*Technical Report*

UCAM-CL-TR-970
ISSN 1476-2986

Number 970

UNIVERSITY OF
CAMBRIDGE

Computer Laboratory

# Transparent analysis of multi-modal embeddings

Anita L. Verő

May 2022

15 JJ Thomson Avenue
Cambridge CB3 0FD
United Kingdom
phone +44 1223 763500

*https://www.cl.cam.ac.uk/*

This technical report is based on a dissertation submitted November 2021 by the author for the degree of Doctor of Philosophy to the University of Cambridge, King’s College.

Some figures in this document are best viewed in colour. If you received a black-and-white copy, please consult the online version if necessary.

Technical reports published by the University of Cambridge Computer Laboratory are freely available via the Internet:

*https://www.cl.cam.ac.uk/techreports/*

ISSN 1476-2986

## Abstract

Vector Space Models of Distributional Semantics – or Embeddings – serve as useful statistical models of word meanings, which can be applied as proxies to learn about human concepts. One of their main benefits is that not only textual, but a wide range of data types can be mapped to a space, where they are comparable or can be fused together.

Multi-modal semantics aims to enhance Embeddings with perceptual input, based on the assumption that the representation of meaning in humans is grounded in sensory experience. Most multi-modal research focuses on downstream tasks, involving direct visual input, such as Visual Question Answering. Fewer papers have exploited visual information for meaning representations when the evaluation tasks involve *no direct visual input*, such as semantic similarity. When such research has been undertaken, the results on the impact of visual information have been often inconsistent, due to the lack of comparison and the ambiguity of intrinsic evaluation.

Does visual data bolster performance on non-visual tasks? If it does, is this only because we add more data or does it convey complementary *quality* information compared to a higher *quantity* of text? Can we achieve comparable performance using small-data if it comes from the right data distribution? Is the modality, the size or the distributional properties of the data that matters? Evaluating on downstream or similarity-type tasks is a good start to compare models and data sources. However, if we want to resolve the ambiguity of intrinsic evaluations and the spurious correlations of downstream results, creating more *transparent* and human *interpretable* models is necessary.

This thesis proposes diverse studies to scrutinize the inner "cognitive models" of Embeddings, trained on various data sources and modalities. Our contribution is threefold. Firstly, we present comprehensive analyses of how various visual and linguistic models behave in semantic similarity and brain imaging evaluation tasks. We analyse the effect of various image sources on the performance of semantic models, as well as the impact of the quantity of images in visual and multi-modal models. Secondly, we introduce a new type of modality: a visually structured, text based semantic representation, lying in-between visual and linguistic modalities. We show that this type of embedding can serve as an efficient modality when combined with low resource text data. Thirdly, we propose and present proof-of-concept studies of a transparent, interpretable semantic space analysis framework.

# Acknowledgements

I am especially thankful to my supervisors, Stephen Clark and Ann Copestake, who guided me on my path to the PhD at different stages and in different ways. I am immensely grateful to Steve for the opportunity of starting a PhD at Cambridge. I learned a lot from our discussions and enjoyed his openness to any out-of-the-box ideas. Ann helped me greatly with organising my work after a break I had to take in the middle of the programme. She helped me clarifying my thoughts with her insightful questions and motivated me to start planning and writing down ideas early. I feel, I greatly benefited from their very different but equally supportive mentoring styles.

I owe special thanks to my collaborators Douwe Kiela, Luana Bulat, Ekaterina Shutova and Christopher Davis, whose intellect and creativity I was lucky to experience first hand.

I feel lucky to have a very supportive family, which helped me through difficult times during the course of this programme. My dad has always showed a great interest in whatever I was doing and often had insightful comments and questions about it too. My mom is always there for me when I have difficulties, which means the world.

My dear friend, Krisztián Gergely, provided invaluable support during the past years for which I will always be grateful for. I would like to thank my good friend, Jonathan Kanen, for his friendship and occasional English corrections. In the last few years I was lucky enough to enjoy the immeasurable support of József Konczer, who not only helped me with finding strength but has always been ready to discuss details of my work as well.

The past years would have been much less bearable without the deep conversations with my dear old friends Klára Békés and Fruzsina Balogh, and my close friends from Cambridge, Akemi Herraez Vossbrink, Paula Fayos Pérez, Eugenia Biral and Kaho Sato.

Finally, I would like to thank all my colleagues in the NLIP group and the visiting guests I had a chance to meet, with whom we had many enlightening and fun conversations in and outside the office.

# Contents

# Chapter 1

# Introduction

The anatomy of human language has long intrigued researchers. In the late twentieth century, Information Technology introduced new, ever improving computational tools which opened a wide range of opportunities to perform empirical investigations on the written and spoken (recorded) realisations of language. This technology gave birth to new fields such as Computational Linguistics and Natural Language Processing (NLP). Data driven analysis of language provided another boost to NLP after the deep learning revolution (or renaissance) in the first half of the 2010s.

The motivations for creating computational models for language are, however, very much varied across communities. Probably, the most dominant branch of research is driven by more – what we may call – *engineering* incentives, and stands by the mission of creating human level language understanding and generating systems. This area has become even more prominent since Machine Learning (ML) – and NLP in particular – has weaved itself into a rapidly developing commercial market. ML and NLP have become ubiquitous in our everyday lives in domains ranging from criminal justice and public policy to healthcare and education [Kaur et al., 2020].

The other – less prominent – direction concerns itself with employing technological tools in order to empirically test research hypotheses about language and cognition or social phenomena. Here, computational models are rather the means than an end, which can generate more knowledge using large scale statistical analysis. This area involves subfields which can be labelled as *Computational Linguistics* or *Computational Sociology.*

The two approaches can differ on the level of applied models as well, which are partially derived from the purpose of investigation. Applied NLP involves more end-to-end models trained for tasks which are close to end-user applications, such as Question Answering, or dialogue systems. More theoretic work often focus on models which are more interpretable and evaluations which are more intrinsic, such as semantic similarity or predicting concept representations in the brain. Machine Learning practitioners cannot debug their models if they do not understand their behaviour [Kaur et al., 2020]. Thus, this type of analytic research can also serve as an important component of a checks as balances system of commercial NLP.

The topic of this thesis is related to the aims of the latter area. We concentrate on *word* semantic models. Even though words primarily acquire their meaning within context and use, thinking in concepts and categories is a basic human strategy by which to operate [Bowker and Star, 2000]. Semantic models of words – and vector space models in particular – provide a compelling instrument for statistical analysis of concepts, realised in language. Therefore, investigations on lexical semantics can be useful for other interdisciplinary research, such as Computational Sociology.

Here, we are concerned with analysing the behaviour as well as the internal "cognitive model" of semantic representations with a focus on multi-modal input. Symbol grounding [Harnad, 1990] or the hypothesis that human semantic representation depends on sensorimotor experience, has been given much attention in the past decades. Dual coding theory [Bucci, 1985], the idea in cognitive science that meaning might be represented in the human brain in multiple modalities has inspired much research in NLP and Computational Linguistics.

Most multi-modal research focus on engineering type of evaluation tasks (and therefore models which perform well on them) which involve direct visual input, such as Visual Question Answering (VQA) [Antol et al., 2015, Srivastava and Salakhutdinov, 2012, Kiros et al., 2014, Socher et al., 2014, Tsai et al., 2019, Lu et al., 2019, Su et al., 2019, Majumdar et al., 2020]. They are usually referential type tasks, in which case the usefulness of visual input is not surprising. Moreover, evaluating solely on downstream tasks is prone to exhibit spurious correlations.

Unlike most studies, this work investigates visual information's contribution to semantic meaning representations when the evaluation tasks involve no direct visual input. Instead of evaluating on referential type tasks like VQA, we are interested in the impact of visual information in higher level *word* and *concept* representations. A minority of papers have exploited visual information for meaning representations when the evaluation tasks involve no direct visual input, such as semantic similarity [Bruni et al., 2014, Kiela and Bottou, 2014, Kiela et al., 2016, Lazaridou et al., 2015, Davis et al., 2019, Lin and Parikh, 2015, Vendrov et al., 2015].

There are three main issues in the literature, which we are addressing in this thesis.

**Problems of Intrinsic Analyses** As a start, we focus on two types of intrinsic evaluation: human judgement based semantic tasks and brain activity prediction. The type of evaluation the community uses has an effect on the model selection process, hence the questions we ask will influence the future direction of model development as well. Working on intrinsic evaluations, such as semantic similarity can positively contribute to both basic research questions about linguistic phenomena as well as developing higher quality end-user applications, by recognising potential pitfalls. However, due to the ambiguous notion of similarity and the low inter-annotator agreement, it is difficult to draw robust conclusions on the differences between models based on solely this type of evaluation [Batchkarov et al., 2016]. To overcome this problem our first key contribution is a comprehensive analysis of multi-modal models. We perform large scale evaluations on different data sources, model architectures and modalities.

**Efficiency of Models and Data** Most multi-modal models require huge image and text training datasets. Our second key contribution is the proposal and analysis of a new type of hybrid modality based on small, structured data, lying in-between visual and linguistic modalities.

**Lack of Model Transparency** A further crucial issue with embeddings (and recent ML models in general) is that the learnt representations are not interpretable for humans. Thus, we are prone to overlook spurious correlations, or data and model biases [Kaur et al., 2020, Hooker, 2021, Bender et al., 2021]. To mitigate this problem, the third main proposal of this work is a framework of *transparent* and *interpretable* analyses of semantic space representations. Interpretability has gained traction in AI in the past few years not just for downstream performance but also for AI Safety and Fairness reasons [Barocas et al., 2019, Bender et al., 2021, Kaur et al., 2020]. We introduce various quantitative and qualitative analyses to understand how our models conceptualise the "world", which depends on model architecture, data source and modality.

