and Brian Davis. 2016. Semantic relatedness for all (languages): A comparative analysis of multilingual semantic relatedness using machine translation. In *European Knowledge Acquisition Workshop*, pages 212–222.

André Freitas, Joao Gabriel Oliveira, Seán O'Riain, Edward Curry, and João Carlos Pereira Da Silva. 2011. Querying linked data using semantic relatedness: a vocabulary independent approach. In *International Conference on Application of Natural Language to Information Systems*, pages 40–51.

Cheryl Frenck-Mestre and Steve Bueno. 1999. Semantic features and semantic categories: Differences in rapid activation of the lexicon. *Brain and Language*, 68(1-2):199–204.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The paraphrase database. In *Proceedings of NAACL–HLT*, pages 758–764.

Peter Gärdenfors. 2004. *Conceptual Spaces: The Geometry of Thought*. MIT press.

Dirk Geeraerts. 2010. *Theories of Lexical Semantics*. Oxford University Press.

Dedre Gentner. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2):155–170.

Dedre Gentner. 2006. Why verbs are hard to learn. *Action Meets Word: How Children Learn Verbs*, pages 544–564.

Dedre Gentner and Arthur B. Markman. 1997. Structure mapping in analogy and similarity. *American Psychologist*, 52(1):45.

Daniela Gerz, Ivan Vulić, Felix Hill, Roi Reichart, and Anna Korhonen. 2016. SimVerb-3500: A large-scale evaluation set of verb similarity. In *Proceedings of EMNLP*, pages 2173–2182.

Gyslain Giguère. 2006. Collecting and analyzing data in multidimensional scaling experiments: A guide for psychologists using SPSS. *Tutorials in Quantitative Methods for Psychology*, 2(1):27–38.

Ana-Maria Giuglea and Alessandro Moschitti. 2006. Semantic role labeling via FrameNet, VerbNet and PropBank. In *Proceedings of ACL–COLING*, pages 929–936.

Anna Gladkova and Aleksandr Drozd. 2016. Intrinsic evaluations of word embeddings: What can we do better? In *Proceedings of REPEVAL*, pages 36–42.

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: What works and what doesn't. In *Proceedings of the NAACL Student Research Workshop*, pages 8–15.

Goran Glavaš, Edoardo Maria Ponti, and Ivan Vulić. 2019. Semantic specialization of distributional word vectors. In *Proceedings of EMNLP: Tutorial Abstracts*.

Goran Glavaš and Jan Šnajder. 2015. Construction and evaluation of event graphs. *Natural Language Engineering*, 21(4):607–652.

Goran Glavaš and Ivan Vulić. 2018a. Discriminating between lexico-semantic relations with the specialization tensor model. In *Proceedings of NAACL–HLT*, pages 181–187.

Goran Glavaš and Ivan Vulić. 2018b. Explicit retrofitting of distributional word vectors. In *Proceedings of ACL*, pages 34–45.

Goran Glavaš and Ivan Vulić. 2019. Generalized tuning of distributional word vectors for monolingual and cross-lingual lexical entailment. In *Proceedings of ACL*, pages 4824–4830.

Lila Gleitman. 1990. The structural sources of verb meanings. *Language Acquisition*, 1(1):3–55.

Lila R. Gleitman, Kimberly Cassidy, Rebecca Nappa, Anna Papafragou, and John C. Trueswell. 2005. Hard words. *Language Learning and Development*, 1(1):23–64.

Lila R. Gleitman and Eric Wanner. 1982. *Language Acquisition: The State of the Art*. CUP Archive.

Adele E. Goldberg. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. University of Chicago Press.

Robert Goldstone. 1994. An efficient method for obtaining similarity data. *Behavior Research Methods, Instruments, & Computers*, 26(4):381–386.

Robert Goldstone and Ji Yun Son. 2012. *Similarity*. Oxford University Press.

Ian J. Goodfellow, Mehdi Mirza, Aaron Courville Da Xiao, and Yoshua Bengio. 2014. An empirical investigation of catastrophic forgeting in gradient-based neural networks. In *Proceedings of ICLR*.

John C. Gower. 1966. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53(3-4):325–338.

Roger Granada, Cassia Trojahn, and Renata Vieira. 2014. Comparing semantic relatedness between word pairs in Portuguese using Wikipedia. In *International Conference on Computational Processing of the Portuguese Language*, pages 170–175.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of LREC*, pages 3483–3487.

Georgia M. Green. 1974. *Semantics and Syntactic Regularity*. Indiana University Press.

Joseph H. Greenberg. 1957. *Essays in Linguistics*. University of Chicago Press.

Jane Grimshaw. 1979. Complement selection and the lexicon. *Linguistic Inquiry*, 10(2):279–326.

Ralph Grishman, Catherine Macleod, and Adam Meyers. 1994. COMLEX syntax: Building a computational lexicon. In *Proceedings of COLING*, pages 268–272.

Ralph Grishman and Beth Sundheim. 1996. Message Understanding Conference - 6: A brief history. In *Proceedings of COLING*, pages 466–471.

Jess Gropen, Steven Pinker, Michelle Hollander, Richard Goldberg, and Ronald Wilson. 1989. The learnability and acquisition of the dative alternation in English. *Language*, pages 203–257.

Maurice Gross. 1975. *Méthodes en syntaxe*. Hermann.

Jeffrey Steven Gruber. 1965. *Studies in Lexical Relations*. Ph.D. thesis, Massachusetts Institute of Technology.

Mohamed Guerssel, Kenneth Hale, Mary Laughren, Beth Levin, and Josie White Eagle. 1985. A cross-linguistic study of transitivity alternations. *CLS*, 21(2):48–63.

Weiwei Guo, Hao Li, Heng Ji, and Mona Diab. 2013. Linking tweets to news: A framework to enrich short text data in social media. In *Proceedings of ACL*, pages 239–249.

Prashant Gupta and Heng Ji. 2009. Predicting unknown time arguments based on cross-event propagation. In *Proceedings of ACL–IJCNLP*, pages 369–372.

Iryna Gurevych. 2005. Using the structure of a conceptual network in computing semantic relatedness. In *Proceedings of ICNLP*, pages 767–778.

Iryna Gurevych. 2006. Thinking beyond the nouns – computing semantic relatedness across parts of speech. *Sprachdokumentation & Sprachbeschreibung*, 28:226.

Iryna Gurevych, Christof Müller, and Torsten Zesch. 2007. What to be? – Electronic career guidance based on semantic relatedness. In *Proceedings of ACL*, pages 1032–1039.

Ulrike Hahn, Nick Chater, and Lucy B. Richardson. 2003. Similarity as transformation. *Cognition*, 87(1):1–32.

Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Ondřej Bojar, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. 2012. Announcing Prague Czech-English Dependency Treebank 2.0. In *Proceedings of LREC*, pages 3153–3160.

Ken Hale and Jay Keyser. 1986. Some transitivity alternations in English. *Anuario del Seminario de Filología Vasca "Julio de Urquijo"*, 20(3):605–638.

Ken Hale and Samuel Jay Keyser. 1987. A view from the middle. *Lexicon Project Working Papers 10*.

Birgit Hamp and Helmut Feldweg. 1997. GermaNet – A lexical-semantic net for German. In *Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*.

Sébastien Harispe, Sylvie Ranwez, Stefan Janaqi, and Jacky Montmain. 2015. Semantic similarity from natural language and ontology analysis. *Synthesis Lectures on Human Language Technologies*, 8(1):1–254.

Kaitlyn Harrigan, Valentine Hacquard, and Jeffrey Lidz. 2016. Syntactic bootstrapping in the acquisition of attitude verbs: think, want and hope. In *Proceedings of WCCFL*, volume 33, pages 196–206.

Zellig S. Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Iren Hartmann, Martin Haspelmath, and Bradley Taylor, editors. 2013. *Valency Patterns Leipzig*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Silvana Hartmann, Judith Eckle-Kohler, and Iryna Gurevych. 2016. Generating training data for semantic role labeling based on label transfer from linked lexical resources. *Transactions of the ACL*, 4:197–213.

Joshua K. Hartshorne, Claire Bonial, and Martha Palmer. 2014. The VerbCorner Project: Findings from Phase 1 of crowd-sourcing a semantic decomposition of verbs. In *Proceedings of ACL*, pages 397–402.

Joshua K. Hartshorne, Timothy J. O'Donnell, Yasutada Sudo, Miki Uruwashi, Miseon Lee, and Jesse Snedeker. 2016. Psych verbs, the linking problem, and the acquisition of language. *Cognition*, 157:268–288.

Kazuma Hashimoto, Caiming Xiong, Yoshimasa Tsuruoka, and Richard Socher. 2017. A joint many-task model: Growing a neural network for multiple NLP tasks. In *Proceedings of EMNLP*, pages 1923–1933.

Samer Hassan and Rada Mihalcea. 2009. Cross-lingual semantic relatedness using encyclopedic knowledge. In *Proceedings of EMNLP*, pages 1192–1201.

Annette Hautli and Miriam Butt. 2011. Towards a computational semantic analyzer for Urdu. In *Proceedings of the 9th Workshop on Asian Language Resources*, pages 71–78.

Bin He, Di Zhou, Jinghui Xiao, Qun Liu, Nicholas Jing Yuan, and Tong Xu. 2019. Integrating graph contextualized knowledge into pre-trained language models. In *Findings of the Association for Computational Linguistics (EMNLP)*, page 2281–2290.

Xiaodong He, Mei Yang, Jianfeng Gao, Patrick Nguyen, and Robert C Moore. 2008. Indirect-HMM-based hypothesis alignment for combining outputs from machine translation systems. In *Proceedings of EMNLP*, pages 98–107.

Harry Helson. 1964. *Adaptation-Level Theory: An Experimental and Systematic Approach to Behavior*. New York.

Harry Helson, Walter C. Michels, and Artie Sturgeon. 1954. The use of comparative rating scales for the evaluation of psychophysical data. *The American Journal of Psychology*, 67(2):321–326.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (GELUs). *CoRR*, abs/1606.08415.

Karin Friberg Heppin and Maria Toporowska Gronostaj. 2012. The rocky road towards a Swedish FrameNet – creating SweFN. In *Proceedings of LREC*, pages 256–261.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of NAACL–HLT*, pages 4129–4138.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Geoff Hollis and Chris Westbury. 2016. The principals of meaning: Extracting semantic dimensions from co-occurrence models of semantics. *Psychonomic Bulletin & Review*, 23(6):1744–1756.

Keith J. Holyoak and Kyunghee Koh. 1987. Surface and structural similarity in analogical transfer. *Memory & Cognition*, 15(4):332–340.

Yu Hong, Jianfeng Zhang, Bin Ma, Jianmin Yao, Guodong Zhou, and Qiaoming Zhu. 2011. Using cross-entity inference to improve event extraction. In *Proceedings of ACL–HLT*, pages 1127–1136.

Yu Hong, Wenxuan Zhou, Jingli Zhang, Guodong Zhou, and Qiaoming Zhu. 2018. Self-regulation: Employing a generative adversarial network to improve event detection. In *Proceedings of ACL*, pages 515–526.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of ICML*, pages 2790–2799.

Michael C. Hout, Stephen D. Goldinger, and Ryan W. Ferguson. 2013. The versatility of SpAM: A fast, efficient, spatial method of data collection for multidimensional scaling. *Journal of Experimental Psychology: General*, 142(1):256.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of ACL*, pages 328–339.

Andrew Hsi, Yiming Yang, Jaime Carbonell, and Ruochen Xu. 2016. Leveraging multilingual training for limited resource event extraction. In *Proceedings of COLING*, pages 1201–1210.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. In *Proceedings of ICML*.

Chu-Ren Huang, Shu-Kai Hsieh, Jia-Fei Hong, Yun-Zhu Chen, I-Li Su, Yong-Xiang Chen, and Sheng-Wei Huang. 2010. Chinese Wordnet: Design, implementation, and application of an infrastructure for cross-lingual knowledge processing. *Journal of Chinese Information Processing*, 24(2):14–23.

Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of ACL*, pages 873–882.

Junjie Huang, Fanchao Qi, Chenghao Yang, Zhiyuan Liu, and Maosong Sun. 2019. COS960: A Chinese word similarity dataset of 960 word pairs. *CoRR*, abs/1906.00247.

Sabine Schulte Im Walde. 2000. Clustering verbs semantically according to their alternation behaviour. In *Proceedings of COLING*, pages 747–753.

Sabine Schulte Im Walde. 2006. Experiments on the automatic induction of German semantic verb classes. *Computational Linguistics*, 32(2):159–194.

Lidija Iordanskaja, Richard Kittredge, and Alain Polguere. 1991. Lexical selection and paraphrase in a meaning-text generation model. In *Natural Language Generation in Artificial Intelligence and Computational Linguistics*, pages 293–312.

Hitoshi Isahara, Francis Bond, Kiyotaka Uchimoto, Masao Utiyama, and Kyoko Kanzaki. 2008. Development of the Japanese WordNet. In *Proceedings of LREC*, pages 2420–2423.

Ray Jackendoff. 1972. *Semantic Interpretation in Generative Grammar*. MIT Press.

Ray Jackendoff. 1983. *Semantics and Cognition*, volume 8. MIT press.

Ray Jackendoff. 1990. *Semantic Structures*. MIT Press, Cambridge, MA.

Peter Jansen. 2004. Lexicography in an Interlingual Ontology: An Introduction to EuroWordNet. *Canadian Undergraduate Journal of Cognitive Science*, 3:1–5.

Mario Jarmasz and Stan Szpakowicz. 2003. Roget's thesaurus and semantic similarity. In *Proceedings of RANLP*, pages 111–120.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of ACL*, pages 3651–3657.

Otto Jespersen. 2013. *A Modern English Grammar on Historical Principles*, volume 5. Routledge.

Hwiyeol Jo and Stanley Jungkyu Choi. 2018. Extrofitting: Enriching word representation and its vector space with semantic lexicons. In *Proceedings of the 3rd Workshop on Representation Learning for NLP*, pages 24–29.

Eric Joanis, Suzanne Stevenson, and David James. 2008. A general feature space for automatic verb classification. *Natural Language Engineering*, 14(3):337–367.

Douglas A. Jones, Robert C. Berwick, Franklin Cho, Zeeshan Khan, Karen T Kohl, Naoyuki Nomura, Anand Radhakrishnan, Ulrich Sauerland, and Brian Ulicny. 1996. Verb classes and alternations in Bangla, German, English, and Korean. Technical report, Massachusetts Institute of Technology.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of EMNLP*, pages 2979–2984.