To address the above problems, we propose, and present proof-of-concept studies of a three-pillar analysis framework of multi-modal embeddings:

1. **Black-Box Performance testing** – How representations of different modalities **perform** on intrinsic evaluation tasks? We extended previous work with the following:
   (a) *Comprehensive analysis* of models across data sources, machine learning models and modalities,
   (b) *New modality* based on small data, lying in-between low level visual information and high level linguistic / symbolic data, and
   (c) *Efficiency* analyses, controlling for data size, data distribution and model size.
2. **Transparency testing – Qualitative / Quantitative structural analysis**: **How** representations of different modalities differ? An analysis of concept structures captured by modalities.
3. **Transparency testing – Independence analysis**: An information-theory based analysis to measure **how much** representations differ?

This thesis was inspired by a series of previous work. They are detailed in Chapter 2 where we introduce the background. To highlight a few influential related work: Kiela et al. in [Kiela et al., 2014] introduced enlightening analyses of multi-modal embeddings. They showcased how image dispersion affects multi-modal embedding performance, and how word concreteness is a relevant factor. Our methodology of structural embedding analysis was partially inspired by [Minnema and Herbelot, 2019] who used various metrics to measure the similarity between a linguistic embedding space and a brain image embeddings space. Our theoretical semantic embedding framework generalises Katrin Erk's definition of distributional models [Erk, 2016]. Our information-theoretical framework and experiments were supported by the work of Zoltán Szabó [Szabó, 2014], who kindly offered consulting on the theoretical background.

Understanding how machine learning models "understand" concepts is a crucial step towards managing model and data bias, which impacts billions of users on a daily basis who interact with AI models on social media platforms, jurisdiction or health care practices. We hope that our methodology for analysing model conceptualisation will inspire other researchers to release more interpretable model analyses, therefore contributing to safer and fairer AI system development.

## 1.1 Key Contributions

The contributions of this thesis can be summarised in three key points:

I. A **comprehensive analysis of multi-modal models** – involving visual and linguistic data – across data sources, model architectures and modalities.

II. Introduction and analysis of a **new type of modality**: a visually structured, text based semantic representation, lying in-between visual and linguistic modalities.

III. Proposing and presenting proof-of-concept studies of a **transparent**, interpretable semantic space analysis framework.

The course of this research and the design of the experiments were led by the pursuit for answering the following questions:

1. How does the source of images affect the performance of multi-modal semantic representations?

2. Does the number of images have an impact on performance?

3. Do previous findings on complementary visual information scale to different types and sizes of linguistic corpora?

4. Does visual data bolster performance only because we add more data or does it convey complementary *quality* information compared to a higher *quantity* of text?

   (a) Can we achieve comparable performance using small-data if it comes from the right data distribution?

5. Can we move beyond performance evaluation? Are there any emergent concepts in embeddings? Can we quantify the difference between the concept structures of semantic spaces?

6. Can we quantify the difference between semantic spaces, based on the useful information they contribute to the meaning representation?

## 1.2 Thesis Outline

Chapter 2 gives an overview of the background and literature in Distributional Semantics, Computer Vision and multi-modal semantics, and also introduces our framework of transparency analysis. Details and discussion of the data sources and evaluation methodology are presented in Chapter 3.

Chapters 4, 5 and 6 involve implementation details and results of experiments, designed to answer the research questions from Section 1.1. Chapters 4 and 5 implement our first and second key contributions I. **comprehensive analysis of multi-modal models** and II. introduction and analysis of a **new type of modality**. The experiments focus on Questions 1, 2 and 3. Section 4.1 addresses Questions 1 and 2, evaluating different visual data sources for semantics, in terms of the impact of image quantity and quality. Section 4.2 introduces a novel structured embedding as a new modality. In Section 4.3 a broader study is presented which, tacking Question 3, aims to perform a wide range of evaluations across several different visual, linguistic and multi-modal models. As an outlook over the application of word embedding initialisations we investigate a textual entailment task in Section 4.4. Chapter 5 provides a more in-depth investigation of the effects of data size and frequency distributions in linguistic and multi-modal embeddings (Questions 4 and 4a).

Finally, in Chapter 6 we implement the third key contribution of this thesis: III. a **transparent**, interpretable semantic space analysis. We address Question 5, where we employ qualitative structural analysis of semantic spaces, and Question 6 by presenting a method for estimating the information different modalities add to the linguistic representations.

A summary, conclusions and ideas for future directions based on this research are discussed in Chapter 7. Appendices A, B, C, D, E and F contain extra results, which were omitted from the main text for space and readability considerations.

## 1.3 Publications

**Content involving thesis material:**

- Anita L. Verő and Ann Copestake. Efficient Multi-Modal Embeddings from Structured Data. *arXiv preprint arXiv:2110.02577*, 2021.
- Douwe Kiela, Anita L. Verő, and Stephen Clark. Comparing Data Sources and Architectures for Deep Visual Representation Learning in Semantics. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-16)*, 2016.

**Thesis-related content:**

- Christopher Davis, Luana Bulat, Anita L. Verő, and Ekaterina Shutova. Deconstructing multimodality: visual properties and visual context in human semantic processing. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (* SEM 2019), pages 118–124*, 2019.
- Christopher Davis, Luana Bulat, Anita L. Verő and Ekaterina Shutova. Modelling Visual Properties and Visual Context in Multimodal Semantics. *In Workshop on Visually Grounded Interaction and Language, NIPS*, Montreal, Canada, 2018.

**Not directly thesis-related content:**

- Douwe Kiela, Luana Bulat, Anita L. Verő and Stephen Clark. Virtual Embodiment: A Scalable Long-Term Strategy for Artificial Intelligence Research. In *NIPS Workshop on Machine Intelligence (MAIN)*, Barcelona, Spain, 2016.

**Software**

- **EmbEval**: The implementation of transparent evaluation methodology and the majority of experiments are available as an open source software[1]. This code was used in Chapters 4, 5 and 6. Details on its usage can be found in the documentation[2].
- **MMFeat - Flickr API**: I implemented a Flickr API and some experiment and demo code into the MMFeat software[3], which is used in Chapter 4.[4]
- **Concept Game**: A two player, collaborative gamified data collection app[5] (See Section 6.2.2.) This code is also publicly available on Github[6].

[1] https://github.com/anitavero/embeval

[2] https://anitavero.github.io/embeval/

[3] https://github.com/douwekiela/mmfeat

[4] https://github.com/anitavero/mmfeat/commits?author=anitavero

[5] http://concept-guessing-game.com/

[6] https://github.com/anitavero/concept_game

# Chapter 2

# Background and Motivation for Interpretable Multi-Modal Word Embedding Analysis

In this chapter we place the thesis into the context of previous work. We explain the motivation for our intrinsic and information-theory based analyses. Furthermore, we introduce the framework and notation used throughout the thesis.

## 2.1 What does Word Meaning Mean, and Why should We Care?

### 2.1.1 Philosophical Accounts

Traditionally, word semantics has been discussed in the framework of lexical competence. According to the externalist view, words have an objective meaning known by a “perfect competent speaker”, however, people are imperfect speakers, hence the difference between our levels of understandings [Kripke, 1972, Putnam, 1970]. This has been criticised by many including Chomsky in 2000 [Chomsky et al., 2000]. The most notable criticism came from the contextualist and pragamatic point of view. Similarly to Wittgenstein [Wittgenstein, 1953, p. 20], it identifies meaning with use, and highlights the contextual nature of word meanings [Grice, 1975, Searle, 1985].

To demonstrate the two opposing positions, take the following example sentence: “There is milk in the fridge”. According to the contextualists: in the context of morning breakfast it will be considered true if there is a carton of milk in the fridge and false if there is a patch of milk on a tray in the fridge, whereas in the context of cleaning up the kitchen truth conditions are reversed [Gasparri and Marconi, 2021]. The externalist could object by challenging the contextualist’s intuitions about truth conditions. “There is milk in the fridge”, she could argue, is true if and only if there is a certain amount (a few molecules will do[1]). The contextualist’s reply is that, in fact, neither the speaker nor the interpreter is aware of such alleged literal content if there is even such a thing.

A cognitive approach characterizes Marconi’s [Marconi, 1997] account of lexical semantic competence. In his view, lexical competence has two aspects: an inferential aspect, underlying performances such as semantically based inference and the command

[1] This example was given in [Gasparri and Marconi, 2021], however, we would point out that there is no such thing as “milk molecules” [Lucey et al., 2017], which supports scepticism towards an extreme externalist approach.

of synonymy, hyponymy and other semantic relations; and a referential aspect, which is in charge of performances such as naming (e.g., calling a horse "horse") and application (e.g., answering the question "Are there any spoons in the drawer?"). According to his theory of individual competence, communication depends both on the uniformity of cognitive interactions with the external world and on communal norms concerning the use of language, together with speakers' deferential attitude toward semantic authorities.

Recanati [Recanati, 2004] has extended the contextualised view with including the history of a word's meaning. He says a word has a "semantic potential" defined as the collection of past uses of a word between source situations (i.e., the circumstances in which a speaker has used a word) and target situations (i.e., candidate occasions of application of the word).

### 2.1.2 (Cognitive) Linguistics and Neuroimaging

At the beginning of the 1970s a new cognitive theory of the mental representation of categories surfaced [Mervis and Rosch, 1981]. It put forward the notion on prototypes which revolutionized the existing approaches to category concepts and was a leading force behind the birth of cognitive linguistics. Later a whole paradigm, called Simulationism emerged with a series of evidence between mental realisation of concepts and sensory-motor activation. For example listening to sentences that describe actions performed with the mouth, hand, or leg activates the visuomotor circuits [Tettamanti et al., 2005]; or odor-related words ("jasmine", "garlic", "cinnamon") differentially activates the primary olfactory cortex [González et al., 2006]. This all lead to theories such as the dual coding hypothesis, which is in relation to the philosophical problem of symbol grounding, discussed in detail in Section 2.4.

**Distributional Hypothesis** According to the summary of [Lenci, 2008], although the linguistic context appears as one of the ingredients of human conceptualization, the emphasis of cognitive semantics is on an intrinsically embodied conceptual representation of aspects of the world, grounded in action and perception systems. On the other hand, the *Contextual Hypothesis* in psychology arguing for a "usage-based" characterization of semantic representations incited linguistics towards statistical corpus analysis. According to Lenci, this view is related to Wittgenstein's claim, i.e. that "the meaning of a word is its use in the language". This led to the *Distributional Hypothesis* (DH) according to which at least certain aspects of the meaning of lexical expressions depend on the distributional properties of semantic similarity between two such expressions. Or as Firth [Firth, 1957] put it, "Words that occur in similar contexts tend to have similar meanings" [Turney, 2010].

There is an increasing evidence towards the "strong" version of DH which does not only assumes *correlation* between semantic content and linguistic distributions. This version is a *cognitive hypothesis* stating that repeated encounters with words in different linguistic contexts eventually lead to the formation of a contextual representation. That is an abstract characterization of the most significant contexts with which the word is used [Lenci, 2008]. Baroni and Lenci found important similarities between distributional models and human-generated properties but also striking differences [Baroni and Lenci, 2008]. Statistical representations of word meaning has since become a prevalent approach forming the basis of computational linguistics. [Boleda, 2020] summarised the reasons behind this in three factors. First, distributional representations are *learnt* from natural language data, scaling up to very large vocabularies, thus providing a coherent system where systematic explorations are possible. Second, recent models involve *high dimen-*

*sional* representations. Third, they use *continuous* values and similarity metrics. Both of the latter allow for rich and nuanced information to be encoded and analysed.

**Concepts, words and senses** In philosophy, historically there has been many different definitions of the term *concept* [Margolis and Laurence, 2021]. We use an empiricist, embodied definition which treat *concepts* as internal human cognitive knowledge representation, which probably involves multi-modal sensory based representation, as mentioned earlier. *Words* are elements of a language with *meaning*. However, human language is ambiguous, so many words can be interpreted in multiple ways depending on the context in which they occur. For instance, consider the following sentences (from [Navigli, 2009]):

(a) I calculated the *interest* rate.

(b) They have an *interest* in music.

The occurrences of the word *interest* in the two sentences clearly denote different meanings: financial earnings and passion, respectively. These different meanings of a word are called *word senses*, which are abstractions over word meanings [Lenci, 2008].

**Neuroimaging** The development of neuroimaging techniques such as PET, fMRI and ERP has provided further means to adjudicate hypotheses about lexical semantic processes in the brain, which has been studied in relation to statistical semantic models, e.g. [Mitchell et al., 2008, Pereira et al., 2018, Handjaras et al., 2016]. Mitchell et al. found correlation between distributional models of word meanings and brain imaging representations in human participants [Mitchell et al., 2008]. Handjaras et al. found that conceptual knowledge in the human brain relies on a distributed, modality-independent cortical representation that integrates the partial category and modality specific information retained at a regional level [Handjaras et al., 2016]. This thesis also complements standard semantic evaluations with tests on neuroimaging datasets, introduced in Section 3.2.2.

**Introducing Model-Concepts** In this thesis – similarly to Lenci and Boleda – we treat distributional semantic models of word meaning as a proxy to empirically investigate "aggregated meanings", which is not the semantic model of any particular individual (and most likely not even a particular society's). Since human concept representations seem at least partially perceptual, we focus on *multi-modal* distributional models involving visual perceptual data. We start from statistical models of word meaning, but we proceed towards more in-depth model interpretation analysis. We investigate whether there are structures in our learnt representations which represent some kind of *conceptualisation* of the machine. We call these **model-concepts**. Model-concepts are different from human cognition. They are also not directly *word meaning* representations as we are looking for further emerging structures / clusters. Since we are studying the fusion of linguistic and perceptual data, model-concepts are assumed to be closer to human concepts than purely text based ones. Throughout the thesis we will use "concept" and "model-concept" interchangeably, as our investigation only involves model-concepts, not human conceptual representations.

We introduce the history of Distributional Semantic models in more detail in Section 2.2, visual models from Computer Vision in Section 2.3 and multi-modal literature in Section 2.4.

## 2.2 Linguistic Embeddings: From Text to Meaning

This section reviews the history of statistical models of word semantics based on text corpora.

### 2.2.1 Distributional Semantics

In Natural Language Processing, word meaning representation models have been primarily inspired by Firth's *distributional hypothesis* [Firth, 1957], saying "Words that occur in similar contexts tend to have similar meanings" [Turney, 2010]. Contemporary corpus-based approaches implement this idea by using vector representations of words also known as distributional semantic models or embeddings. The representation vector of each word can be computed from the co-occurrence frequencies with other terms in the same context. Here, we give a short overview of the development of distributional semantic models; for a detailed survey, see Clark's book chapter in The Handbook of Contemporary Semantic Theory [Clark, 2015] or a more recent overview of Distributional Models of Word Meaning by Lenci [Lenci, 2018].

The history of word representations by vectors goes back to Karen Spärck Jones' 1967 work in Computational Linguistics who first used a principled technique for comparing contexts [Spärck Jones, 1967]. Vector representation was widely popularised for the document retrieval problem in Information Retrieval [Schütze et al., 2008]. At the beginning, both the query and the documents were represented with a "bag of words", i.e., a vector of word frequencies. This was a successful model despite the fact that it does not account for word order. To circumvent bias towards frequent words, weighted versions have been introduced, such as the *term frequency-inverse document frequency (tf-idf)* based on the frequency of terms in a document, and the inverse of the number of documents in which a term occurs. One useful way to think about document vectors is in terms of *term-document* matrix. This way, rows can correspond to document vectors, whereas columns are word representations. A popular method was to apply a dimensionality reduction technique on such matrices, such as singular value decomposition (SVD). The application of SVD to the term-document matrix was introduced by Deerwester et al. [Deerwester et al., 1990], who called the method Latent Semantic Analysis (LSA). The name comes from the intuition that LSA teases out a latent meaning from the co-occurrence data, by clustering words along a small number — typically a few hundred — of semantic, or topical, dimensions [Turney, 2010].

From the *term-document* matrix we can easily arrive to the concept of *term-term* matrix. Instead of treating the document as the context similar words co-occur in, we can narrow it down to a smaller window around a word. This way the elements of a matrix are the frequency of two words occurring in the same context window. To normalise raw frequencies using Positive Pointwise Mutual Information (PPMI) of two words $(w1, w2)$ is a popular method:

$$\text{PPMI}(w1, w2) = \max(\log_2 \frac{P(w1, w2)}{P(w1)P(w2)}, 0). \tag{2.1}$$

Applying SVD can also be useful on these type of matrices.

Representing the meaning of multiple-word phrases or sentences, still proves to be a challenging problem. Many researchers have studied compositional semantics using vector operations on word vectors [Mitchell and Lapata, 2010] or tensor based representations [Clark, 2015].

### 2.2.2 Shallow Networks

Recent research has presented several neural network-based approaches to learn word vector representations. Such distributed representations have become known as embeddings. The most well known and widely used models were introduced by Mikolov et al. [Mikolov et al., 2013a, Mikolov et al., 2013b] and have become popular as part of the *word2vec* toolkit. They introduced two models, both consisting of a shallow, two-layer neural network which learns an approximation of co-occurrence statistics [Levy and Goldberg, 2014b]. They train a neural network to predict neighbouring words, in doing so learning dense embeddings for the words. It is much faster than SVD and easy to train.

The skip-gram (SG) model [Mikolov et al., 2013b] learns to predict the words that can occur in the context of a target word. Its objective function is as follows:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c,c\neq 0}\log p(w_{t+j}|w_t) \tag{2.2}$$

where $T$ is the size of the corpus, $c$ is the context window size, $w_i$ is a word, $(1 <= i <= T)$.

Let $d$ be the embedding dimension, $V$ the vocabulary. The model learns two embeddings, or lookup matrices: 1) an input embedding $W \in \mathbb{R}^{d\times|V|}$, where column $i$ gives the embedding $v_i$ of size $1 \times d$ for word $w_i$ in the vocabulary 2) an output embedding $W' \in \mathbb{R}^{|V|\times d}$, where row $i$ is a $d \times 1$ embedding $v'_i$ for word $w_i$ in $V$. $v'_O$ and $v_I$ are the "input" and "output" vector representations of $w$. The probability of a word occurring in a context is given by the softmax function:

$$p(w_O|w_I) = \frac{\exp(v'_O \cdot v_I)}{\sum_{j=1}^{|V|}\exp(v'_j \cdot v_I)} \tag{2.3}$$

This architecture is illustrated in Figure 2.1.

Because of the denominator term, training this model directly would be computationally infeasible. For this reason Mikolov et al. introduced the trick of hierarchical softmax and skip-gram with negative sampling (SGNS).