Kamila M. Jozwik, Nikolaus Kriegeskorte, and Marieke Mur. 2016. Visual features as stepping stones toward semantics: Explaining object similarity in IT and perception with non-negative least squares. *Neuropsychologia*, 83:201–226.

Dan Jurafsky. 2000. *Speech & Language Processing*. Pearson Education India.

David Jurgens and Ioannis Klapaftis. 2013. SemEval-2013 Task 13: Word sense induction for graded and non-graded senses. In *Proceedings of SEMEVAL*, pages 290–299.

Aishwarya Kamath, Jonas Pfeiffer, Edoardo Maria Ponti, Goran Glavaš, and Ivan Vulić. 2019. Specializing distributional vectors of all words for lexical entailment. In *Proceedings of the 4th Workshop on Representation Learning for NLP*, pages 72–83.

Jonathan Ben Kamp. 2019. Statistical modeling at the syntax-semantics interface: Exploiting automatically induced lexical classes evaluated through variational bayesian inference. Master's thesis, Utrecht University.

Ron Kaplan and Joan Bresnan. 1982. Lexical functional grammar: A formal system for grammatical representation. In Joan Bresnan, editor, *The Mental Representation of Grammatical Relations*, pages 173–281. MIT Press.

Daisuke Kawahara and Martha Palmer. 2014. Single classifier approach for verb sense disambiguation based on generalized features. In *LREC*, pages 4210–4213.

Daisuke Kawahara, Daniel Peterson, and Martha Palmer. 2014. A step-wise usage-based method for inducing polysemy-aware verb classes. In *Proceedings of ACL*, pages 1030–1040.

Douwe Kiela, Felix Hill, and Stephen Clark. 2015. Specializing word embeddings for similarity or relatedness. In *Proceedings of EMNLP*, pages 2044–2048.

Jeong-uk Kim, Younggyun Hahm, and Key-Sun Choi. 2016a. Korean FrameNet expansion based on projection of Japanese FrameNet. In *Proceedings of COLING*, pages 175–179.

Jin-Dong Kim, Tomoko Ohta, Kanae Oda, and Jun'ichi Tsujii. 2008. From text to pathway: Corpus annotation for knowledge acquisition from biomedical literature. In *Proceedings of the 6th Asia-Pacific Bioinformatics Conference*, pages 165–175.

Joo-Kyung Kim, Gokhan Tur, Asli Celikyilmaz, Bin Cao, and Ye-Yi Wang. 2016b. Intent detection using semantically enriched word embeddings. In *IEEE Spoken Language Technology Workshop (SLT)*, pages 414–419.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of ICLR*.

Barbara Ann Kipfer. 2009. *Roget's 21st Century Thesaurus (3rd Edition)*. Philip Lief Group.

Karin Kipper, Hoa Trang Dang, and Martha Stone Palmer. 2000. Class-based construction of a verb lexicon. In *Proceedings of AAAI*, pages 691–696.

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2006. Extending VerbNet with novel verb classes. In *Proceedings of LREC*, pages 1027–1032.

Karin Kipper Schuler. 2005. *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. Ph.D. thesis, University of Pennsylvania.

Svetlana Kiritchenko and Saif Mohammad. 2017. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of ACL*, pages 465–470.

Svetlana Kiritchenko and Saif M. Mohammad. 2016. Capturing reliable fine-grained sentiment associations by crowdsourcing and best–worst scaling. In *Proceedings of NAACL–HLT*, pages 811–817.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, and Agnieszka Grabska-Barwinska. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing Universal Dependencies universally. In *Proceedings of EMNLP–IJCNLP*, pages 2779–2795.

Miloslav Konopík, Ondřej Pražák, and David Steinberger. 2017. Czech dataset for semantic similarity and relatedness. In *Proceedings of RANLP*, pages 401–406.

Anna Korhonen, Yuval Krymolowski, and Ted Briscoe. 2006. A large subcategorization lexicon for natural language processing applications. In *Proceedings of LREC*, pages 1015–1020.

Anna Koufakou and Jason Scott. 2020. Lexicon-enhancement of embedding-based approaches towards the detection of abusive language. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 150–157.

Hideki Kozima. 1994. *Computing Lexical Cohesion as a Tool for Text Analysis*. Ph.D. thesis, The University of Electro-Communications, Chōfu, Tokyo.

Nikolaus Kriegeskorte and Marieke Mur. 2012. Inverse MDS: Inferring dissimilarity structure from multiple item arrangements. *Frontiers in Psychology*, 3:245.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A. Bandettini. 2008. Representational similarity analysis – Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2(4).

Joseph B. Kruskal and Myron Wish. 1978. *Multidimensional Scaling*. 11. Sage.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. 2020. The Open Images Dataset V4. *International Journal of Computer Vision*, 128(7):1956–1981.

John Lafferty, Andrew McCallum, and Fernando C.N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML*, pages 282–289.

George Lakoff and Mark Johnson. 1980. The metaphorical structure of the human conceptual system. *Cognitive Science*, 4(2):195–208.

George Lakoff and Mark Johnson. 1999. *Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought*, volume 4. University of Chicago Press.

Thomas K. Landauer and Susan T. Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211.

Ronald W. Langacker. 2008. *Cognitive Grammar: A Basic Introduction*. OUP USA.

Maria Lapata. 1999. Acquiring lexical generalizations from corpora: A case study for diathesis alternations. In *Proceedings of ACL*, pages 397–404.

Mary Laughren. 1988. Toward a lexical representation of Warlpiri verbs. In *Thematic Relations*, pages 215–242. Brill.

Anne Lauscher, Olga Majewska, Leonardo F. R. Ribeiro, Iryna Gurevych, Nikolai Rozanov, and Goran Glavaš. 2020a. Common sense or world knowledge? Investigating adapter-based knowledge injection into pretrained transformers. In *Proceedings of Deep Learning Inside Out (DeeLIO): The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 43–49.

Anne Lauscher, Ivan Vulić, Edoardo Maria Ponti, Anna Korhonen, and Goran Glavaš. 2020b. Specializing unsupervised pretraining models for word-level semantic similarity. In *Proceedings of COLING*, pages 1371–1383.

Anne Lederer, Henry Gleitman, and Lila Gleitman. 1995. Verbs of a feather flock together: Semantic information in the structure of maternal speech. *Beyond Names for Things: Young Children's Acquisition of Verbs*, 277.

Benoit Lemaire and Guy Denhiere. 2006. Effects of high-order co-occurrences on word semantic similarity. *Current Psychology Letters. Behaviour, Brain & Cognition*, 1(18).

Alessandro Lenci. 1998. The structure of predication. *Synthese*, 114(2):233–276.

Ben Lengerich, Andrew Maas, and Christopher Potts. 2018. Retrofitting distributional embeddings to knowledge graphs with functional relations. In *Proceedings of COLING*, pages 2423–2436.

Ira Leviant and Roi Reichart. 2015a. Judgment language matters: Multilingual vector space models for judgment language aware lexical semantics. *CoRR, abs/1508.00106*.

Ira Leviant and Roi Reichart. 2015b. Separated by an un-common language: Towards judgment language informed vector space modeling. *CoRR, abs/1508.00106*.

Beth Levin. 1985. Lexical semantics in review. *Lexicon Project Working Papers 1*.

Beth Levin. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press.

Beth Levin. 2015. Semantics and pragmatics of argument alternations. *Annual Review of Linguistics*, 1:63–83.

Gary M. Levine, Jamin B. Halberstadt, and Robert L. Goldstone. 1996. Reasoning and the weighting of attributes in attitude judgments. *Journal of Personality and Social Psychology*, 70(2):230.

Yoav Levine, Barak Lenz, Or Dagan, Ori Ram, Dan Padnos, Or Sharir, Shai Shalev-Shwartz, Amnon Shashua, and Yoav Shoham. 2020. SenseBERT: Driving some sense into BERT. In *Proceedings of ACL*, pages 4656–4667.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of ACL*, pages 302–308.

Jianguo Li and Chris Brew. 2008. Which are the best features for automatic verb classification. In *Proceedings of ACL–HLT*, pages 434–442.

Min Li, Jian-er Chen, Jian-xin Wang, Bin Hu, and Gang Chen. 2008. Modifying the DPClus algorithm for identifying protein complexes based on new topological structures. *BMC Bioinformatics*, 9(1):398.

Min Li, Dongyan Li, Yu Tang, Fangxiang Wu, and Jianxin Wang. 2017. CytoCluster: A cytoscape plugin for cluster analysis and visualization of biological networks. *International Journal of Molecular Sciences*, 18(9):1880.

Mu Li, Muhua Zhu, Yang Zhang, and Ming Zhou. 2006. Exploring distributional similarity based models for query spelling correction. In *Proceedings of ACL–COLING*, pages 1025–1032.

Qi Li, Heng Ji, and Liang Huang. 2013. Joint event extraction via structured prediction with global features. In *Proceedings of ACL*, pages 73–82.

Wenjie Li, Kam-Fai Wong, and Chunfa Yuan. 2003. A design of temporal event extraction from Chinese financial news. *International Journal of Computer Processing of Oriental Languages*, 16(01):21–39.

Constantine Lignos, Vasumathi Raman, Cameron Finucane, Mitchell Marcus, and Hadas Kress-Gazit. 2015. Provably correct reactive control from natural language. *Autonomous Robots*, 38(1):89–105.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of ACL–COLING*, pages 768–774.

Liping You and Kaiying Liu. 2005. Building Chinese FrameNet database. In *Proceedings of International Conference on Natural Language Processing and Knowledge Engineering*, pages 301–306.

Thomas Lippincott, Laura Rimell, Karin Verspoor, and Anna Korhonen. 2013. Approaches to verb subcategorization for biomedicine. *Journal of Biomedical Informatics*, 46(2):212–227.

Hugo Liu and Push Singh. 2004. ConceptNet – A practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.

Jian Liu, Yubo Chen, Kang Liu, Wei Bi, and Xiaojiang Liu. 2020a. Event extraction as machine reading comprehension. In *Proceedings of EMNLP*, pages 1641–1651.

Jian Liu, Yubo Chen, Kang Liu, Jun Zhao, et al. 2018. Event detection via gated multilingual attention mechanism. In *Proceedings of AAAI*, pages 4865–4872.

Mei-chun Liu, Ting-yi Chiang, et al. 2008. The construction of Mandarin VerbNet: A frame-based study of statement verbs. *Language and Linguistics*, 9(2):239–270.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019a. Linguistic knowledge and transferability of contextual representations. In *Proceedings of NAACL–HLT*, pages 1073–1094.

Qianchu Liu, Diana McCarthy, Ivan Vulić, and Anna Korhonen. 2019b. Investigating cross-lingual alignment methods for contextualized embeddings with token-level evaluation. In *Proceedings of CoNLL*, pages 33–43.

Quan Liu, Hui Jiang, Si Wei, Zhen-Hua Ling, and Yu Hu. 2015. Learning semantic word embeddings based on ordinal knowledge constraints. In *Proceedings of ACL–IJCNLP*, pages 1501–1511.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020b. K-BERT: Enabling language representation with knowledge graph. In *Proceedings of AAAI*, volume 34, pages 2901–2908.

Xiao Liu, Heyan Huang, and Yue Zhang. 2019c. Open domain event extraction using neural latent variable models. In *Proceedings of ACL*, pages 2860–2871.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019d. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Hector Llorens, Estela Saquete, and Borja Navarro. 2010. TIPSem (English and Spanish): Evaluating CRFs and semantic roles in TempEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 284–291.

Stuart Lloyd. 1982. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137.

Iñigo Lopez-Gazpio, Montse Maritxalar, Aitor Gonzalez-Agirre, German Rigau, Larraitz Uria, and Eneko Agirre. 2017. Interpretable semantic textual similarity: Finding and explaining differences between sentences. *Knowledge-Based Systems*, 119:186–199.

Jordan J. Louviere, Terry N. Flynn, and Anthony Alfred John Marley. 2015. *Best-Worst Scaling: Theory, Methods and Applications*. Cambridge University Press.

Jordan J. Louviere and George G. Woodworth. 1991. Best-worst scaling: A model for the largest difference judgments. Technical report, University of Alberta.

Minh-Thang Luong, Richard Socher, and Christopher D. Manning. 2013. Better word representations with recursive neural networks for morphology. In *Proceedings of CoNLL*, pages 104–113.

Stephen J. Lupker. 1984. Semantic priming without association: A second look. *Journal of Verbal Learning and Verbal Behavior*, 23(6):709–733.

John Lyons. 1968. *Introduction to Theoretical Linguistics*, volume 510. Cambridge University Press.

James MacQueen. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297.

Olga Majewska, Diana McCarthy, Jasper van den Bosch, Nikolaus Kriegeskorte, Ivan Vulić, and Anna Korhonen. 2020a. Spatial multi-arrangement for clustering and multi-way similarity dataset construction. In *Proceedings of LREC*, pages 5751–5760.

Olga Majewska, Diana McCarthy, Jasper van den Bosch, Nikolaus Kriegeskorte, Ivan Vulić, and Anna Korhonen. 2021a. Semantic data set construction from human clustering and spatial arrangement. *Computational Linguistics*, 47(1):69–116.

Olga Majewska, Diana McCarthy, Ivan Vulić, and Anna Korhonen. 2018a. Acquiring verb classes through bottom-up semantic verb clustering. In *Proceedings of LREC*, pages 952–958.

Olga Majewska, Ivan Vulić, Goran Glavaš, Edoardo Maria Ponti, and Anna Korhonen. 2021b. Verb knowledge injection for multilingual event processing. In *Proceedings of ACL*, pages 6952–6969.

Olga Majewska, Ivan Vulić, Diana McCarthy, Yan Huang, Akira Murakami, Veronika Laippala, and Anna Korhonen. 2018b. Investigating the cross-lingual translatability of VerbNet-style classification. *Language Resources and Evaluation*, 52(3):771–799.

Olga Majewska, Ivan Vulić, Diana McCarthy, and Anna Korhonen. 2020b. Manual clustering and spatial arrangement of verbs for multilingual evaluation and typology analysis. In *Proceedings of COLING*, pages 4810–4824.

Rila Mandala, Takenobu Tokunaga, and Hozumi Tanaka. 1998. The use of WordNet in information retrieval. In *Usage of WordNet in Natural Language Processing Systems*, pages 31–37.

Nathan Mantel. 1967. The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27(2 Part 1):209–220.

Arthur B. Markman and Dedre Gentner. 1993. Structural alignment during similarity comparisons. *Cognitive Psychology*, 25(4):431–467.

Arthur B. Markman and Edward J. Wisniewski. 1997. Similar and different: The differentiation of basic-level categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23(1):54.