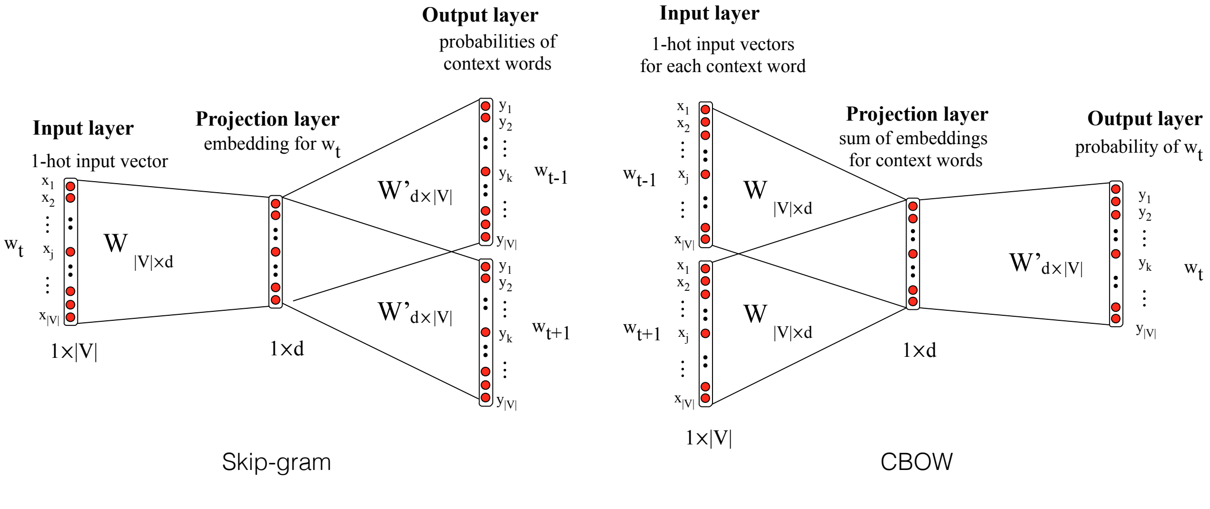


Figure 2.1: Skip-gram and CBOW architectures.[2]

Since we have two embeddings $v_j$ and $v'_j$ for each word $w_j$ we can either just use, $v_j$, sum or concatenate them.

[2] https://web.stanford.edu/~jurafsky/li15/lec3.vector.png

If we multiply $WW'^T$, we get a matrix $M$, each entry $m_{ij}$ corresponding to some association between input word $i$ and output word $j$. Levy and Goldberg [Levy and Goldberg, 2014b] show that skip-gram reaches its optimum just when this matrix is a shifted version of the PMI matrix:

$$WW'^T = M^{PMI} - \log k \tag{2.4}$$

Thus, skip-gram is implicitly factoring a shifted version of the PMI matrix, into the two embedding matrices.

In the other model of Mikolov et al., called Continuous Bag of Words (CBOW) [Mikolov et al., 2013a], a similar training happens, except instead of predicting the context around a word in a window, the objective is to predict the middle word in the context window. The two model architectures are illustrated in Figure 2.1.

Global Vectors model (GloVe) [Pennington et al., 2014] aims to learn a version of the PMI matrix which is weighted toward more frequent word context pairs. They theorise that the fact that their model can be optimised directly as opposed to the on-line training of SGNS, it introduces more global frequency information. However, Levy and Goldberg showed, that after tuning hyperparameters, it does not produce any performance gain [Levy et al., 2015].

Other versions of skip-gram have been proposed such as a dependency-based word embedding [Levy and Goldberg, 2014a], where instead of using a simple sliding window as the context, a window goes through the dependency graph of each word as the context.

Deep Recurrent Neural Networks [Bengio et al., 2003, Bahdanau et al., 2015, Cho et al., 2014, Kiros et al., 2015, Wang and Jiang, 2015, Rocktäschel et al., 2016] and Transformers with self-attention [Peters et al., 2018, Radford et al., 2018, Devlin et al., 2019, Yang et al., 2019] have appeared in the forefront of NLP research in the past few years. They achieve state-of-the-art performance on various sentence level tasks, included in the GLUE multi-task benchmark for Natural Language Understanding [Wang et al., 2018a]. The tasks involve textual entailment, sentiment analysis, paraphrasing and question answering. Since the main objectives of this thesis were creating and testing a framework for comprehensive, transparent and interpretable semantic analysis, we use the smallest possible models which allow us to incorporate visual embeddings, thus studying multi-modality. Therefore, in this work we apply shallow network type models, as visual embeddings fit into them more easily then into count based models, while being the simplest neural models. Due to the few parameters of these models, they are also much easier to train than bigger neural models, allowing us to run comprehensive studies across several datasets and model types. Throughout this work we use SGNS and FastText, which uses the CBOW model, with versions extended with subword information [Mikolov et al., 2018]. Furthermore, we use different versions of PMI in Section 6.4 for analysing our training corpora. Applying our framework for the latest transformer type models would be a straightforward application of this thesis. Although running broad-scale analysis is much more challenging using these large models, it would be interesting to see how attentions affects multi-modal fusion.

## 2.3 Visual Embeddings: From Images to Meaning

Our research focuses on the most efficient fusion of vision and language for meaning representations. Thus we revise the basics of Computer Vision approaches for encoding images as well as state-of-the-art models in Section 2.3.1, which we rely on.

Similar to language embeddings, representing the content of an image or a video also involves producing a vector representation. This is expected to capture a compressed representation of interesting features over the high dimensional, raw pixel input that corresponds to human semantic constructs. This can include low level features such as edges and corners, or higher level ones such as objects of an image or temporal patterns on a video. The selection of these features, however, is not a trivial task. Traditional Computer Vision methods applied hand-crafted features similar to the above mentioned edge and corner detectors from which they could build a Bag-of-words type model [Sivic and Zisserman, 2003].

Neural Networks revolutionized this area as well with the introduction of Convolutional Neural Networks (CNNs). These are biologically inspired networks motivated by the visual cortex [Lecun et al., 1998]. They are capable of learning high level features gradually by exploiting a deep structure where every layer learns a higher abstraction based on the lower ones. Such networks can be trained for many different tasks such as object classification [Simonyan and Zisserman, 2014, Krizhevsky et al., 2012, Szegedy et al., 2015, He et al., 2016], image segmentation [Kendall et al., 2017] or action recognition [Sharma et al., 2015]. The learned vectors proved to be a good basis for learning high performing image embeddings [Kiela and Bottou, 2014].

The core building block of such networks is the *convolutional layer*. This refers to the mathematical convolution of a filter function across the pixels of an image. In traditional Computer Vision this filter function (or kernel) was crafted manually, whereas in a CNN it is learned from data. Down-sampling and learning compressed local (globally invariant) features is done by the *pooling* layers. CNNs usually involve fully connected layers on the top and activation functions similar to other neural networks. They are usually trained with an objective for a supervised task, such as object classification.

Figure 2.2 illustrates the architecture of LeNet [LeCun et al., 1989], the first CNN successfully trained by back-propagation to classify hand-written digits. It performed better than manual coefficient design, and was suited to a broader range of image recognition problems. Thus, it became the foundation of modern Computer Vision.

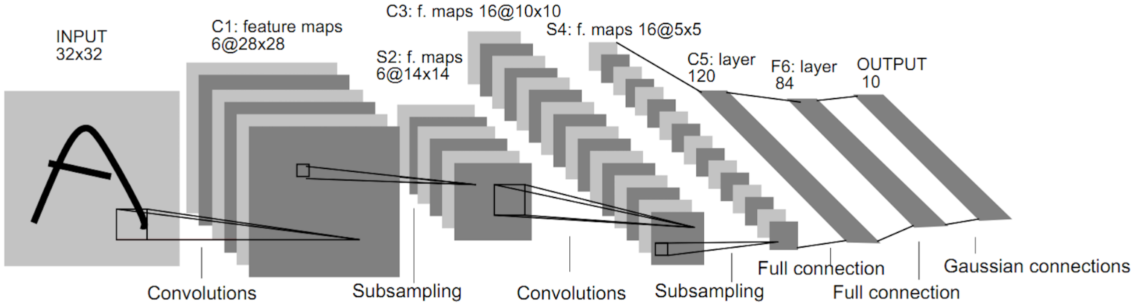


Figure 2.2: Architecture of the LeNet-5 for digit recognition. Each plane is a feature map i.e. a set of units whose weights are constrained to be identical.

### 2.3.1 CNN Models

In our study, CNN models serve the role of encoding images into visual word semantic vectors. We used four architectures which differ in size and structure. See Table 2.1 for an overview.

| | **AlexNet** | **GoogLeNet** | **VGGNet** | **ResNet** |
|---|---|---|---|---|
| ILSVRC winner | 2012 | 2014 | 2015 | 2015 |
| #Layers | 7 | 22 | 19 | 152 |
| #Parameters (million) | ~60 | ~6.7 | ~144 | ~6.8 |
| Receptive field size | $11 \times 11$ | $1 \times 1, 3 \times 3,$<br>$5 \times 5$ | $3 \times 3$ | $3 \times 3$ |
| Fully connected layers | Yes | No | Yes | Yes |

Table 2.1: Network architectures. Layer counts only include layers with parameters.

**AlexNet** The network by Krizhevsky [Krizhevsky et al., 2012] introduces the following network architecture: first, there are five convolutional layers, followed by two fully-connected layers, where the final layer is fed into a softmax which produces a distribution over the class labels. All layers apply rectified linear units (ReLUs) [Nair and Hinton, 2010] and use dropout for regularization [Hinton et al., 2012]. This network won the ILSVRC 2012 ImageNet classification challenge.