Lara Martin, Prithviraj Ammanabrolu, Xinyu Wang, William Hancock, Shruti Singh, Brent Harrison, and Mark Riedl. 2018. Event representations for automated story generation with deep neural nets. *Proceedings of AAAI*, 32(1).

Yuval Marton, Chris Callison-Burch, and Philip Resnik. 2009. Improved statistical machine translation using monolingually-derived paraphrases. In *Proceedings of EMNLP*, pages 381–390.

Cyprien de Masson d'Autume, Sebastian Ruder, Lingpeng Kong, and Dani Yogatama. 2019. Episodic memory in lifelong language learning. In *Proceedings of NeurIPS*, pages 13122–13131.

Jiří Materna. 2012. LDA-Frames: An unsupervised approach to generating semantic frames. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 376–387.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. In *Proceedings of NeurIPS*, pages 6294–6305.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of ACL*, pages 3428–3448.

Ken McRae, Todd R. Ferretti, and Liane Amyote. 1997. Thematic roles as verb-specific concepts. *Language and Cognitive Processes*, 12(2-3):137–176.

Ken McRae, Saman Khalkhali, and Mary Hare. 2012. Semantic and associative relations in adolescents and young adults: Examining a tenuous dichotomy. In V. F. Reyna, S. B. Chapman, M. R. Dougherty, and J. Confrey, editors, *The Adolescent Brain: Learning, Reasoning, and Decision Making*, pages 39–66. American Psychological Association.

Marina Meila and Jianbo Shi. 2001. A random walks view of spectral segmentation. In *AISTATS*.

Oren Melamud, David McClosky, Siddharth Patwardhan, and Mohit Bansal. 2016. The role of context types and dimensionality in learning word embeddings. In *Proceedings of NAACL-HLT*, pages 1030–1040.

Paola Merlo and Suzanne Stevenson. 2001. Automatic verb classification based on statistical distributions of argument structure. *Computational Linguistics*, 27(3):373–408.

Meryem M'hamdi, Marjorie Freedman, and Jonathan May. 2019. Contextualized cross-lingual event trigger extraction with minimal resources. In *Proceedings of CoNLL*, pages 656–665.

Rada Mihalcea and Dan I. Moldovan. 2000. AutoASC – a system for automatic acquisition of sense tagged corpora. *International Journal of Pattern Recognition and Artificial Intelligence*, 14(01):3–17.

Nives Mikelić Preradović and Damir Boras. 2013. Semi-automatic verb valence frame assignment through VerbNet classification. In *Proceedings of TSD*, pages 492–500.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of LREC*, pages 52–55.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Proceedings of NeurIPS*, pages 3111–3119.

Dmitrijs Milajevs and Sascha Griffiths. 2016. A proposal for linguistic similarity datasets based on commonality lists. In *Proceedings of REPEVAL*, pages 127–133.

George A. Miller. 1995. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41.

George A. Miller and Walter G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

Eleni Miltsakaki. 2009. Matching readers' preferences and reading skills with appropriate web texts. In *Proceedings of EACL*, pages 49–52.

Eleni Miltsakaki, Rashmi Prasad, Aravind K. Joshi, and Bonnie L. Webber. 2004. The Penn Discourse Treebank. In *Proceedings of LREC*.

Anne-Lyse Minard, Manuela Speranza, Ruben Urizar, Begona Altuna, Marieke Van Erp, Anneleen Schoen, and Chantal Van Son. 2016. MEANTIME, the NewsReader multilingual event and time corpus. In *Proceedings of LREC*, pages 4417–4422.

Andriy Mnih and Koray Kavukcuoglu. 2013. Learning word embeddings efficiently with noise-contrastive estimation. In *Proceedings of NIPS*, pages 2265–2273.

Dan I. Moldovan and Rada Mihalcea. 2000. Using WordNet and lexical operators to improve internet searches. *IEEE Internet Computing*, 4(1):34–43.

Padraic Monaghan, Morten H. Christiansen, and Stanka A. Fitneva. 2011. The arbitrariness of the sign: Learning advantages from the structure of the vocabulary. *Journal of Experimental Psychology: General*, 140(3):325.

Jaouad Mousser. 2010. A large coverage verb taxonomy for Arabic. In *Proceedings of LREC*, pages 2675–2681.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. Counter-fitting word vectors to linguistic constraints. In *Proceedings of NAACL–HLT*, pages 142–148.

Nikola Mrkšić, Ivan Vulić, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young. 2017. Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the ACL*, 5:309–324.

David Mueller, Nicholas Andrews, and Mark Dredze. 2020. Sources of transfer in multilingual named entity recognition. In *Proceedings of ACL*, pages 8093–8104.

Marieke Mur, Mirjam Meys, Jerzy Bodurka, Rainer Goebel, Peter A. Bandettini, and Nikolaus Kriegeskorte. 2013. Human object-similarity judgments reflect and transcend the primate-IT object representation. *Frontiers in Psychology*, 4:128.

Agnieszka Mykowiecka, Małgorzata Marciniak, and Piotr Rychlik. 2018. SimLex-999 for Polish. In *Proceedings of LREC*, pages 2398–2402.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of ACL*, pages 216–225.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Claire Nédellec, Robert Bossy, Jin-Dong Kim, Jung-jae Kim, Tomoko Ohta, Sampo Pyysalo, and Pierre Zweigenbaum. 2013. Overview of BioNLP shared task 2013. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 1–7.

Vladimir P. Nedjalkov. 1976. Diathesen und Satzstruktur im Tschuktschischen. *Studia Grammatica*, 13:181–213.

Douglas L. Nelson, Cathy L. McEvoy, and Thomas A. Schreiber. 2004. The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, 36(3):402–407.

Tamás Nepusz, Haiyuan Yu, and Alberto Paccanaro. 2012. Detecting overlapping protein complexes in protein-protein interaction networks. *Nature Methods*, 9(5):471.

Ponrudee Netisopakul, Gerhard Wohlgenannt, and Aleksei Pulich. 2019. Word similarity datasets for Thai: Construction and evaluation. *IEEE Access*, 7:142907–142915.

Mark E.J. Newman. 2005. A measure of betweenness centrality based on random walks. *Social Networks*, 27(1):39–54.

Vuong M. Ngo, Tru H. Cao, and Tuan Le. 2018. WordNet-based information retrieval using common hypernyms and combined features. *arXiv preprint arXiv:1807.05574*.

Kim Anh Nguyen, Maximilian Köper, Sabine Schulte im Walde, and Ngoc Thang Vu. 2017. Hierarchical embeddings for hypernymy detection and directionality. In *Proceedings of EMNLP*, pages 233–243.

Kim Anh Nguyen, Sabine Schulte im Walde, and Ngoc Thang Vu. 2016a. Integrating distributional lexical contrast into word embeddings for antonym-synonym distinction. In *Proceedings of ACL*, pages 454–459.

Thien Huu Nguyen, Kyunghyun Cho, and Ralph Grishman. 2016b. Joint event extraction via recurrent neural networks. In *Proceedings of NAACL–HLT*, pages 300–309.

Thien Huu Nguyen and Ralph Grishman. 2015. Event detection and domain adaptation with convolutional neural networks. In *Proceedings of ACL–IJCNLP*, pages 365–371.

Thien Huu Nguyen and Ralph Grishman. 2018. Graph convolutional networks with argument-aware pooling for event detection. In *AAAI*, volume 18, pages 5900–5907.

Hamed Nili, Cai Wingfield, Alexander Walther, Li Su, William Marslen-Wilson, and Nikolaus Kriegeskorte. 2014. A toolbox for representational similarity analysis. *PLoS Computational Biology*, 10(4):e1003553.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of ACL*, pages 4658–4664.

Robert M. Nosofsky. 1992. Similarity scaling and cognitive process models. *Annual review of Psychology*, 43(1):25–53.

Laura R. Novick. 1988. Analogical transfer, problem similarity, and expertise. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14(3):510.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2020. What the [MASK]? Making sense of language-specific BERT models. *CoRR*, abs/2003.02912.

Diarmuid Ó Séaghdha and Ann Copestake. 2008. Semantic classification with distributional kernels. In *Proceedings of COLING*, pages 649–656.

Kyoko Ohara. 2012. Semantic annotations in Japanese FrameNet: Comparing frames in Japanese and English. In *Proceedings of LREC*, pages 1559–1562.

Masataka Ono, Makoto Miwa, and Yutaka Sasaki. 2015. Word embedding-based antonym detection using thesauri and distributional information. In *Proceedings of NAACL–HLT*, pages 984–989.

Bryan Orme. 2009. MaxDiff analysis: Simple counting, individual-level logit, and HB. *Sawtooth Software*.

Dominique Osborne, Shashi Narayan, and Shay B. Cohen. 2016. Encoding prior knowledge with eigenword embeddings. *Transactions of the ACL*, 4:417–430.

Charles E. Osgood. 1952. The nature and measurement of meaning. *Psychological Bulletin*, 49(3):197.

Pinar Öztürk, Mila Vulchanova, Christian Tumyr, Liliana Martinez, and David Kabath. 2011. Assesing the feature-driven nature of similarity-based sorting of verbs. *Polibits*, (43):15–22.

Sebastian Padó and Mirella Lapata. 2007. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

Karel Pala and Aleš Horák. 2008. Can complex valency frames be universal? In *Proceedings of RASLAN*, pages 41–49.

Martha Palmer. 2009. Semlink: Linking PropBank, VerbNet and FrameNet. In *Proceedings of the Generative Lexicon Conference*, pages 9–15.

Martha Palmer, Paul Kingsbury, and Daniel Gildea. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Alexander Panchenko, Dmitry Ustalov, Nikolay Arefyev, Denis Paperno, Natalia Konstantinova, Natalia Loukachevitch, and Chris Biemann. 2016. Human and machine judgements for Russian semantic relatedness. In *International Conference on Analysis of Images, Social Networks and Texts*, pages 221–235.

Anna Papafragou, Kimberly Cassidy, and Lila Gleitman. 2007. When we think about thinking: The acquisition of belief verbs. *Cognition*, 105(1):125–165.

Allen Parducci. 1965. Category judgment: A range-frequency model. *Psychological Review*, 72(6):407.

Terence Parsons. 1990. *Events in the Semantics of English*, volume 5. MIT press Cambridge, MA.

Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2015. PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proceedings of ACL–IJCNLP*, pages 425–430.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of EMNLP*, pages 1532–1543.

David Pesetsky. 1987. Binding problems with experiencer verbs. *Linguistic Inquiry*, 18(1):126–140.

Matthew Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. In *Proceedings of ACL*, pages 1756–1765.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL–HLT*, pages 2227–2237.

Matthew Peters, Mark Neumann, Robert L. Logan IV, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge enhanced contextual word representations. In *Proceedings of EMNLP–IJCNLP*, pages 43–54.

Daniel Peterson, Jordan Boyd-Graber, Martha Palmer, and Daisuke Kawahara. 2016. Leveraging VerbNet to build corpus-specific verb clusters. In *Proceedings of the 5th Joint Conference on Lexical and Computational Semantics*, pages 102–107.

Daniel Peterson, Susan Brown, and Martha Palmer. 2020. Verb class induction with partial supervision. In *Proceedings of AAAI*, volume 34, pages 8616–8623.

Daniel Peterson and Martha Palmer. 2018. Bayesian verb sense clustering. In *Proceedings of AAAI*, volume 32, pages 5398–5405.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2020a. AdapterFusion: Non-destructive task composition for transfer learning. *arXiv preprint arXiv:2005.00247*.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020b. AdapterHub: A framework for adapting transformers. In *Proceedings of EMNLP*, pages 46–54.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020c. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of EMNLP*, pages 7654–7673.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: The word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of NAACL–HLT*, pages 1267–1273.

Mohammad Taher Pilehvar, Dimitri Kartsaklis, Victor Prokhorov, and Nigel Collier. 2018. Card-660: Cambridge Rare Word Dataset - a reliable benchmark for infrequent word representation models. In *Proceedings of EMNLP*, pages 1391–1401.

Steven Pinker. 1989. *Learnability and Cognition: The Acquisition of Argument Structure*. MIT press.

Steven Pinker. 1994. How could a child use verb syntax to learn verb semantics. *Lingua*, 92(1-4):377–410.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of ACL*, pages 4996–5001.

Claudia Plant, Stefan J. Teipel, Annahita Oswald, Christian Böhm, Thomas Meindl, Janaina Mourao-Miranda, Arun W. Bokde, Harald Hampel, and Michael Ewers. 2010. Automated detection of brain atrophy patterns based on MRI for the prediction of Alzheimer's disease. *Neuroimage*, 50(1):162–174.

Carl Pollard and Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. University of Chicago Press.

Edoardo Maria Ponti and Anna Korhonen. 2017. Event-related features in feedforward neural networks contribute to identifying causal relations in discourse. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 25–30.

Edoardo Maria Ponti, Ivan Vulić, Goran Glavaš, Nikola Mrkšić, and Anna Korhonen. 2018. Adversarial propagation and zero-shot cross-lingual transfer of word vector specialization. In *Proceedings of EMNLP*, pages 282–293.

Edoardo Maria Ponti, Ivan Vulić, Goran Glavaš, Roi Reichart, and Anna Korhonen. 2019. Cross-lingual semantic specialization via lexical relation induction. In *Proceedings of EMNLP–IJCNLP*, pages 2206–2217.

Alexander Popov, Kiril Simov, and Petya Osenova. 2019. Know your graph. State-of-the-art knowledge-based WSD. In *Proceedings of RANLP*, pages 949–958.

Quentin Pradet, Laurence Danlos, and Gaël de Chalendar. 2014. Adapting VerbNet to French using existing resources. In *Proceedings of LREC*, pages 1122–1126.

Nives Mikelic Preradovic, Damir Boras, and Sanja Kisicek. 2009. CROVALLEX: Croatian verb valence lexicon. In *Proceedings of the 31st International Conference on Information Technology Interfaces*, pages 533–538.

James Pustejovsky, José M. Castano, Robert Ingria, Roser Sauri, Robert J. Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R. Radev. 2003. TimeML: Robust specification of event and temporal expressions in text. *New Directions in Question Answering*, 3:28–34.

James Pustejovsky, Robert Ingria, Roser Sauri, José M. Castaño, Jessica Littman, Robert J. Gaizauskas, Andrea Setzer, Graham Katz, and Inderjeet Mani. 2005. The specification language TimeML. In *The Language of Time: A reader*, pages 545–557.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *SCIENCE CHINA Technological Sciences*.

George B. Rabinowitz. 1975. An introduction to nonmetric multidimensional scaling. *American Journal of Political Science*, pages 343–390.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *OpenAI Technical Report*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich, and Shaul Markovitch. 2011. A word at a time: Computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th International Conference on World Wide Web*, pages 337–346.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Altaf Rahman and Vincent Ng. 2011. Coreference resolution with world knowledge. In *Proceedings of ACL–HLT*, pages 814–824.

Bo Ralph. 1980. Relative semantic complexity in lexical units. In *Proceedings of COLING*, pages 115–121.

Vasumathi Raman, Constantine Lignos, Cameron Finucane, Kenton Lee, Mitch Marcus, and Hadas Kress-Gazit. 2013. Sorry Dave, I'm afraid i can't do that: Explaining unachievable robot tasks using natural language. In *Robotics: Science and Systems*.

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. 2018. Efficient parametrization of multi-domain deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8119–8127.

Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of IJCAI*, pages 448–453.

Philip Resnik and Mona T. Diab. 2000. Measuring verb similarity. In *Proceedings of the 22nd Annual Meeting of the Cognitive Science Society (CogSci 2000)*, pages 399–404.

Laura Rimell, Thomas Lippincott, Karin Verspoor, Helen L. Johnson, and Anna Korhonen. 2013. Acquisition and evaluation of verb subcategorization resources for biomedicine. *Journal of Biomedical Informatics*, 46(2):228–237.

Miguel Rios, Wilker Aziz, and Lucia Specia. 2011. TINE: A metric to assess MT adequacy. In *Proceedings of the 6th Workshop on Statistical Machine Translation*, pages 116–122.

Alan Ritter, Oren Etzioni, and Sam Clark. 2012. Open domain event extraction from Twitter. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1104–1112.

Anna Rogers, Olga Kovaleva, Matthew Downey, and Anna Rumshisky. 2020a. Getting closer to AI complete question answering: A set of prerequisite real tasks. In *Proceedings of AAAI*, pages 8722–8731.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020b. A primer in BERTology: what we know about how BERT works. *Transactions of the ACL*.

Peter Mark Roget. 1911. *Roget's Thesaurus of English Words and Phrases*. TY Crowell Company.

Brian H. Ross. 1987. This is like that: The use of earlier problems and the separation of similarity effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13(4):629.

Brian H. Ross. 1989. Distinguishing types of superficial similarities: Different effects on the access and use of earlier problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15(3):456.

Herbert Rubenstein and John B. Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

Abdulgabbar Saif, Mohd Juzaiddin Ab Aziz, and Nazlia Omar. 2014. Evaluating knowledge-based semantic measures on Arabic. *International Journal on Communications Antenna and Propagation*, 4(5):180–194.

Manfred Sailer and Stella Markantonatou. 2018. *Multiword Expressions: Insights from a Multi-lingual Perspective*. Language Science Press.

Yuya Sakaizawa and Mamoru Komachi. 2018. Construction of a Japanese word similarity dataset. In *Proceedings of LREC*, pages 948–951.

Claude Sammut and Geoffrey I. Webb. 2011. *Encyclopedia of Machine Learning*. Springer Science & Business Media.

Edward Sapir. 1985. *Culture, Language and Personality: Selected Essays*, volume 342. University of California Press.

Sebastian Sauppe. 2016. Verbal semantics drives early anticipatory eye movements during the comprehension of verb-initial sentences. *Frontiers in Psychology*, 7:95.

Ferdinand de Saussure. 1916. *Course in General Linguistics*. McGraw-Hill Book Company.

Carolina Scarton and Sandra Alusio. 2012. Towards a cross-linguistic VerbNet-style lexicon for Brazilian Portuguese. In *Proceedings of the LREC 2012 Workshop on Creating Cross-language Resources for Disconnected Languages and Styles*, pages 11–18.

Carolina Scarton, Lin Sun, Karin Kipper Schuler, Magali Sanches Duran, Martha Palmer, and Anna Korhonen. 2014. Verb clustering for Brazilian Portuguese. In *Proceedings of CICLing*, pages 25–39.

Michael Schmitz, Robert Bart, Stephen Soderland, Oren Etzioni, et al. 2012. Open language learning for information extraction. In *Proceedings of EMNLP–CoNLL*, pages 523–534.

Patrick Schober, Christa Boer, and Lothar A. Schwarte. 2018. Correlation coefficients: Appropriate use and interpretation. *Anesthesia & Analgesia*, 126(5):1763–1768.

Erich Schubert, Jörg Sander, Martin Ester, Hans Peter Kriegel, and Xiaowei Xu. 2017. DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems (TODS)*, 42(3):1–21.

Roy Schwartz, Roi Reichart, and Ari Rappoport. 2015. Symmetric pattern based word embeddings for improved word similarity prediction. In *Proceedings of CoNLL*, pages 258–267.

John R. Searle and Daniel Vanderveken. 1985. *Foundations of Illocutionary Logic*. CUP Archive.

Joao Sedoc, Derry Wijaya, Masoud Rouhizadeh, Andy Schwartz, and Lyle Ungar. 2017. Deriving verb predicates by clustering verbs with arguments. *arXiv preprint arXiv:1708.00416*.

Lei Sha, Feng Qian, Baobao Chang, and Zhifang Sui. 2018. Jointly extracting event triggers and arguments by dependency-bridge RNN and tensor-based argument interaction. In *Proceedings of AAAI*, pages 5916–5923.

Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. 2003. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11):2498–2504.

Lanbo She and Joyce Chai. 2017. Interactive learning of grounded verb semantics towards human-robot communication. In *Proceedings of ACL*, pages 1634–1644.

Roger N. Shepard. 1962. The analysis of proximities: Multidimensional scaling with an unknown distance function. *Psychometrika*, 27(2):125–140.

Roger N. Shepard. 1980. Multidimensional scaling, tree-fitting, and clustering. *Science*, 210(4468):390–398.

Roger N. Shepard. 1987. Toward a universal law of generalization for psychological science. *Science*, 237(4820):1317–1323.

Lei Shi and Rada Mihalcea. 2005. Putting pieces together: Combining FrameNet, VerbNet and WordNet for robust semantic parsing. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 100–111.

Vered Shwartz, Omer Levy, Ido Dagan, and Jacob Goldberger. 2015. Learning to exploit structured resources for lexical inference. In *Proceedings of CoNLL*, pages 175–184.

Chenglei Si, Shuohang Wang, Min-Yen Kan, and Jing Jiang. 2019. What does BERT learn from multiple-choice reading comprehension datasets? *arXiv preprint arXiv:1910.12391*.

W. Kyle Simmons, Stephan B. Hamann, Carla L. Harenski, Xiaoping P. Hu, and Lawrence W. Barsalou. 2008. fMRI evidence for word association and situated simulation in conceptual processing. *Journal of Physiology*, 102(1-3):106–119.

Matthew Sims, Jong Ho Park, and David Bamman. 2019. Literary event detection. In *Proceedings of ACL*, pages 3623–3634.

Push Singh, Thomas Lin, Erik T Mueller, Grace Lim, Travell Perkins, and Wan Li Zhu. 2002. Open Mind Common Sense: Knowledge acquisition from the general public. In *OTM Confederated International Conferences "On the Move to Meaningful Internet Systems"*, pages 1223–1237.

Dan I. Slobin. 1996. From "thought and language" to "thinking for speaking". In J. Gumperz and S. Levinson, editors, *Rethinking Linguistic Relativity*, pages 70–96. Cambridge University Press.

Małgorzata Smereczniak. 2018. Czas i aspekt jako podstawowe kategorie czasowników. Wybrane zagadnienia w ujęciu glottodydaktycznym. In Marcin Maciołek, editor, *Tęczowa gramatyka języka polskiego w tabelach*, pages 26–32. Szkoła Języka i Kultury Polskiej, Uniwersytet Śląski.

Neal Snider and Mona Diab. 2006. Unsupervised induction of modern standard Arabic verb classes using syntactic frames and LSA. In *Proceedings of ACL–COLING*, pages 795–802.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and fast – but is it good? Evaluating non-expert annotations for natural language tasks. In *Proceedings of EMNLP*, pages 254–263.

Zhiyi Song, Ann Bies, Stephanie Strassel, Tom Riese, Justin Mott, Joe Ellis, Jonathan Wright, Seth Kulick, Neville Ryant, and Xiaoyi Ma. 2015. From light to rich ERE: Annotation of entities, relations, and events. In *Proceedings of the The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 89–98.

Ugur Sopaoglu and Gonenc Ercan. 2016. Evaluation of semantic relatedness measures for Turkish language. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 600–611.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An open multilingual graph of general knowledge. In *Proceedings of AAAI*, volume 31.

Rohini K. Srihari, Zhongfei Zhang, and Aibing Rao. 2000. Intelligent indexing and semantic retrieval of multimodal documents. *Information Retrieval*, 2(2):245–275.

Sofia Stamou, Kemal Oflazer, Karel Pala, Dimitris Christoudoulakis, Dan Cristea, Dan Tufis, Svetla Koeva, George Totkov, Dominique Dutoit, and Maria Grigoriadou. 2002. BalkaNet: A multilingual semantic network for the Balkan languages. In *Proceedings of the International Wordnet Conference*, pages 21–25.

Carlos Subirats and Hiroaki Sato. 2004. Spanish FrameNet and FrameSQL. In *Workshop on Building Lexical Resources from Semantically Annotated Corpora (LREC)*.

Saku Sugawara, Pontus Stenetorp, Kentaro Inui, and Akiko Aizawa. 2020. Assessing the benchmarking capacity of machine reading comprehension datasets. In *Proceedings of AAAI*, pages 8918–8927.

Lin Sun. 2013. *Automatic Induction of Verb Classes Using Clustering*. Ph.D. thesis, University of Cambridge.

Lin Sun and Anna Korhonen. 2009. Improving verb clustering with automatically acquired selectional preferences. In *Proceedings of EMNLP*, pages 638–647.

Lin Sun, Anna Korhonen, and Yuval Krymolowski. 2008a. Automatic classification of English verbs using rich syntactic features. In *Proceedings of IJCNLP*, pages 769–774.

Lin Sun, Anna Korhonen, and Yuval Krymolowski. 2008b. Verb class discovery from rich syntactic data. In *Proceedings of CICLing*, pages 16–27.

Lin Sun, Anna Korhonen, Thierry Poibeau, and Cédric Messiant. 2010. Investigating the cross-linguistic potential of VerbNet-style classification. In *Proceedings of COLING*, pages 1056–1064.

Lin Sun, Diana McCarthy, and Anna Korhonen. 2013. Diathesis alternation approximation for verb clustering. In *Proceedings of ACL*, pages 736–741.

Robert S. Swier and Suzanne Stevenson. 2004. Unsupervised semantic role labelling. In *Proceedings of EMNLP*, pages 95–102.

Mohamed Ali Hadj Taieb, Torsten Zesch, and Mohamed Ben Aouicha. 2020. A survey of semantic relatedness evaluation datasets and procedures. *Artificial Intelligence Review*, 53(6):4407–4448.

Leonard Talmy. 1985. Lexicalization patterns: Semantic structure in lexical forms. *Language Typology and Syntactic Description*, 3:57–149.

Leonard Talmy. 2000. *Toward a Cognitive Semantics*, volume 2. MIT press.

Donna Tatsuki. 1998. Basic 2000 words – synonym match 1. *Interactive JavaScript Quizzes for ESL Students*.

Wilson L. Taylor. 1953. "Cloze procedure": A new tool for measuring readability. *Journalism Bulletin*, 30(4):415–433.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019a. BERT rediscovers the classical NLP pipeline. In *Proceedings of ACL*, pages 4593–4601.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, and Dipanjan Das. 2019b. What do you learn from context? Probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316*.

Lucien Tesnière. 1959. *Éléments de syntaxe structurale.* Klincksieck.

Paul Thompson, Syed A. Iqbal, John McNaught, and Sophia Ananiadou. 2009. Construction of an annotated corpus to support biomedical information extraction. *BMC Bioinformatics*, 10(1):349.

Robert Tibshirani, Guenther Walther, and Trevor Hastie. 2001. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423.

Antonio Toral, Stefania Bracale, Monica Monachini, Claudia Soria, et al. 2010. Rejuvenating the Italian WordNet: Upgrading, standardising, extending. In *Proceedings of the 5th International Global WordNet Conference.*

Tiago T. Torrent, Collin F. Baker, Oliver Czulo, Kyoko Ohara, and Miriam R.L. Petruck. 2020. *Proceedings of the International FrameNet Workshop 2020: Towards a Global, Multilingual FrameNet.* ELRA.

Tiago Timponi Torrent, Maria Margarida Salomão, Fernanda Campos, Regina Braga, Ely Matos, Maucha Gamonal, Julia Gonçalves, Bruno Souza, Daniela Gomes, and Simone Peron. 2014. Copa 2014 FrameNet Brasil: A frame-based trilingual electronic dictionary for the Football World Cup. In *Proceedings of COLING*, pages 10–14.

Ágoston Tóth. 2013. How similar: Word similarity judgments in english and Hungarian. Technical report, Technical report.

Julien Tourille. 2018. *Extracting Clinical Event Timelines: Temporal Information Extraction and Coreference Resolution in Electronic Health Records.* Ph.D. thesis, Université Paris-Saclay.

Thanh N. Tran, Ron Wehrens, and Lutgarde M.C. Buydens. 2006. KNN-kernel density-based clustering for high-dimensional multivariate data. *Computational Statistics & Data Analysis*, 51(2):513–525.

Yulia Tsvetkov, Manaal Faruqui, Wang Ling, Guillaume Lample, and Chris Dyer. 2015. Evaluation of word vector representations by subspace alignment. In *Proceedings of EMNLP*, pages 2049–2054.

John C. Turner, Michael A. Hogg, Penelope J. Oakes, Stephen D. Reicher, and Margaret S. Wetherell. 1987. *Rediscovering the Social Group: A Self-Categorization Theory.* Basil Blackwell.

Peter D. Turney. 2001. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of ECML*, pages 491–502.

Peter D. Turney. 2006. Similarity of semantic relations. *Computational Linguistics*, 32(3):379–416.

Peter D. Turney. 2012. Domain and function: A dual-space model of semantic relations and compositions. *Journal of Artificial Intelligence research*, 44:533–585.

Peter D. Turney, Michael L. Littman, Jeffrey Bigham, and Victor Shnayder. 2003. Combining independent modules to solve multiple-choice synonym and analogy problems. *arXiv preprint cs/0309035*.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

Amos Tversky. 1977. Features of similarity. *Psychological Review*, 84(4):327.

Amos Tversky and Itamar Gati. 1978. Studies of similarity. *Cognition and Categorization*.

Amos Tversky and Itamar Gati. 1982. Similarity, separability, and the triangle inequality. *Psychological review*, 89(2):123.