**GoogLeNet** The ILSVRC 2014 challenge winning GoogLeNet [Szegedy et al., 2015] uses “inception modules” as a network-in-network method [Lin et al., 2013] for enhancing model discriminability for local patches within the receptive field. It uses much smaller receptive fields and explicitly focuses on efficiency: while it is much deeper than AlexNet, it has fewer parameters. Its architecture consists of two convolutional layers, followed by inception layers that culminate into an average pooling layer that feeds into the softmax decision. That is, it has no fully connected layers. Dropout is only applied on the final layer. All connections use rectified units.

**VGGNet** The ILSVRC 2015 ImageNet classification challenge was won by VGGNet [Simonyan and Zisserman, 2014]. Like GoogLeNet, it is much deeper than AlexNet and uses smaller receptive fields. It has many more parameters than the other networks. It consists of a series of convolutional layers followed by the fully connected ones. All layers are rectified and dropout is applied to the first two fully connected layers.

**ResNet** ResNet [He et al., 2016] revolutionized the CNN architectural race by introducing the concept of residual learning in CNN and devised an efficient methodology for training of deep nets. He et al. proposed a 152-layers deep CNN, which won the ILSVRC 2015 competition. ResNet, which was 20 and 8 times deeper than AlexNet and VGG respectively, showed less computational complexity than previously proposed nets. They empirically showed that ResNet with 50/101/152 layers has less error on image classification task than 34 layers plain net.

These networks were selected because they are very well-known in the Computer Vision community. They exhibit interesting qualitative differences in terms of their depth (i.e., the number of layers), the number of parameters, regularization methods and the use of fully connected layers. They have all been winning network architectures in the ILSVRC ImageNet classification challenges[3].

[3] `https://image-net.org/challenges/LSVRC/`

## 2.4 Multi-modal Semantics

### 2.4.1 Symbol Grounding

Despite their undeniable success, textual embeddings have their own limitations regarding the grounding of meaning to the outside world, often referred to as Harnard's *symbol grounding problem* [Harnad, 1990]. Similarly, Computer Vision research has reached a point where leveraging non-visual common sense knowledge is necessary for further improvement even on purely vision based applications. It is motivated by an insight from cognitive science (Section 2.1.2): the human semantic representation of symbols (e.g., words or objects) is based on multi-modal sensory inputs perceived on a lifelong basis [Roy, 2005].

When it comes to applications and models the question arises: What do we mean by grounding in practice? In what way can multi-modal data contribute to meaning representations? We can distinguish between two main approaches for grounding:

*Referential grounding* refers to the task of determining the referent that a word denotes in the context of the other modality (e.g., a specific object in an image). The core issue here is finding a mapping between the two spaces [Lazaridou et al., 2016].

In contrast, *representational grounding* addresses the problem of multi-modal semantics: Representing the grounded meaning of a word in the sense of fusing different modalities into one, richer semantic representation [Bruni et al., 2014].

While all these results are promising some fundamental questions are still unexplored.

**Non-Visual Tasks** Most work focuses on evaluation tasks (and therefore on models which perform well on them) which involve direct visual input. These are usually referential type tasks such as Visual Question Answering (VQA) [Srivastava and Salakhutdinov, 2012, Kiros et al., 2014, Socher et al., 2014, Tsai et al., 2019, Lu et al., 2019, Su et al., 2019, Majumdar et al., 2020]. In these cases the usefulness of visual input is not surprising. Fewer papers have exploited visual information for representational grounding, when the evaluation tasks involve no direct visual input, such as semantic similarity [Bruni et al., 2014, Kiela and Bottou, 2014, Kiela et al., 2016, Lazaridou et al., 2015, Davis et al., 2019, Lin and Parikh, 2015, Vendrov et al., 2015]. Lin [Lin and Parikh, 2015] introduced a fill-in-the-blank task, which has been done, however, using abstract images. A further interesting proposal relates to the so-called order-embeddings, a general hierarchical framework for hypernymy, textual entailment, and image captioning [Vendrov et al., 2015]. However, it still does not involve a thorough investigation of multi-modal fusion possibilities. Some papers including [Kiela and Bottou, 2014, Kiela et al., 2016, Lazaridou et al., 2015, Davis et al., 2019] perform intrinsic analysis of multi-modal embeddings. However, the reasons for the impact of visual information are not well understood, for we see only correlations on intrinsic evaluation tasks.

This work investigates visual information's contribution to meaning representations on evaluation tasks involving no direct visual input. We aim to showcase a proof-of-concept framework for deeper analysis of unsupervised multi-modal representations. We study the concepts which emerge in grounded meaning representations.

**Cost of Data** All the mentioned tasks require huge image datasets with expensive human annotation. In the case of multi-modal tasks these annotations are even more difficult to acquire, since annotating combinations of texts and images/videos can be even more complicated and time consuming than in the uni-modal cases.

We try to circumvent the problem of the costs by studying model and data size efficiency (introduced in Section 2.7.2) as well as alternatives for new modalities based on small data (Section 2.5).

### 2.4.2 Early-, Late- and Mid-fusion

In the literature, we can find three ways for performing the fusion of textual and perceptual information:

- In *early fusion*, one learns a joint representation from the two spaces, then computes a function for the specific task (e.g., cosine distance for measuring semantic relatedness) [Lazaridou et al., 2015, Kottur et al., 2015].
- *Mid-fusion* techniques learn separate representations for each modalities, then combine them into a multi-modal representation, finally they compute the function for the task [Kiela et al., 2014].
- *Late fusion* methods also learn uni-modal representations separately, then compute a function for each modality individually, and combine function outputs at the end [Silberer and Lapata, 2014].

Figure 2.3 illustrates the three types of fusion techniques. In this work we focus on mid-fusion based models since it allows us to study the information preserved in the individual modalities.

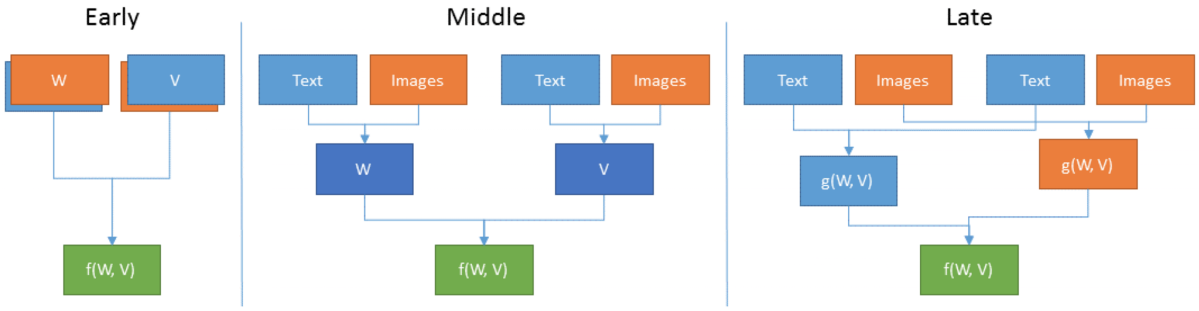


Figure 2.3: Fusion methods for combining textual and perceptual information. $V$ and $W$ are representations learnt from either Text or Images. $f$ is a function that fuses two representations in Early and Middle fusion. In Late fusion $f$ combines the outputs of functions $g$ which embed uni-modal data. (Figure is borrowed from the "Multimodal Learning and Reasoning" ACL 2016 tutorial[4].)

### 2.4.3 Multi-modal RNNs and Transformers

Neural networks and recurrent networks have been used on multi-modal input since they got popular, even going back to Boltzmann machines [Srivastava and Salakhutdinov, 2012, Kiros et al., 2014]. They were mainly tested on image retrieval and caption generation tasks. Architectures, such as Tree RNNs have also been applied to cross-modal tasks [Socher et al., 2014].

The latest NLP models have also inspired the creation of new multi-modal representations. Tsai et al. [Tsai et al., 2019] developed a multi-modal Transformer model using

[4] `http://multimodalnlp.github.io/mlr_tutorial.png`

cross-modal attention and tested it on sentiment analysis tasks in videos. Lu et al. [Lu et al., 2019] created ViLBERT, a multi-modal model based on BERT. They pre-trained it on Conceptual Captions dataset and then transferred it to multiple vision-and-language tasks — visual question answering, visual common-sense reasoning, referring expressions, and caption-based image retrieval.

## 2.5 Structured Embeddings: Motivation for a New Modality

The multi-modal framework we introduce in this thesis can be used to any modalities (such as text, image, video, audio). In the experimental part of this work we focus on fusing linguistic and visual information. As we saw in the previous section, ample research exploited large visual datasets and CNN models with increasingly large number of parameters. This is a fairly expensive way of injecting visual information into meaning representations.

The second key contribution of this thesis is thoroughly exploring a structured visual dataset, called Visual Genome [Krishna et al., 2016], and the way it can enrich meaning representations. Visual Genome contains images with bounding box annotations as well as text annotation in a graph structure (it is detailed, among all the other datasets we use, in Section 3.1). This would be beneficial for two reasons. First, structured data can serve as a bridge over the semantic gap between low level image data and high level symbolic information in text. Secondly, it can provide a small data alternative to big data driven models, which could become the basis of essential tools in situations where a huge amount of text is not available, but where more structured data could be easier to collect.

By exploiting this textual dataset based on a visual structure, this work introduces a new type of embedding, which we consider as a new, hybrid modality. In the next section we introduce our general framework of modalities. The new embedding modality called Structured Embeddings will be introduced in Section 2.6.1. The details of its creation is explained in Section 4.2.