Mohammed Nazim Uddin, Trong Hai Duong, Ngoc Thanh Nguyen, Xin-Min Qi, and Geun Sik Jo. 2013. Semantic similarity measures for enhancing information retrieval in folksonomies. *Expert Systems with Applications*, 40(5):1645–1653.

Zdenka Uresova, Eva Fucíková, Eva Hajicová, and Jan Hajic. 2018. Creating a verb synonym lexicon based on a parallel corpus. In *Proceedings of LREC*, pages 1432–1437.

Zdeňka Urešová, Eva Fučíková, and Jana Šindlerová. 2016. CzEngVallex: A bilingual Czech-English valency lexicon. *The Prague Bulletin of Mathematical Linguistics*, 105(1):17–50.

Zdeňka Urešová, Jan Štěpánek, Jan Hajič, Jarmila Panevova, and Marie Mikulová. 2014. PDT-vallex: Czech valency lexicon linked to treebanks. Technical report, Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics.

Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. 2013. SemEval-2013 task 1: TempEval-3: Evaluating time expressions, events, and temporal relations. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*, pages 1–9.

Bui Van Tan, Nguyen Phuong Thai, and Pham Van Lam. 2017. Construction of a word similarity dataset and evaluation of word similarity techniques for Vietnamese. In *9th International Conference on Knowledge and Systems Engineering (KSE)*, pages 65–70.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of NeurIPS*, pages 6000–6010.

Zeno Vendler. 1967. *Linguistics in Philosophy*. Cornell University Press.

Zeno Vendler. 1972. *Res cogitans*. Cornell University Press.

Marc Verhagen and James Pustejovsky. 2008. Temporal processing with the TARSQI toolkit. In *COLING 2008: Companion volume: Demonstrations*, pages 189–192.

Marc Verhagen, Roser Saurí, Tommaso Caselli, and James Pustejovsky. 2010. SemEval-2010 task 13: TempEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 57–62.

Zygmunt Vetulani, Marek Kubis, and Tomasz Obrębski. 2010. PolNet — Polish WordNet: Data and tools. In *Proceedings of LREC*, pages 3793–3797.

Loïc Vial, Benjamin Lecouteux, and Didier Schwab. 2018. Improving the coverage and the generalization ability of neural word sense disambiguation through hypernymy and hyponymy relationships. *arXiv preprint arXiv:1811.00960*.

Åke Viberg. 1984. The verbs of perception: A typological study. *Explanations for Language Universals*, pages 123–162.

V. Vijayarajan, M. Dinakaran, Priyam Tejaswin, and Mayank Lohani. 2016. A generic framework for ontology-based information retrieval and image retrieval in web data. *Human-centric Computing and Information Sciences*, 6(1):18.

Elisabeth Villalta. 2000. Spanish subjunctive clauses require ordered alternatives. In *Semantics and Linguistic Theory*, volume 10, pages 239–256.

Elisabeth Villalta. 2008. Mood and gradability: An investigation of the subjunctive mood in spanish. *Linguistics and Philosophy*, 31(4):467–522.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: BERT for Finnish. *CoRR*, abs/1912.07076.

Andreas Vlachos, Anna Korhonen, and Zoubin Ghahramani. 2009. Unsupervised and constrained Dirichlet process mixture models for verb clustering. In *Proceedings of the EACL Workshop on Geometrical Models of Natural Language Semantics*, pages 74–82.

Ralf Vogel. 2016. Optimal constructions. In G. Legendre, M. Putnam, H. de Swart, and E. Zaroukian, editors, *Optimality-Theoretic Syntax, Semantics, and Pragmatics: From Uni- to Bidirectional Optimization*, pages 55–77. Oxford University Press.

Stella Vosniadou and Andrew Ortony. 1989. *Similarity and Analogical Reasoning*. Cambridge University Press.

Piek Vossen. 1998. A multilingual database with lexical semantic networks. *Dordrecht: Kluwer Academic Publishers*, 10:978–94.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Ivan Vulić, Simon Baker, Edoardo Maria Ponti, Ulla Petti, Ira Leviant, Kelly Wing, Olga Majewska, Eden Bar, Matt Malone, Thierry Poibeau, et al. 2020a. Multi-SimLex: A large-scale evaluation of multilingual and cross-lingual lexical semantic similarity. *Computational Linguistics*, pages 1–51.

Ivan Vulić, Goran Glavaš, Nikola Mrkšić, and Anna Korhonen. 2018. Post-specialisation: Retrofitting vectors of words unseen in lexical resources. In *Proceedings of NAACL–HLT*, pages 516–527.

Ivan Vulić, Douwe Kiela, and Anna Korhonen. 2017a. Evaluation by association: A systematic study of quantitative word association evaluation. In *Proceedings of EACL*, pages 163–175.

Ivan Vulić and Anna Korhonen. 2018. Injecting lexical contrast into word vectors by guiding vector space specialisation. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 137–143.

Ivan Vulić and Nikola Mrkšić. 2018. Specialising word vectors for lexical entailment. In *Proceedings of NAACL–HLT*, pages 1134–1145.

Ivan Vulić, Nikola Mrkšić, and Anna Korhonen. 2017b. Cross-lingual induction and transfer of verb classes based on word vector space specialisation. In *Proceedings of EMNLP*, pages 2546–2558.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020b. Probing pretrained language models for lexical semantics. In *Proceedings of EMNLP*, pages 7222–7240.

Ivan Vulić, Simone Paolo Ponzetto, and Goran Glavaš. 2019. Multilingual and cross-lingual graded lexical entailment. In *Proceedings of ACL*, pages 4963–4974.

Ivan Vulić, Roy Schwartz, Ari Rappoport, Roi Reichart, and Anna Korhonen. 2017c. Automatic selection of context configurations for improved class-specific word representations. In *Proceedings of CoNLL*, pages 112–122.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, relation, and event extraction with contextualized span representations. In *Proceedings of EMNLP–IJCNLP*, pages 5784–5789.

Chang Wang, Aditya Kalyanpur, James Fan, Branimir K. Boguraev, and D.C. Gondek. 2012a. Relation extraction and scoring in DeepQA. *IBM Journal of Research and Development*, 56(3.4):9–1.

Jianxin Wang, Min Li, Jianer Chen, and Yi Pan. 2011a. A fast hierarchical clustering algorithm for functional modules discovery in protein interaction networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(3):607–620.

Jianxin Wang, Jun Ren, Min Li, and Fang-Xiang Wu. 2012b. Identification of hierarchical and overlapping functional modules in PPI networks. *IEEE Transactions on NanoBioscience*, 11(4):386–393.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Cuihong Cao, Daxin Jiang, Ming Zhou, et al. 2020a. K-Adapter: Infusing knowledge into pretrained models with adapters. *arXiv preprint arXiv:2002.01808*.

Shike Wang, Yuchen Fan, Xiangying Luo, and Dong Yu. 2020b. SHIKEBLCU at SemEval-2020 task 2: An external knowledge-enhanced matrix for multilingual and cross-lingual lexical entailment. In *Proceedings of the 14th Workshop on Semantic Evaluation*, pages 255–262.

Xiang Wang, Yan Jia, Bin Zhou, Zhao-Yun Ding, and Zheng Liang. 2011b. Computing semantic relatedness using Chinese Wikipedia links and taxonomy. *Journal of Chinese Computer Systems*, 32(11):2237–2242.

Xiaozhi Wang, Xu Han, Zhiyuan Liu, Maosong Sun, and Peng Li. 2019a. Adversarial training for weakly supervised event detection. In *Proceedings of NAACL–HLT*, pages 998–1008.

Xiaozhi Wang, Ziqi Wang, Xu Han, Wangyi Jiang, Rong Han, Zhiyuan Liu, Juanzi Li, Peng Li, Yankai Lin, and Jie Zhou. 2020c. MAVEN: A Massive General Domain Event Detection Dataset. In *Proceedings of EMNLP*, pages 1652–1671.

Yuxuan Wang, Wanxiang Che, Jiang Guo, Yijia Liu, and Ting Liu. 2019b. Cross-lingual BERT transformation for zero-shot dependency parsing. In *Proceedings of EMNLP–IJCNLP*, pages 5721–5727.

Douglas H Wedell. 1995. Contrast effects in paired comparisons: Evidence for both stimulus-based and response-based processes. *Journal of Experimental Psychology: Human Perception and Performance*, 21(5):1158.

Julie Elizabeth Weeds. 2003. *Measures and Applications of Lexical Distributional Similarity*. Ph.D. thesis, University of Sussex.

Tingting Wei, Yonghe Lu, Huiyou Chang, Qiang Zhou, and Xianyu Bao. 2015. A semantic approach for text clustering using WordNet and lexical chains. *Expert Systems with Applications*, 42(4):2264 – 2275.

Aaron Steven White, Rachel Dudley, Valentine Hacquard, and Jeffrey Lidz. 2014. Discovering classes of attitude verbs using subcategorization frame distributions. In *Proceedings of NELS*, volume 43, pages 249–260.

Benjamin Lee Whorf. 1956. *Language, Thought, and Reality*. MIT press.

Janyce Wiebe, Tom O'Hara, and Rebecca Bruce. 1998. Constructing Bayesian networks from WordNet for word-sense disambiguation: Representational and processing issues. In *US Army Conference on Applied Statistics*, page 67.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. From paraphrase database to compositional paraphrase model and back. *Transactions of the ACL*, 3:345–358.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2016. CHARAGRAM: Embedding words and sentences via character n-grams. In *Proceedings of EMNLP*, pages 1504–1515.

Susan Windisch Brown, Dmitriy Dligach, and Martha Palmer. 2011. VerbNet class assignment as a WSD task. In *Proceedings of IWCS*, pages 85–94.

Ludwig Wittgenstein. 1953. *Philosophical Investigations*. Basil Blackwell.

Maryanne Wolf and Catherine J. Stoodley. 2008. *Proust and the Squid: The Story and Science of the Reading Brain*. Harper Perennial New York.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv*, abs/1910.03771.

Alison Wray. 2015. Why are we so sure we know what a word is? In John R. Taylor, editor, *The Oxford Handbook of the Word*, pages 725–750. Oxford University Press.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of EMNLP-IJCNLP*, pages 833–844.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Yunfang Wu and Wei Li. 2016. Overview of the NLPCC-ICCPOL 2016 shared task: Chinese word similarity measurement. In *Natural Language Understanding and Intelligent Applications*, pages 828–839. Springer.

Wenhan Xiong, Jingfei Du, William Yang Wang, and Veselin Stoyanov. 2019. Pretrained encyclopedia: Weakly supervised knowledge-pretrained language model. *arXiv preprint arXiv:1912.09637*.

XTAG Research Group. 2001. A lexicalized tree adjoining grammar for English. Technical report, IRCS, University of Pennsylvania.

Chang Xu, Yalong Bai, Jiang Bian, Bin Gao, Gang Wang, Xiaoguang Liu, and Tie-Yan Liu. 2014. RC-NET: A general framework for incorporating knowledge into word representations. In *Proceedings of ACM*, pages 1219–1228.

Haoran Yan, Xiaolong Jin, Xiangbin Meng, Jiafeng Guo, and Xueqi Cheng. 2019. Event detection with multi-order graph convolution and aggregated attention. In *Proceedings of EMNLP–IJCNLP*, pages 5766–5770.

Dongqiang Yang and David M. W. Powers. 2006. Verb similarity on the taxonomy of WordNet. In *Proceedings of the 3rd International WordNet Conference (GWC-06)*, pages 121–128.

Sen Yang, Dawei Feng, Linbo Qiao, Zhigang Kan, and Dongsheng Li. 2019a. Exploring pre-trained language models for event extraction and generation. In *Proceedings of ACL*, pages 5284–5294.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019b. XLNet: Generalized autoregressive pretraining for language understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Proceedings of NeurIPS*, pages 5753–5763.

Mo Yu and Mark Dredze. 2014. Improving lexical embeddings with semantic knowledge. In *Proceedings of ACL*, pages 545–550.

Stella X. Yu and Jianbo Shi. 2003. Multiclass spectral clustering. In *Proceedings of ICCV*, page 313.

Michelle Yuan, Mozhi Zhang, Benjamin Van Durme, Leah Findlater, and Jordan Boyd-Graber. 2019. Interactive refinement of cross-lingual word embeddings. In *Proceedings of EMNLP*, pages 5984–5996.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of ACL*, pages 4791–4800.

Torsten Zesch and Iryna Gurevych. 2006. Automatically creating datasets for measures of semantic relatedness. In *Proceedings of the Workshop on Linguistic Distances*, pages 16–24.

Jian-Xin Zhang, Xue-Dong Du, and Bin-Guo Wang. 2019a. Semantic representation based on clustering and attention mechanism to identify deceptive comment models. *Journal of Computers*, 30(4):130–139.

Mozhi Zhang, Yoshinari Fujinuma, Michael J. Paul, and Jordan Boyd-Graber. 2020. Why overfitting isn't always bad: Retrofitting cross-lingual word embeddings to dictionaries. In *Proceedings of ACL*, pages 2214–2220.

Tongtao Zhang and Heng Ji. 2018. Event extraction with generative adversarial imitation learning. *arXiv preprint arXiv:1804.07881*.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019b. ERNIE: Enhanced language representation with informative entities. In *Proceedings of ACL*, pages 1441–1451.

Fanghua Zheng. 2018. A corpus-based multidimensional analysis of linguistic features of truth and deception. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, pages 841–848.

Junru Zhou and Hai Zhao. 2019. Head-Driven Phrase Structure Grammar parsing on Penn Treebank. In *Proceedings of ACL*, pages 2396–2408.

Liming Zhou, Philip K. Hopke, and Prasanna Venkatachari. 2006. Cluster analysis of single particle mass spectra measured at Flushing, NY. *Analytica chimica acta*, 555(1):47–56.

Zi Huang, Kam-Fai Wong, Wenjie Li, D. Song, and P. Bruza. 2003. Back to the future: A logical framework for temporal information representation and inferencing from financial news. In *Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering*, pages 95–101.

Arnold M. Zwicky. 1971. In a manner of speaking. *Linguistic Inquiry*, 2(2):223–233.