## 2.6 Generalisation of Embeddings: Proposed Framework and Formalism

In this work we use a general notion of Embedding, which refers to a vector space representation of word meanings. The weights of each vector, however, can be set by any machine learning algorithm, trained on any data type, such as text, images, sound, structured datasets etc. The only criterion for calling a vector space a word embedding space is that we find an interpretation of the dataset where it represents words.

We formally define Semantic Embedding models as tuples of their relevant parameters. We generalise Katrin Erk's definition of distributional models [Erk, 2016] to include word representations based on other modalities as well. We denote modality by $m \in \{L, V, S\}$, which can take the value of linguistic $L$, visual $V$ or structural $S$. The parameters of a semantic embedding model of modality $m$ are the following: A set of $T$ target words that receive vector representations, a set $O_m$ of observable context items in a dataset $D_m$, an *extraction function* $X_m$ which chooses relevant contexts in which to look for context items, and a *mapping function* $A_m$, which maps from target and context items to a $d_m$ dimensional space $\mathbb{R}^{d_m}$. The mapping for all target elements is represented by an Embedding matrix $E_m \in \mathbb{R}^{|T| \times d_m}$.

$T$ is an arbitrary set of words, $D_m$ is a set of data items. $D_m$ includes target representations $r \in D_m$ with a relation to $t \in T$ target elements $r \sim t$. $O_m$ is all the potential target contexts in the dataset: $O_m : T \rightarrow \mathcal{P}(D_m), O_m(t) = \{U \subset D_m \,|\, \exists r \in U, r \sim t\}$, where $\mathcal{P}$ is the power set. The *extraction function* $X_m$ returns "relevant" context items from $O_m$ to each target element from $T$ – that is it returns a mapping from target/context item pairs to numbers in $\mathbb{N}$, representing a *relevance score* of context pairs: $X_m : T \rightarrow (O_m(T) \rightarrow \mathbb{N})$. We use "relevance" here in a fairly general sense: it can for example be co-occurrence counts within a text window, image search engine result relevance, or scores based on other prior assumptions about relevancy in the the dataset, such as graph neighbourhood, which we will exploit for structured data. The *mapping function* $A_m$ is a combination of a (usually machine learning) algorithm and any further pre- and post-processing method which together takes the output of $X_m$ and turns it into a mapping from targets to real values, $A_m : (T, D_m, X_m) \rightarrow (T \rightarrow \mathbb{R}^{d_m})$. The output mapping is represented by a matrix, called an Embedding $E_m \in \mathbb{R}^{|T| \times d_m}$, which is a vector space consisting of vector representations for each target word in $T$.

In summary, we define Semantic Embedding models for a modality $m$ as tuples comprising the sets of target elements, observable context items, the dataset, the extraction function, the mapping function, and the embedding dimensionality:

$$\mathcal{S}_m = \langle T, O_m, D_m, X_m, A_m, d_m \rangle \tag{2.5}$$

The output of the model is the learnt embedding $E_m$.

For example a Google Image based Semantic Embedding model would have the following parameters:

$$\mathcal{S}_G = \langle T, O_G, D_G, X_G, A_G, d_G \rangle \tag{2.6}$$

where $T$ is our target vocabulary and $D_G$ is the dataset consisting of words from $T$ and Google Image Search results for each $t \in T$. $O_G$ are all the potential subsets of image results for a given word $t$ in Google Image Search. For example we can use any number of images from the search results. The extraction function $X_G$ selects which contexts we chose, e.g., it selects the first 10 image results in Google Search Engine's relevance order. $A_G$ will include a CNN network which maps each image to a vector representation, plus an aggregation function which creates one image vector representation for each word $t$. In this case, $d_G$ will be the dimensionality of the last layer of the CNN network which we use as image representation. Thus, it will be the dimensionality of our learnt Google Image Embedding $E_G$.

Note that in general, $A_m$ is a very broad notation. It can involve any learning algorithm. If our training data is text for example, it can involve any traditional count based methods, shallow or deep neural networks or any other type of method which maps targets from a dataset with an extraction function to choose relevant contexts, to a vector representation.

In the next section we will introduce three types of Semantic Embedding models which we study in this thesis.

### 2.6.1 Embedding Modalities

In this work we are going to distinguish between three different types of embedding for each modalities $m \in \{L, V, S\}$, which are produced by three class of semantic embedding models varying in all parameters but $T$:

**Linguistic Embeddings** $E_L \in \mathbb{R}^{|T| \times d_L}$ are vector spaces which are learnt by an algorithm $A_L$ trained on large text data $D_L$. The learning algorithm can be any of the standard shallow neural models, which approximate co-occurrence statistics of words, such as SGNS, CBOW or FastText. $X_L$ corresponds to co-occurrence counts for target/context word pairs within a context window around target words.

**Visual Embeddings** $E_V \in \mathbb{R}^{|T| \times d_V}$ consist of vectors which have been trained on images $D_V$, which are associated to words by $X_V$ (e.g. images labelled with words). In this case the learning algorithm is typically a CNN network (see Section 2.3) which has a specified architecture for learning abstract patterns from image data. However, after mapping images to a vector space, we need a method which associates one vector to a word. In our case we usually have multiple image results for a word, hence this method has to be a vector aggregation, such as element-wise maximum, mean or median (discussed in Sections 4.1 and 4.3). The learning algorithm and the aggregation method together constitutes $A_V$.

**Structured Embeddings** $E_S \in \mathbb{R}^{|T| \times d_S}$ are the result of an $X_S$ which extracts relevant contexts from data $D_S$ which has a more developed structure than raw text or images on the internet. These datasets usually involve some manual design and labour for the collection, therefore they are much smaller in terms of the used computer memory in bytes. One example is Visual Genome Scene Graph annotations (introduced in Section 3.1.2), which we study in detail in Chapters 4, 5 and 6. $A_S$ is a similar algorithm to $A_L$, trained on the extracted pairs, with co-occurrence statistics.

$d_L, d_S$ depend on the output size of the shallow network model in use, usually equals to 300. $d_V$ is the size of the last layer of a CNN network.

The combination of the above embedding types can happen using one of the three fusion techniques (Section 2.4.2). Throughout this thesis we will use mid-fusion as it allows us to examine the information coming from each embeddings more easily. We denote multi-modal embeddings by $E_{m_1} + E_{m_2}, m_1, m_2 \in \{L, V, S\}, m_1 \neq m_2$.

## 2.7 Modalities as Partial Observers of Meaning

The ancient Indian parable called *Blind men and an elephant* tells a story of a group of blind men who have never come across an elephant before and who learn and conceptualise what the elephant is like by touching it. Their observations go as follows in James Baldwin's English version[5]:

> *…The first one happened to put his hand on the elephant's side. "Well, well!" he said, "now I know all about this beast. He is exactly like a wall."*
>
> *The second felt only of the elephant's tusk. "My brother," he said, "you are mistaken. He is not at all like a wall. He is round and smooth and sharp. He is more like a spear than anything else."*
>
> *The third happened to take hold of the elephant's trunk. "Both of you are wrong," he said. "Anybody who knows anything can see that this elephant is like a snake."…*

[5] https://americanliterature.com/author/james-baldwin/short-story/the-blind-men-and-the-elephant

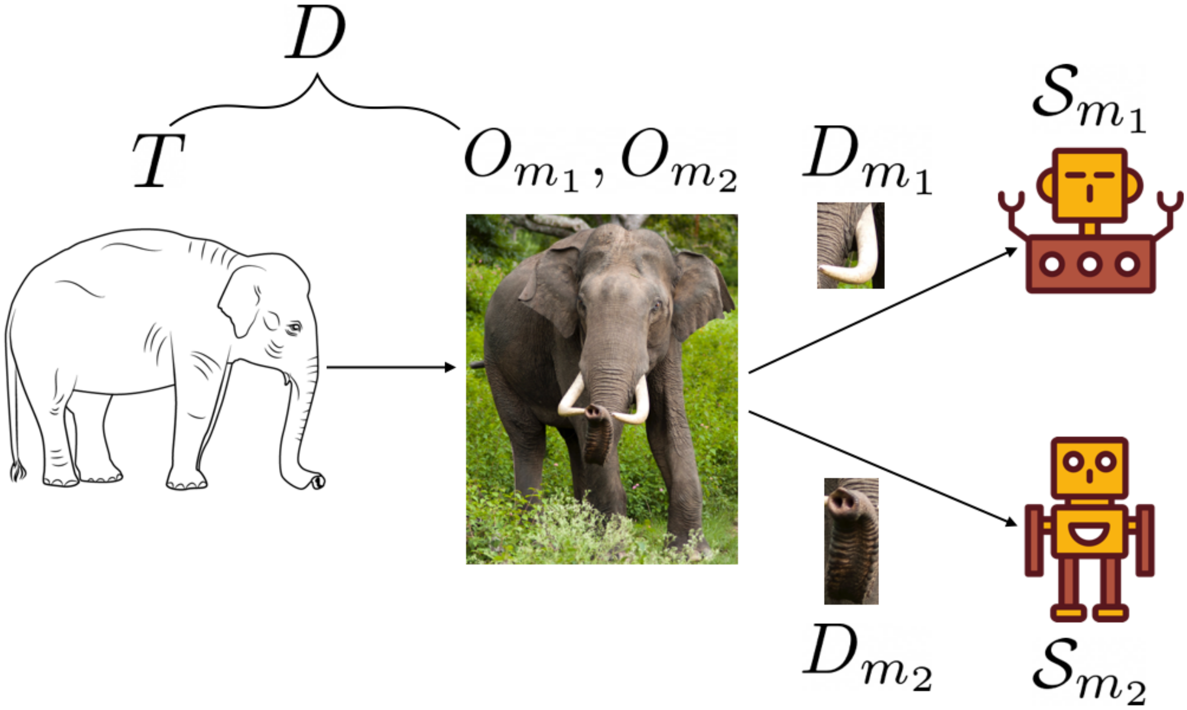


Figure 2.4: Modalities and the elephant. Illustration of the Semantic Embedding models for different modalities, which include different perspectives. Data $D$ includes the target concept $T$ of the *elephant* plus the observable contexts $O_{m_1}, O_{m_2}$, which are the *trunk* and a *tusk*. Each of the two Semantic Embedding models $\mathcal{S}_{m_1}, \mathcal{S}_{m_2}$ receives the data from their different perspectives: $D_{m_1} = (T, O_{m_1})$ and $D_{m_2} = (T, O_{m_2})$ respectively.

As for another person, whose hand was upon its leg, said, the elephant is a pillar like a tree. For the fifth whose hand reached its ear, it seemed like a kind of fan. The last one who felt its tail, described it as a rope.

Will they be able to combine their observations into one description more accurate than any of their individual ones? Or will they just disagree and become more confused than they had been?

If the blind men were touching different objects, or were in completely different universes, they would probably struggle to reach an agreement. Since, however, they are feeling the same animal, they do have a common ground, which is at first hidden from them, but which they have a chance to comprehend better together through collaboration. It only makes sense to collaborate if none of them is already an elephant expert, or talking about a completely irrelevant or random subject. Similarly, our Semantic Embedding models have a chance to combine their knowledge if done properly. Figure 2.4 presents an illustration of our multi-modal framework, with one target concept of the *elephant* and two Semantic Embedding models with different perspectives.[6]

Analogously to the imperfect lexical competence framework, mentioned in Section 2.1.1, we treat modalities as partial observers of meaning. Like the men above, we assume that they have different perspectives on the same object. This object in our case is word meaning, or rather an aggregated statistical representation of words at a specific point in time (described in Section 2.1.2).

[6] Icons made by Good Ware (`https://www.flaticon.com/authors/good-ware`) from `www.flaticon.com`. Photo of an Indian elephant is from Wikipedia (`http://web.archive.org/web/20210907113830/https://de.wikipedia.org/wiki/Datei:Elephas_maximus_%28Bandipur%29.png`), elephant drawing is from `http://web.archive.org/web/20210907105456/https://www.drawingtutorials101.com/how-to-draw-an-indian-elephant`.

Using the notation before, let's say we have $[\mathcal{S}_{m_1}, \dots, \mathcal{S}_{m_M}]$ Semantic Embedding models of $M$ different modalities. We assume:

1. **Common ground**: Each of them captures *some* aspect of word meanings. That is, we assume that the vector weights of none of the learnt embeddings $[E_{m_1}, \dots, E_{m_M}]$ are random.

2. **Perspectives**: They do not share the same knowledge, they represent different perspectives.

3. **Imperfect knowledge**: None of them has perfect knowledge: none of the Semantic Embedding models is an oracle which represents the ground truth.

In some versions of the parable the men get into a disagreement (or a fight of various degree of violence depending on the version), in others they learn that they were all partially correct and partially wrong. In the following, we will search for the best way to ensure our models of different modalities can collaborate in the most effective way.

### 2.7.1 Background and Motivation for Model Transparency

From the existing multi-modal literature we know that combining textual and visual modalities can collaborate and improve performance in various cases. Most work, however, evaluates solely on tasks, such as semantic similarity or downstream tasks, such as Visual Question Answering (VQA). It has been shown by many researchers that this traditional way of evaluating models in Machine Learning is prone to various flaws, which can be fatally misleading for the field. Kuhnle in [Kuhnle, 2020, Chapter 2] gave a comprehensive discussion of these problems. Built on this we summarise the issues in the following categories:

**Black-Box Model Performance** Since the recent deep learning revolution, ML evaluation appeared to be solely concerned with beating benchmarks on downstream-tasks, while the models are often treated as black-boxes. This often lead to models which learn "weird behaviour". For example vision models may rely on the image background to recognise an object [Ponce et al., 2006], blind spots of deep CNNs [Zhang et al., 2018], or neural models mistranslate low-frequency words into context-fitting but content-changing alternatives [Arthur et al., 2016]. Good evaluation performance on one task often does not transfer to downstream tasks either. In Section 4.4 we also present our own finding that a deep LSTM with randomly initialised input word vectors performs on par with an input of pretrained word embeddings on a Textual Entailment task (SNLI). Zhang and Bowman found the related phenomenon of high performing random initialized LSTM models [Zhang and Bowman, 2018]. This is in line with current findings considering the recent transformer type models which are shown to be far from solving general tasks (e.g., document question answering). Rather, these models are overfitting to the quirks of particular datasets [Yogatama et al., 2019]. This all leads us to conclude that looking at only performance improvements between models are mostly meaningless without further analysis.

**Dataset Bias** Data in the context of ML is supposed to convey patterns which are characteristic for a certain task. Kuhnle defines dataset bias as coincidental systematic artefacts in the data which are not characteristic of the task in question. Because of this incidentality, using such datasets as training data can result in unintentional behaviour.

For instance Wang et al. [Wang et al., 2018b] found that image captioning models for MS-COCO [Lin et al., 2014] can learn to produce reasonable captions merely by knowing about the objects in an image while ignoring, for instance, their location and relation. On VQA tasks ***modality bias*** has been shown, which refers to the systematic tendency that one modality suffices to infer the correct output with high confidence. Multiple examples were reported, such as a language-only model which completely ignores the image but can answer almost half of the questions correctly [Zhang et al., 2016]. Agrawal et al. [Agrawal et al., 2016] observed how seemingly well-performing models jump to conclusions after only the first few question words, thus concluding that they fail at complete question and image understanding. Although, Kuhnle does not include ***ethical bias*** in his definition, we think it could fit into it, by including ethical goals into our task definition. The field of AI fairness is shifting towards concentrating on *harms* rather than *bias* in the political sense [Barocas et al., 2019, p. 136-143], however, after including mitigating harm in our task objective we can use Kuhnle's data bias definition. There is a line of research on cultural stereotypes reflected in word embeddings [Barocas et al., 2019, p. 141]. Even though word embeddings per se do not correspond to any linguistic or decision-making task, analysing them before incorporating them into applications is a crucial step from an ethical point of view as well.

**Model Bias** Hooker in [Hooker, 2021] argued that bias materialises not only in data but in the algorithms as well. She argues that the key reason why model design choices amplify algorithmic bias is because notions of fairness often coincide with how underrepresented protected features are treated by the model. Most real-world data naturally have a skewed distribution with a small number of well-represented features and a "long-tail" of features that are relatively underrepresented. The skew in feature frequency leads to disparate error rates on the underrepresented attribute.

**Problems of Metrics** Lastly, evaluating meaning representations is inherently limited by the methods and possibilities of human annotation collection. On top of this, as mentioned in [Kuhnle, 2020, p. 23-24] evaluations are often prone to statistical flaws of interpreting performance scores, such as missing baseline scores, reported confidence intervals with no reference or explanation, and lacking formal comparison/hypothesis testing [Faruqui et al., 2016].

**Solutions** A range of papers have been published recently which attempt to fix some of the identified evaluation issues. Several attempts have been made to fixing data, however Torralba and Efros [Torralba and Efros, 2011] argued that such a process is likely doomed to result in a "vicious cycle" of ad hoc improvements, unless one reconsiders the underlying mechanisms which cause undesired dataset bias. Artificial data and unit testing [Fouhey and Zitnick, 2014, Johnson et al., 2017, Kuhnle and Copestake, 2017] is a promising paradigm to amend ML evaluations. Probing is a recently increasingly popular approach to "stress-testing" involving testing the model on solving an auxiliary predictive task and testing the sensitivity of the model output to modifications of the input [Conneau et al., 2018, Voita and Titov, 2020]. Approaches for interpretable models and post-hoc model explanation techniques are also growing areas [Ghorbani et al., 2019, Kaur et al., 2020].

In the next section we propose transparency analysis as an extension of the above proposed solutions aiming to prevent "vicious cycles" by promoting a more informed model development process.

### 2.7.2 Transparency Testing and Efficient Multi-Modal Fusion

A key objective of this thesis is to propose and demonstrate a framework for overcoming the inconsistency of multi-modal results. Our approach is somewhat related to the probing paradigm and partially inspired by interpretability research and cognitive science. Beyond "stress-tests" for our models we propose to extend standard evaluation techniques with an in-depth model and data analysis. We propose both going wider towards a more comprehensive model comparison across modalities and data sources, as well as deeper into studying the "cognition" of our models. We choose to analyse our datasets and models in a transparent way, which could serve as a preprocessing step before performing data or model debiasing. We propose performing and automating such data and model analyses, in order to prevent "vicious cycles" of ad hoc improvements, mentioned in the previous section.

We postulate that amending performance evaluation with more in-depth **transparency testing** of semantic models are a useful way of developing more efficient and also safer models. Getting to know our models inner "cognitive models" can be a way towards AI methods, which are capable of communicating their reasoning and also potential biases towards humans. This would make them easier to debug and maintain safely in the future.

We propose an embedding analysis leaning on three pillars. We postulate that they together form a comprehensive, interpretable semantic analysis but none of them are sufficient on their own. The three types of analysis are categorised in black/transparency testing and are aiming to answer the following questions:

1. *Performance testing*: **Black-Box** testing – How representations of different modalities **perform** on evaluation tasks trained on different datasets?