Appendix A

Clustering Algorithms

Unsupervised cluster analysis is a widely used exploratory method for identifying patterns and structure in unlabelled data. The goal of the clustering task is to group a set of data items (i.e., observations) so that items belonging to the same cluster are more similar (based on some criterion) to each other than to items in the other clusters. Since the choice of the clustering algorithm and optimal parameter settings (e.g., the number of expected clusters, distance function) depends on the properties of the analysed data and the end goal, cluster analysis is usually an iterative process where multiple algorithm and parameter configurations are tested before attaining the desired output. Provided below is a brief overview of the clustering algorithms used in this thesis.

k-means clustering

One of the classic clustering algorithms is k-means (Lloyd, 1982; MacQueen, 1967), a Euclidean distance-based method which splits the data into k flat clusters, using partitioning representatives which correspond to the mean of each cluster (i.e., not actual points drawn from the data). The goal is to separate a set of n data points $(x_1, x_2, ..., x_n)$ into k groupings ($\mathbf{C} = \{C_1, C_2, ..., C_k\}$) of equal variance, minimising the *inertia*, i.e., the within-cluster sum-of-squares, which serves as a measure of internal coherence of clusters and is computed as:

$$\sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||^2 \tag{A.1}$$

where μ_j is the mean of the points in the cluster C_i (i.e., its centroid). The algorithm performs clustering by first choosing the initial centroids, and then looping between

two steps: (i) assigning each data point to its nearest centroid, and (ii) creating new centroids based on the mean value of all the data points assigned previously to each centroid, until the difference between the old and new centroid is below some threshold value (i.e., until the centroids stabilise).

Agglomerative clustering

Agglomerative clustering techniques are hierarchical methods which organise the data points into a dendrogram from the ground up, starting from individual data points and recursively merging them into clusters. The merging strategy depends on the chosen linkage method: the algorithm successively joins pairs of clusters that minimally increase a given linkage distance. Amongst the popular linkage criteria are: (a) single-linkage, which minimises the distance between the two closest data points in a pair of clusters, (b) average-linkage, where the average of the distance between all data points in pairs of clusters is minimised and (c) complete linkage, which minimises the maximum distance between data points in cluster pairs. Further, Ward's linkage method (d) minimises the sum of squared differences in all clusters, by finding a pair of clusters at each step which brings about the minimum increase in total within-cluster variance after merging – a criterion similar to the variance-minimising objective function used in k-means, albeit differently tackled. The result is a nested cluster structure, and partitionings of different granularity can be derived by cutting the dendrogram at different levels.

Spectral clustering

In spectral methods, the clustering process involves two phases, dimensionality reduction and clustering in the low dimensional space. The symmetric affinity matrix W_{ij} representing pairwise similarities between all data points is first embedded into a Euclidean space, and then the clustering is performed on the components of the eigenvectors of the matrix. Spectral methods view the similarity matrix as an adjacency matrix of an undirected graph G over the set of n data points, corresponding to the nodes in the graph and connected by weighted edges, with weights representing the pairwise similarity between a pair of data points. Viewed from this perspective, the clustering task is reduced to the problem of finding optimal cuts in the graph, so that weakly connected areas of nodes (i.e., connected by edges with small weights) represent the partitions of the data (Aggarwal and Reddy, 2014). The goal is therefore to minimise a function of the weights across the partitions, constructed in terms of the adjacency matrix and a degree matrix, where all entries are zero except for the

diagonal values, each equal to the sum of the weights of the incident edges (Sammut and Webb, 2011). The degree of a vertex is therefore: $d_i = \sum_{j=1}^{N} w_{ij}$.

A variation of the method using the MNCut algorithm has been proposed by Meila and Shi (2001), which has been shown to produce strong results in the task of word-level clustering (Scarton et al., 2014; Sun and Korhonen, 2009; Sun et al., 2010, 2013; Vulić et al., 2017b). The MNCut algorithm transforms the affinity matrix W into a stochastic matrix P:

$$P = D^{-1}W \tag{A.2}$$

where D is the degree matrix. Following Sun and Korhonen (2009), let $V = \{v_n\}_{n=1}^{N}$ be the set of data points, and $I = \{I_k\}_{k=1}^{K}$ be a disjoint partition of K classes into which the data points are to be clustered. Meila and Shi (2001) showed that if matrix P has the K leading eigenvectors that are piecewise constant[1] with respect to a partition I^* and their eigenvalues are not zero, then I^* minimises the multiway normalised cut (MNCut), i.e., the sum of transition probabilities across different clusters. This criterion thus finds the data partition where the random walks are most likely to happen within the same cluster (Sun and Korhonen, 2009). Considering that, in practice, the leading eigenvectors of the matrix P are not piecewise constant, the partition can be obtained by employing a clustering algorithm such as k-means to find the approximately equal elements in the eigenvectors.

DBSCAN

The Density-Based Spatial Clustering of Applications with Noise algorithm (DBSCAN) is a non-parametric clustering method first introduced by Ester et al. (1996), which detects clusters by identifying high concentrations of data points in the data space. The popularity of the algorithm stems from its ability to discover clusters of different sizes and shapes, with little to no information as to the structure and patterns in the data (Tran et al., 2006): it makes no assumptions upfront about the number of clusters in the dataset, nor the variance within the groupings. Originally designed to handle spatial data, it has since been applied to various data types across different research areas (De Oliveira et al., 2011; Plant et al., 2010; Zhang et al., 2019a; Zhou et al., 2006, *inter alia*).

The core notion on which the DBSCAN cluster model is based is that of *core points*, i.e., the points within high-density areas, defined in terms of two parameters: the *Eps-neighbourhood* (ϵ) of a point and the minimum number of points (*MinPts*) in an

[1] The eigenvector v is piecewise constant with respect to I if $v(i) = v(j) \forall i, j \in I_k$ and $k \in 1, 2...K$.

Eps-neighbourhood of that point. *Core points* are data items with more neighbours than the *MinPts* threshold within the ϵ radius. Points on the edge of a cluster (non-core points) are referred to as *border points*. The algorithm aims to find high-density areas, separated by lower density areas, such that for every point p in a cluster C there is a point q in C so that p is inside of the *Eps-neighbourhood* ϵ of q and ϵ contains at least *MinPts* points. In other words, all points found within radius ϵ of a *core point* are considered members of the same cluster as the *core point* (i.e., they are *direct density reachable*). Further, if any of these points are *core points* themselves, their neighbourhoods are transitively included (*direct density reachable*) (Schubert et al., 2017). Data points that are not *density reachable* from any core point (directly or transitively) do not belong to any cluster and are considered as noise. The DBSCAN algorithm linearly scans the dataset for unprocessed data points, assigning non-core points to *noise* and, whenever a core point is detected, iteratively expanding the point's neighbours and adding them to the cluster.

Appendix B

Verb Class Induction Through Bottom-up Semantic Clustering

B.1 Classification Guidelines

The annotators were presented with the following classification guidelines, along with the list of 267 verbs in their native language, at the start of the task:

Here is a long list of verbs, with one verb in each line.

Please put them together in groups where you feel they are used to express similar or related meanings. For example you may feel 'throw, kick, punch' are related, or 'speak, talk, write'. These groups can be broader (more members) or narrower (fewer members) but any group must have at least 3-5 members. Aim for cohesive small groups if possible and you can add a 'relationship link' from each group to any other groups if you feel there are relationships between the two groups. The relationship could be similar-to (bidirectional) or broader-than (unidirectional). Any verbs you cannot find a good place for, please put in a 'Miscellaneous' group. There is no problem with putting a verb in more than one class if it fits all, for example because a verb may have several different meanings.

We suggest using Microsoft Excel or a related spreadsheet program (e.g. Google Sheets) to constantly have an overview of current groups. The expected output is: (i) groups of verbs according to your own criteria (see above), (ii) relationship links between groups as also discussed above. To facilitate the linking, you can provide simple labels for each group, e.g., Group 1, Group 2.

There is not necessarily a fully correct solution to this task and a perfect grouping. It is perfectly reasonable to use your intuition or gut feeling as a native speaker while working on this task.

Appendix C

Semantic Dataset Construction from Clustering and Spatial Arrangement

C.1 Semantic Clustering and Spatial Arrangement Task Guidelines

C.1.1 Guidelines for Phase 1: Semantic Clustering

The purpose of the task is to group verbs according to their meaning, putting similar or related verbs in the same groups. The aim is to create broad classes of verbs expressing similar and related meanings.

*Look for similar meanings of verbs, and ignore similarity of sound or letters making up the word.

There are 825 verbs[1] in the entire sample, which you will need to divide into smaller groups by placing them in circles displayed on the screen. As a rule of thumb, aim for broad classes of roughly 30-50 verbs. In some cases, groups may be even more numerous – but try not to go beyond 80. Rarely, you may end up identifying a very concrete, narrow class of verbs (i.e., related to a specific meaning domain), which you will want to keep separate, despite it being smaller than the recommended class size – it is permissible to do so, if the verbs do not seem to fit with any other class.

[1]Provided is the English sample size. Note that the exact number of verbs in the sample varied from language to language (see Table 5.1).

*There is not necessarily a fully correct solution to this task and a perfect grouping. It is perfectly reasonable to use your intuition or gut feeling as a native speaker while working on this task.

The Structure of the Task

Task 1: Qualification
The aim of this task is to make sure that the participant carrying out the task understood the instructions and to give them a chance to familiarise themselves with the interface. It should only take a few minutes to complete. You will be asked to group 7 verbs according to their similarity, putting similar verbs in the same circles.

Task 2: Verb clustering
This is the actual verb clustering experiment. You will be automatically directed it to it once you've successfully completed the qualification task. In this task, you will need to group 825 verbs (listed alphabetically in a queue), putting similar and related verbs in the same circles.

The Interface

Task 1 and Task 2 have exactly the same setup. Upon starting the task, you will see three white circles and a list of verbs displayed in a queue at the bottom. You can scroll through the list clicking on the arrows (») on the right. The aim is to group the verbs displayed at the bottom according to their meaning, placing similar or related verbs in the same circles.

The three circles you see upon starting the task are: 'trash', 'copy' and 'new category'. In order to place a verb in a circle, click on one of them, drag it over a circle and drop it there. For example, let's say you spot the verb 'boil' in the list. In order to place it in a group, drag it onto the 'new category' circle. When you drop it there, you will be prompted to name the category. The easiest thing to do is to number them, as the theme of the category may change as you encounter new verbs. The category name is just for yourself, and the only thing that matters to us is to see which verbs are grouped together (regardless of the category 'theme'). You may want to keep track of the category names/numbers in a text file or on a piece of paper, so that it's easier to remember what kind of general theme you wanted that category to have (and modify it as you go along and add new verbs) – but this is up to you.

Next, let's say you spot the verb 'simmer' later on. To add it to the group containing 'boil', simply drag it and drop it onto that circle.

It is possible that in some cases you will want to place one verb in more than one category. Let's say you spot the verb 'draw' and you decide that you want to place it in an 'art'-related circle and in a 'pulling'-verbs circle. In order to duplicate the verb 'draw', drag it onto the 'copy' circle and drop it there. This way, a copy of 'draw' will be added to the verb queue at the bottom and you can drag both onto the circles of your choice. The copying can be repeated as many times as you need. If you copy a verb too many times by accident and want to discard one of the copies, simply drag it onto the 'trash' circle.

The cross visible next to the edge of a category circle allows you to delete the category completely. When a category gets deleted, all the items from that circle move back to the queue at the bottom, and can be grouped again. When you click on the cross, before the category gets deleted, you will be asked to confirm your choice ('Are you sure you want to remove 'X'?). Be careful not to delete the Trash category – however, if you do remove it by mistake, you can create it again, placing the item you want to get rid of onto a 'new category' circle and naming it 'trash'. Keep in mind that you will need to bin all the items you previously placed in 'trash' again (as they will have moved to the queue at the bottom).

Below the queue of verbs, in the bottom bar you will see how many verbs you have already placed (0/825). You will also see the 'Save progress', 'Help', and 'Full screen' button. It is recommended to do the task in the Full screen mode, to improve visibility.

Time

Task 1 should only take a few minutes to complete. Task 2 can take 2-3 hours. You will be able to save your progress at any point and are encouraged to complete it over several sessions, to reduce the fatigue. We recommend, however, to start the task with enough time freed up over a course of several days to complete it at fairly short time intervals, over the course of no more than a week. The shorter the breaks between sessions, the easier the task, as it's easier to remember the verbs already seen and the reasoning behind the created categories. (Taking note of the general theme of each category in a separate file/on a piece of paper may help, as suggested above).

Saving Progress

To save progress, click on the green 'Save progress' button in the bottom bar.

Feedback

After completing the task, you will be asked to provide feedback on the task. Apart from general comments, you will be asked to list any classes of verbs that you feel are related. When classifying verbs, you may feel like two groups of verbs which you separate into two classes are in fact related to one another. Please take a note of these, noting down their name/number, e.g., Class 1 — Class 6; feel free to comment on how these classes relate to one another (e.g., perhaps one is broader the other narrower, or the verbs in one describe specific types of actions, and the other contains more general terms).

C.1.2 Guidelines for Phase 2: Spatial Similarity Judgments

In this task, we ask you to arrange verbs in a circular arena according to the similarity of their meaning, putting similar verbs closer together and those expressing less similar concepts further apart. The relative positions and distances between the words in the circle will reflect the degree of similarity between them: the more similar the words are, the closer together they should be placed; the more dissimilar they are, the further apart.

*Please consider only similar meanings of verbs, and ignore similarity of sound or letters making up the word.

*There is not necessarily a fully correct solution to this task and a perfect grouping. It is perfectly reasonable to use your intuition or gut feeling as a native speaker while working on this task.

The Structure of the Task

Task 1: Qualification
Here we want to see if you understood the instructions and give you a chance to familiarise yourself with the interface. It should take a few minutes to complete. You will be asked to arrange 7 verbs according to their similarity, putting similar verbs closer together and less similar verbs further apart. Based on your arrangement of verbs we may decide not to let you start the full experiment. It is a way for us to make sure only native speakers who fully understood the task will participate, not to waste anyone's time.

Task 2: Verb Similarity Multi-arrangement Task

This is the actual verb arrangement experiment. You will be directed to it once you have successfully completed the qualification experiment. This phase consists of several of the same type of tasks, differing only in the set of words to be arranged. Each task is separate and in each you will work on a single set of words. Once you've finished arranging them in the circular arena, the task will be completed and you will be able to start the next one – again, working on a different set of words from scratch (although there may be some words you have seen in previous tasks).

The Interface

The qualification and the actual experiment have exactly the same setup. When starting the task, you will see a white circle in the middle of the screen and a set of words (roughly from 25 to 100, depending on the task) displayed around the circle. You will need to move each verb onto the circle and arrange them so that similar verbs are close together and dissimilar verbs are further apart. Please make sure you use the entire arena and all of the space available.

In order to move a verb, click on it and drag it onto the circular arena. You will notice that when you hover the cursor over a word it will be highlighted in bold and black. Once you've dropped a word onto the arena, you will be able to click on it again and move it again as many times as you like. Feel free to keep rearranging the words until you are satisfied with the arrangement. If you want to move several verbs at the same time, drag a box around the words you want to move and then click on one of them to get hold of the selected items – and drag them to the intended location.

At the bottom of the screen, below the circular arena, there is a 'Place all items' button, which will change into a green 'Finish' button once you have arranged all words within the circle. You can still adjust the arrangement at that point, until you are happy with it. If you don't want to make any more changes, click on the 'Finish' button. This will take you to another circular arena, this time surrounded by a subset of words from the first arrangement. This is because the program 'zooms in' on the verbs which you put closer together in the first trial, to give you more space to arrange them again, in a less crowded space. Again, drag the verbs onto the circle, putting similar verbs closer together and those less similar further apart. Please use the entire arena on each trial. This may mean that you will need to spread out the same words more in consecutive trials, as there will be fewer of them and you will have more space. Only the relations between distances on a single trial (i.e., not the absolute on-screen

distances) are meaningful, so approach each trial individually. Once you're done with this arena, click on the 'Finish' button and move onto the next trial, and so on.

These repeated arrangements allow the program to collect similarity information about all the different words you work on, so do not get discouraged if you have to arrange the same group of words over and over again. Once enough similarity information is collected, the task will end and you will be done with the given set of words. When you feel ready to tackle a new set of words, simply start the next task.

If you ever need to consult these instructions during the task, click on the 'Help' button at the bottom of the screen. The 'Full screen' button will allow you to work in full-screen mode, which is highly recommended to ensure good visibility.

Time

The duration of the experiment will depend on how many tasks (i.e., how many different sets of words) you will want to complete. Each individual task can take roughly from 30 to 60 minutes. You are encouraged to take as many breaks between tasks as you wish, but it is best to work on a single task and wordset within one session, as this will keep the words fresh in your memory and make the task easier. We recommend to start work on the experiment with enough time freed up over a course of several days to complete it in the course of no more than a week.

Saving Progress

Your progress will save each time you click on the 'Finish' button. Therefore, it's recommended to take breaks right after clicking 'Finish', so that no work is lost if you accidentally close the browser. If you close the window after clicking 'Finish', you will be able to go back to it later by reopening it and start where you left off.

Feedback

After completing the task, you will be asked to provide feedback on the task. Please share with us your thoughts about the difficulty of the task and your experience using the interface.

C.2 Phase 1: Cluster Distributions

Figure C.1 shows cluster size distributions across the 10 annotators who completed the rough clustering task (Phase 1).

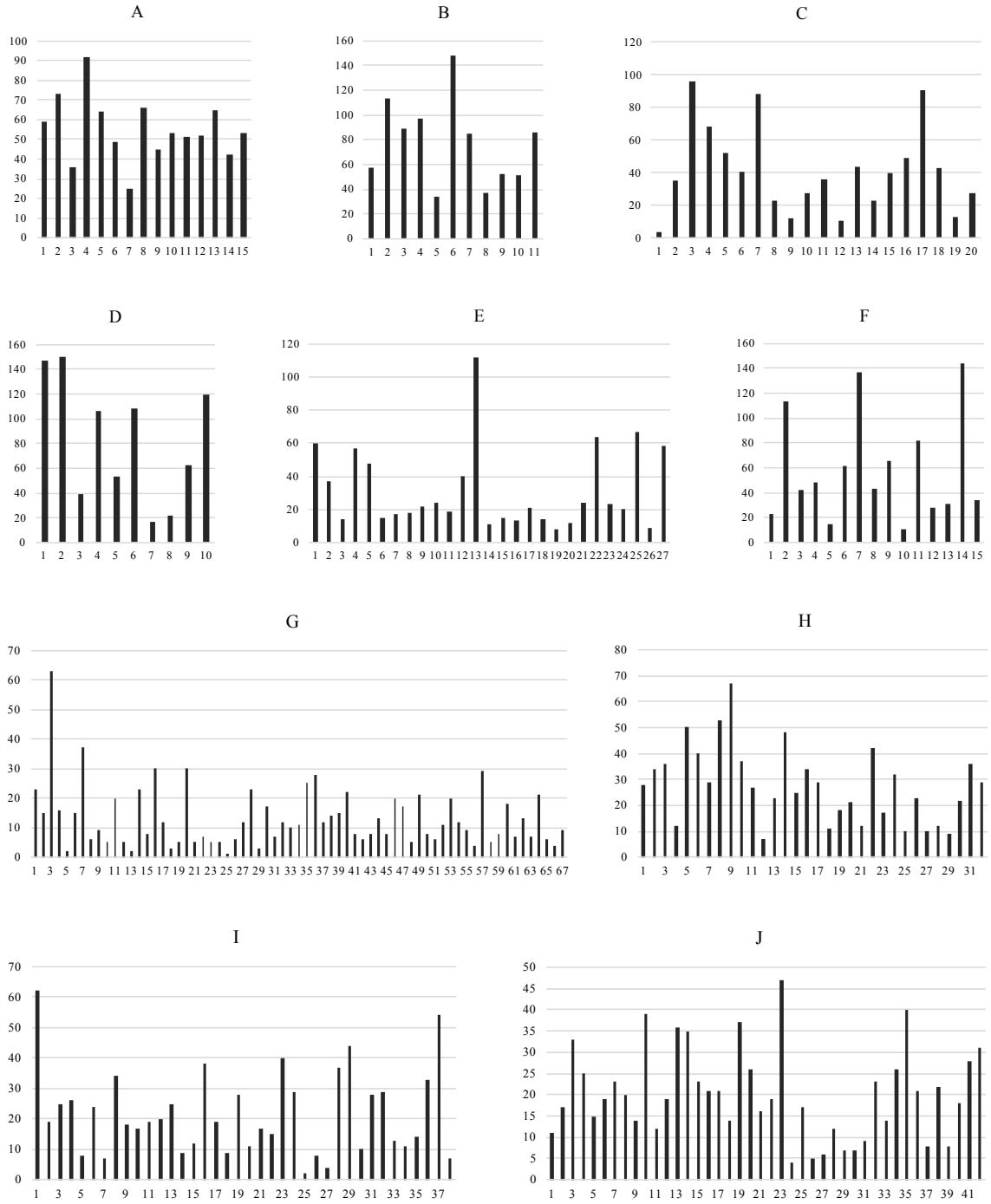


Fig. C.1 Phase 1 cluster size distributions across 10 annotators (A-J).

C.3 Phase 2: Comparison of Individual Arrangements

Figures C.2 and C.3 compare the output of PCoA on pairs of representational dissimilarity matrices for Classes 11 and 3, respectively, for pairs of annotators with the highest (top) and lowest (bottom) average pairwise agreement with the mean of all annotators.

C.4 PCoA on Class 10

Figure C.4 includes additional views of the 3-dimensional projection of the distances captured by the representational dissimilarity matrix for Class 10 (verbs of motion) shown in Figure 4.11.

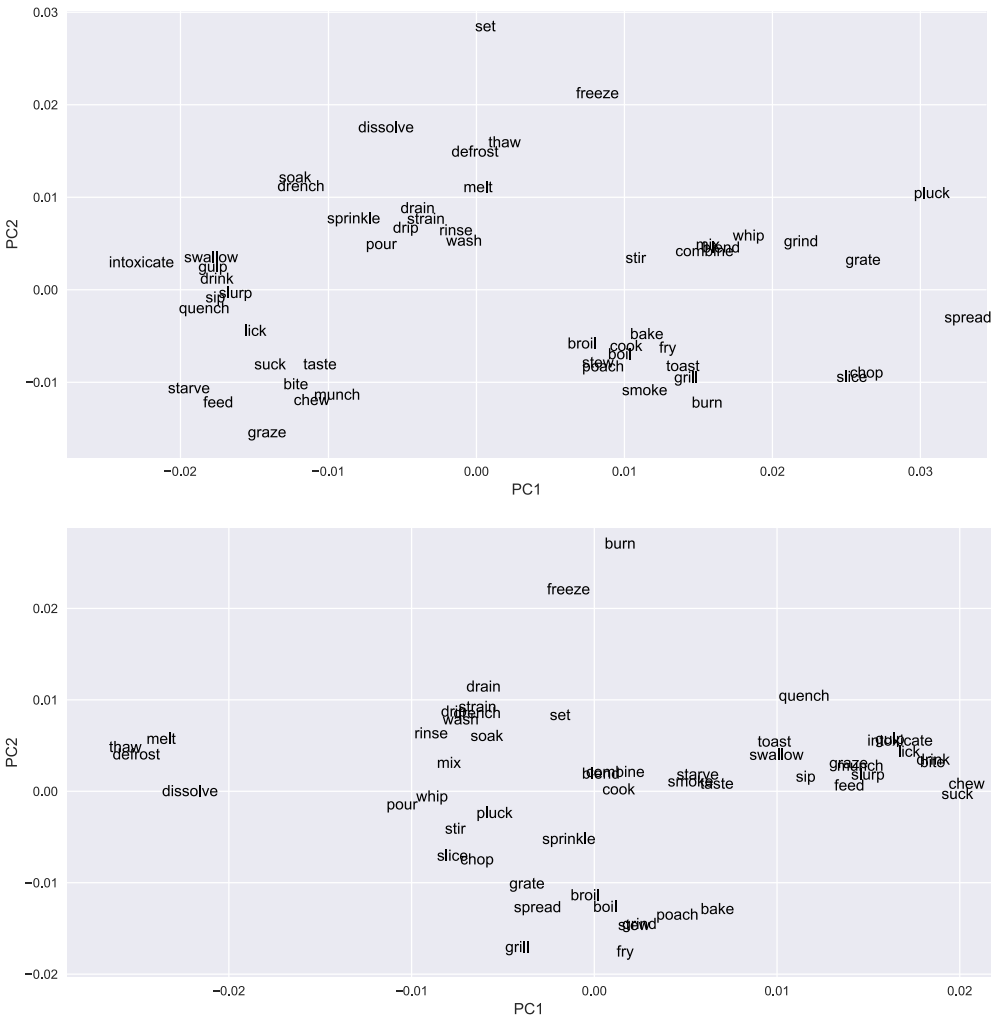


Fig. C.2 Visualisation of PCoA applied to Phase 2 distance matrices for Class 11 verbs from two annotators with the highest (top) and lowest (bottom) average pairwise agreement with the mean of all annotators.

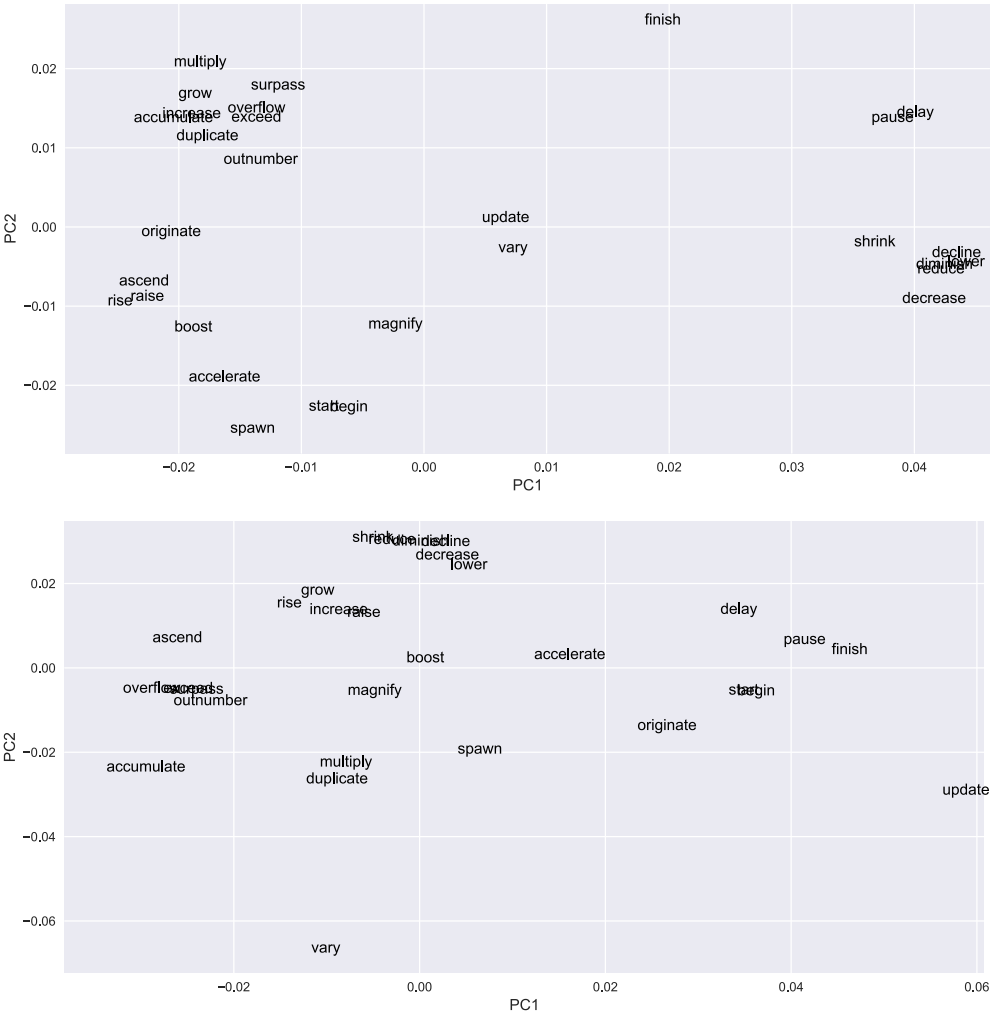


Fig. C.3 Visualisation of PCoA applied to Phase 2 distance matrices for Class 3 verbs from two annotators with the highest (top) and lowest (bottom) average pairwise agreement with the mean of all annotators, and the lowest pairwise correlation between themselves.

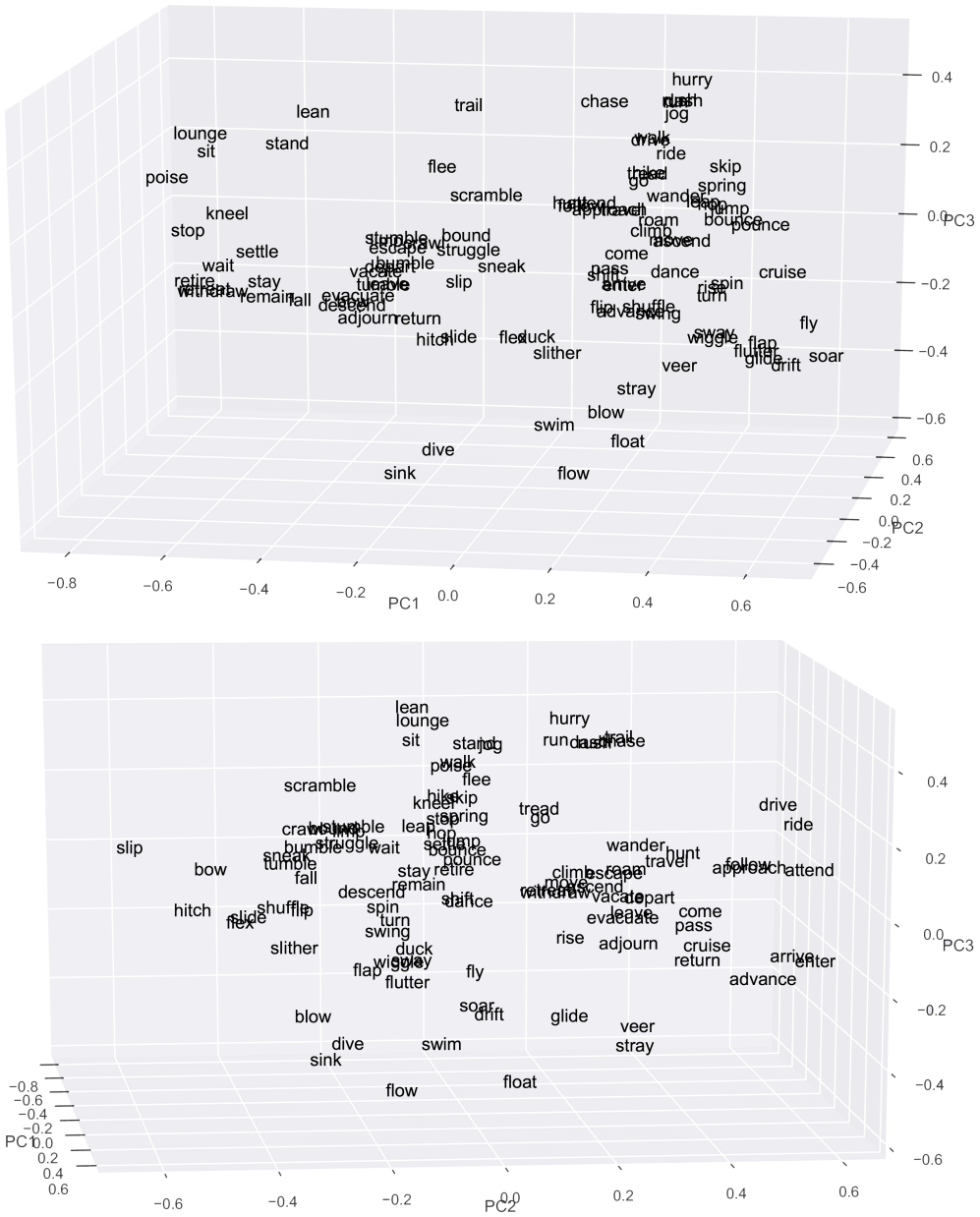


Fig. C.4 Visualisation of PCoA applied to the Phase 2 distance matrix for verbs of motion (Class 10), rotated to highlight individual dimensions.

C.5 Supplementary Results

Table C.1 presents additional results of the evaluation of BERT embeddings computed *in isolation* using different lexical representation extraction configurations. To identify the strongest performing configuration (the results reported in Section 4.8) I fine-tuned the following parameters: (1) the choice of hidden layers to average over (L_0 refers to the input embedding layer, $\leq L_n$ refers to the Transformer layers included in layer-wise averaging, inclusive of L_n), (2) the choice of special tokens (`[CLS]` and `[SEP]`) included in the subword averaging step. I report the subset of results on SpA-Verb and SimVerb-3500 data (including the subset of shared pairs and the thresholded set of SpA-Verb) to illustrate the impact of the different configurations of lexical representation extraction parameters on performance on both datasets.

Configuration		SV-3500	SV∩SpA$_{SVs}$	SV∩SpA$_{SpAs}$	SpA-Verb	SpA-Verb-THR
L	**SPEC**					
BERT-BASE						
L_0	+	0.219	0.142	0.096	0.167	0.195
	−	0.318	0.214	0.169	0.212	0.220
$\leq L_4$	+	0.222	0.139	0.107	0.175	0.204
	−	0.338	0.221	0.195	0.239	0.246
$\leq L_6$	+	0.204	0.130	0.093	0.158	0.187
	−	0.338	0.224	0.207	0.235	0.240
$\leq L_8$	+	0.181	0.107	0.079	0.149	0.177
	−	0.315	0.202	0.198	0.219	0.224
BERT-LARGE						
L_0	+	0.214	0.141	0.087	0.156	0.175
	−	0.314	0.221	0.165	0.214	0.225
$\leq L_4$	+	0.190	0.133	0.053	0.158	0.191
	−	0.331	0.248	0.191	0.229	0.222
$\leq L_6$	+	0.184	0.127	0.038	0.145	0.180
	−	0.319	0.240	0.188	0.224	0.215
$\leq L_8$	+	0.184	0.127	0.037	0.141	0.174
	−	0.319	0.240	0.193	0.221	0.214
BERT-LARGE-WWM						
L_0	+	0.211	0.128	0.090	0.167	0.191
	−	0.347	0.245	0.198	0.230	0.242
$\leq L_4$	+	0.226	0.142	0.110	0.177	0.202
	−	0.390	0.295	0.249	0.248	0.258
$\leq L_6$	+	0.210	0.131	0.104	0.160	0.189
	−	0.396	0.307	0.257	0.237	0.246
$\leq L_8$	+	0.211	0.133	0.107	0.156	0.186
	−	0.395	0.312	0.258	0.224	0.234

Table C.1 Evaluation results across different lexical representation extraction configurations on SimVerb-3500 and SpA-Verb datasets and the subset of shared pairs (cf. Table 4.7). L_0 refers to the input embedding layer; $\leq L_n$ refers to embeddings computed by averaging representations over all Transformer layers up to and inclusive of the nth layer. For each layer averaging configuration I consider two configurations of special tokens (column SPEC): one where special tokens `[CLS]` and `[SEP]` are included (+) and one where they are excluded (−) from the subword embedding averaging step. All scores are Spearman's ρ correlations.

Appendix D

Verb Knowledge Acquisition for Multilingual Evaluation

D.1 Representation Models

I provide URLs to the models used in this study in Table D.1 below. For all languages, I used the pretrained uncased BERT-base models. I also evaluate 300-dimensional `fastText` vectors (Mikolov et al., 2018), trained on Common Crawl and Wikipedia data of each language using an extension of the CBOW `word2vec` model (Mikolov et al., 2013b) with position-weights over 10 training epochs, with character n-grams of length 5, window size of 5, and 10 negative examples.

D.2 External Corpora

The corpora used to extract sentential contexts for the *in context* BERT word embeddings are listed below (Table D.2). I randomly sampled 1 million sentences of maximum sequence length 512 from each monolingual corpus.

D.3 WALS Features

Table D.3 lists the morphological, syntactic and lexical typological features from the World Atlas of Language Structures (WALS) (https://wals.info) used in cross-lingual comparisons in Section 5.4.1 (Figure 5.3), selected based on the availability of corresponding entries for the languages in the sample.

Language	Model	URL
Chinese	BERT	https://huggingface.co/bert-base-chinese
Finnish	BERT	https://huggingface.co/TurkuNLP/bert-base-finnish-uncased-v1
Italian	BERT	https://huggingface.co/dbmdz/bert-base-italian-uncased
	+XXL	https://huggingface.co/dbmdz/bert-base-italian-xxl-uncased
Japanese	BERT	https://huggingface.co/cl-tohoku/bert-base-japanese
	+WWM	https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking
Polish	BERT	https://huggingface.co/dkleczek/bert-base-polish-uncased-v1
Multilingual	BERT	https://huggingface.co/bert-base-multilingual-uncased
all	FT	https://fasttext.cc/docs/en/crawl-vectors.html

Table D.1 Links to the models used in this study. For each language, I used the uncased BERT-base model(s) (including the variant with whole word masking (+WWM) for Japanese and the XXL Italian BERT-base model trained on a larger (81GB) corpus) and 300-dimensional `fastText` (FT) vectors available for that language.

Language	Corpus	URL	Word segmenter
Chinese	United Nations Parallel Corpus	http://opus.nlpl.eu/UNPC.php	https://github.com/fxsjy/jieba
Finnish	Europarl	http://opus.nlpl.eu/Europarl.php	-
Italian	Europarl	http://opus.nlpl.eu/Europarl.php	-
Japanese	Polyglot Wikipedia	https://sites.google.com/site/rmyeid/projects/polyglot?authuser=0	https://github.com/Kensuke-Mitsuzawa/JapaneseTokenizers
Polish	Europarl	http://opus.nlpl.eu/Europarl.php	-

Table D.2 Links to the external corpora used for extraction of N sentences for computing BERT representations *in context* and the word segmenters used, where appropriate.

ID	Feature	ID	Feature
26A	Prefixing vs. Suffixing in Inflectional Morphology	86A	Order of Genitive and Noun
29A	Syncretism in Verbal Person/Number Marking	87A	Order of Adjective and Noun
33A	Coding of Nominal Plurality	88A	Order of Demonstrative and Noun
36A	The Associative Plural	89A	Order of Numeral and Noun
40A	Inclusive/Exclusive Distinction in Verbal Inflection	90A	Order of Relative Clause and Noun
44A	Gender Distinctions in Independent Personal Pronouns	91A	Order of Degree Word and Adjective
45A	Politeness Distinctions in Pronouns	92A	Position of Polar Question Particles
46A	Indefinite Pronouns	95A	Relationship between the Order of Object and Verb and the Order of Adposition and Noun Phrase
47A	Intensifiers and Reflexive Pronouns	96A	Relationship between the Order of Object and Verb and the Order of Relative Clause and Noun
48A	Person Marking on Adpositions	97A	Relationship between the Order of Object and Verb and the Order of Adjective and Noun
49A	Number of Cases	100A	Alignment of Verbal Person Marking
50A	Asymmetrical Case-Marking	101A	Expression of Pronominal Subjects
51A	Position of Case Affixes	102A	Verbal Person Marking
52A	Comitatives and Instrumentals	103A	Third Person Zero of Verbal Person Marking
53A	Ordinal Numerals	104A	Order of Person Markers on the Verb
64A	Nominal and Verbal Conjunction	105A	Ditransitive Constructions: The Verb 'Give'
69A	Position of Tense-Aspect Affixes	107A	Passive Constructions
70A	The Morphological Imperative	112A	Negative Morphemes
71A	The Prohibitive	115A	Negative Indefinite Pronouns and Predicate Negation
72A	Imperative-Hortative Systems	116A	Polar Questions
74A	Situational Possibility	129A	Hand and Arm
75A	Epistemic Possibility	130A	Finger and Hand
76A	Overlap between Situational and Epistemic Modal Marking	138A	Tea
80A	Verbal Number and Suppletion	143A	Order of Negative Morpheme and Verb
81A	Order of Subject, Object and Verb	143E	Preverbal Negative Morphemes
82A	Order of Subject and Verb	143F	Postverbal Negative Morphemes
83A	Order of Object and Verb	143G	Minor morphological means of signaling negation
84A	Order of Object, Oblique, and Verb	144A	Position of Negative Word With Respect to Subject, Object, and Verb
85A	Order of Adposition and Noun Phrase		

Table D.3 WALS typological features considered in cross-lingual comparisons.

Appendix E

Verb Knowledge Injection for Multilingual Event Processing

E.1 Adapter Training: Hyperparameter Search

I experimented with $n \in \{10, 15, 20, 30\}$ training epochs, as well as an early stopping approach using validation loss on a small held-out validation set as the stopping criterion, with a patience argument $p \in \{2, 5\}$; I found the adapters trained for the full 30 epochs to perform most consistently across tasks.

The size of the training batch varies based on the value of k negative examples generated from the starting batch B of positive pairs: e.g., by generating $k = 3$ negative examples for each of 8 positive examples in the starting batch I end up with a training batch of total size $8 + 3 * 8 = 32$. I experimented with starting batches of size $B \in \{8, 16\}$ and found the configuration $k = 3$, $B = 16$ to yield the strongest results (reported in this thesis).

E.2 STM Training Details

The STM is trained using the sets of English positive examples from each lexical resource (Table 6.1). Negative examples are generated using controlled sampling (see §6.2.2), using a $k = 2$ [*cc*] configuration, ensuring that generated negatives do not constitute positive constraints in the global set. I use the pretrained 300-dimensional static distributional word vectors computed on Wikipedia data using the `fastText` model (Bojanowski et al., 2017), cross-lingually aligned using the RCSLS model of Joulin et al. (2018), to obtain the shared cross-lingual embedding space for each source-

			+FN$_{seq}$	+VN$_{seq}$
AR	mBERT-zs	T-id	16.1	15.2
		T-cl	15.1	14.1
		ARG-id	1.2	1.1
		ARG-cl	1.0	1.0
	AR-BERT	T-id	**70.5**	69.1
		T-cl	**65.0**	63.7
		ARG-id	**32.9**	30.2
		ARG-cl	**29.5**	27.6
	AR-mBERT	T-id	64.6	65.5
		T-cl	55.6	**57.1**
		ARG-id	24.6	23.4
		ARG-cl	20.5	19.9
ZH	mBERT-zs	T-id	**41.6**	**39.9**
		T-cl	**29.6**	27.8
		ARG-id	4.6	**7.6**
		ARG-cl	4.0	**6.4**
	ZH-BERT	T-id	75.6	75.7
		T-cl	**69.0**	68.5
		ARG-id	26.8	26.1
		ARG-cl	25.9	25.0
	ZH-mBERT	T-id	72.6	72.6
		T-cl	**64.1**	62.2
		ARG-id	**27.0**	24.9
		ARG-cl	**25.8**	23.9

Table E.1 Results on Arabic and Chinese ACE test sets for sequential fine-tuning setup for zero-shot (zs) transfer with **mBERT** and vtrans transfer approach with language-specific BERT (**AR-BERT / ZH-BERT**) or mBERT, on noisily translated FN/VN data (§6.2.4). F1 scores averaged over 5 runs; significant improvements (paired t-test; $p < 0.05$) over both baselines marked in bold.

target language pairing. The STM is trained using the Adam optimiser (Kingma and Ba, 2015), a learning rate $l = 1e-4$, a batch size of 32 (positive and negative) training examples, for a maximum of 10 iterations. I set the values of remaining training hyperparameters as in Ponti et al. (2019), i.e., the number of specialisation tensor slices $K = 5$ and hidden size of the specialised vectors $h = 300$.

E.3 Additional Results

Table E.1 includes the results for the sequential fine-tuning setup for Task 2 (ACE) in Arabic and Chinese.