2. *Qualitative / Quantitative structural analysis*: **Transparency** testing – **How** representations of different modalities **differ**?

3. *Independence analysis*: **Transparency** testing: **How much** representations **differ**?

By learning about how and how much our different embeddings $E_L, E_V, E_S$ differ while looking at the performance scores, we can reach a conclusion on:

> *What is the most efficient way of combining our different resources*?

**Efficiency** What do we mean by efficiency? *Performance testing* is only one way to account for efficiency. When we hold a machine learning model to be efficient depends on our costs and resources. Data is often a limited resource, so in most cases it makes sense to take data size into account. Required computational resources, running times and electricity costs are also important factors to consider. Efficiency in the context of economic footprint was famously thematised by Bender et al. [Bender et al., 2021]. In this work we account for performance, data size and distribution as well as model size, as these are metrics we could easily control for. Including hardware, electricity costs and running time could be a relevant extension of our studies.

None of the three types of analysis on their own is sufficient to answer the above question, but together they have a potential for providing meaningful insight in the anatomy of multi-modal semantic models.

In Chapter 3 we will discuss the details of our approach to all three types of analysis. In the following sections we introduce our framework for transparency analysis of multi-modal models.

### 2.7.3 "Cognitive Model" of Embeddings: How do Models Conceptualise?

As the second pillar, or the first transparency analysis, we ask the question whether each of these vector spaces represent meaningful concepts as clusters, and how these concept structures relate to each other?

Comparing semantic spaces is central in Lexical Semantic Change (LSC). Dubossarsky et al. introduced *Temporal Referencing*[7] for robust modelling of LSC on diachronic corpora [Dubossarsky et al., 2019]. They treat all time-specific corpora $C_a, C_b, \ldots, C_n$ as one corpus $C$ and learn word representations on the full corpus. However, they first replace each target word $w \in C_t$ with a time-specific token $w_t$. This way, they learn one single space that contains a vector for each target-time pair $w_t$, which may be compared directly without the need for mapping different spaces to each other.

In Statistical Machine Translation the comparison of semantic spaces has occurred in order to perform unsupervised learning of bilingual lexicons. Artetxe et al. [Artetxe et al., 2018] developed a cross-lingual word embedding mapping in order to align two languages without the need of parallel corpora. They propose a self-learning method based on the observation that, given the similarity matrix of all words in the vocabulary, each word has a different distribution of similarity values. Their assumption is that two equivalent words in different languages should have similar distributions.

Minnema and Herbelot [Minnema and Herbelot, 2019] used various metrics to measure the similarity between a linguistic embedding space and a brain image embeddings space. Besides testing pairwise and rank correlation between vectors for the same word from the two spaces, their metrics included Nearest Neighbour structure of the two spaces and Representational Similarity Analysis (Pearson correlation between their respective similarity matrices). The latter is somewhat related to the method of Artetxe et al. [Artetxe et al., 2018], as they also initialise with correlation matrices of the two vector spaces – which, in their case, correspond to linguistic spaces of two different languages. Dubossarsky et al. [Dubossarsky et al., 2019] also performed nearest neighbour analysis in the Lexical Semantic Change context.

As regards measurements, such as nearest neighbour, in high dimensional vector spaces, one has to take the threat of the curse of dimensionality into account. Dinu et al. [Dinu et al., 2015] showed that nearest neighbour suffers from the hubness problem. This phenomenon is known to occur as an effect of the curse of dimensionality, and causes a few points (known as hubs) to be nearest neighbours of many other points [Radovanović et al., 2010]. This is a problem because these hub vectors tend to be near a high proportion of items, pushing their correct labels (e.g., words which are semantically similar) down the neighbour list.

Concept based interpretability analysis using clustering is a new area in ML, which is related to our approach in spirit. Ghorbani et al. [Ghorbani et al., 2019] introduced post-training analysis of computer vision models using clustering of image segments. Clustering and visualisations have been previously used for multi-modal embedding analysis in [Gupta et al., 2019].

As a *qualitative / quantitative structural analysis* we will employ standard clusterization metrics, which is most related to [Minnema and Herbelot, 2019] and cluster visualisations somewhat similar to [Gupta et al., 2019]. Unlike previous work, we will zoom even further into our embeddings and perform a thorough qualitative cluster analysis along with visualisations to discover *model-concepts* (introduced in Section 2.1.2), and analyse a new structured embedding type (Section 2.5). This will be complemented with an

[7] `https://github.com/Garrafao/TemporalReferencing`

information-theoretical analysis framework, which we introduce in the following sections.

### 2.7.4 Information Theory Background

The third pillar of our semantic analysis seeks the answer to the question: **How much** semantic embeddings $E_m$ of different modalities **differ**? We reformulate this questions as follows: *How much extra information we gain if we combine two modalities?* We could also phrase it this way: *How much less confused a model $\mathcal{S}_{m_1}$ gets after combining it with another $\mathcal{S}_{m_2}$?* We reach out for the help of information-theory to formalise our question. We start with a review of the basics then formulate our approach.

The standard unit of information in computer science is the bit. The most widespread way of measuring information is the Shannon entropy [MacKay, 2003], introduced by Claude Shannon in 1948 [Shannon, 2001]. In information theory, the entropy of a random variable is the average level of "information", "surprise", or "uncertainty" inherent in the variable's possible outcomes. Shannon was searching for an information measure with the following conditions: Let $p$ be a probability of an event, then

1. $H(p)$ is monotonically decreasing in $p$.
2. $H(p) \geq 0$: information is a non-negative quantity.
3. $H(1) = 0$: events that always occur do not communicate information.
4. $H(p_1, p_2) = H(p_1) + H(p_2)$: the information learned from independent events is the sum of the information learned from each event.

Shannon discovered that the only suitable choice of $H$, where $X = x_1, \dots x_n$ is a random variable and $P(X)$ is a probability mass function, is:

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log_b P(x_i) \tag{2.7}$$

where $b$ is the base of the logarithm used ($b = 2$ measures information in bit).

We rely on the concept of Mutual Information, which is intimately linked to entropy. It is also known as Information Gain and measures the information that two random variables, $X$ and $Y$ share: It measures how much knowing one of these variables reduces uncertainty about the other. Using the entropy it is defined by:

$$I(X, Y) = H(X) - H(X|Y) \tag{2.8}$$

where $H(X|Y)$ is the conditional entropy [MacKay, 2003].

Let $(X, Y)$ be a pair of continuous random variables with values over the space $\mathcal{X} \times \mathcal{Y}$. If their joint distribution is $P_{X,Y}$ and the marginal distributions are $P_X$ and $P_Y$, the mutual information is defined as

$$I(X, Y) = \int_{\mathcal{X}} \int_{\mathcal{Y}} P_{X,Y}(x, y) \log \left( \frac{P_{X,Y}(x, y)}{P_X(x) P_Y(y)} \right) \tag{2.9}$$

It follows that

$$I(X, Y) = D_{KL}(P_{X,Y} || P_X \otimes P_Y) \tag{2.10}$$

where $D_{KL}$ is the Kullback–Leibler divergence:

$$D_{KL}(P||Q) = \int_{\mathbb{R}^d} dP \log \frac{dP}{dQ} \tag{2.11}$$

If $p(x)$ and $q(x)$ are densities then

$$D_{KL}(p||q) = \int_{\mathbb{R}^d} p(x) \log \frac{p(x)}{q(x)} dx. \tag{2.12}$$

### 2.7.5 Proposal for Measuring Independence of Embeddings

The phenomena that human multi-modal sensory information fusion happens in a statistically optimal fashion has been studied in Cognitive Psychology [Ernst and Banks, 2002]. Ernst et al. found that humans combine visual and haptic information in proportion to their uni-modal variance. Interestingly, not directly analogous, but somewhat related is the finding of Kiela et al. [Kiela et al., 2014] for multi-modal (visuolinguistic) word embeddings. They filtered visual input for words based on the corresponding images' *dispersion*, which measures the average pairwise distances of image vectors for a word. They found that filtering out "noisy" images improved on the multi-modal representation. This does not necessarily mean that one should ignore all new conflicting information, but highlights that it is possible to add more data to the system and having worse performance. In this thesis, we are pursuing a deeper understanding of the exact circumstances under which visual information enhances meaning representations and when it does not, by learning more about the relationship between semantic spaces of different modalities.

The informativeness of new data has been studied in learning pure linguistic embeddings as well. Kabbach et al. [Kabbach et al., 2019] developed a method to train word embeddings on a smaller corpus with maximal information gain, after pretraining them on a large corpus. Their model is designed to simulate new word acquisition by an adult speaker who already masters a substantial vocabulary. Their system uses a pretrained CBOW as this "background knowledge" which they then use to train an SGNS on a much smaller data in a way that the context is maximally informative (has minimal entropy) given the previous knowledge.

To our knowledge we are first to propose measuring the *independence* of different modalities by estimating the Mutual Information between their embedding spaces. In order to do so, we treat each embedding space $E_{m_i}$ as a vector space, representing samples from a multivariate random distribution. By estimating the mutual information we can compare which embedding pairs differ more from each other. We would like to know, whether the perspective of $E_S$ or $E_V$ is "farther" from $E_L$; which one is "more independent"? Let us reformulate our three assumptions on partial observers from Section 2.7 in the information-theoretical framework. Let $m_i, m_j, m_k$ be modalities, where $i, j, k$ are distinct, then:

1. **Common ground**: Neither two embeddings $E_{m_i}, E_{m_j}$ are completely independent, as they have all learnt some pattern related to the same hidden concepts in a language: $I(E_{m_i}, E_{m_j}) \neq 0$.
2. **Perspectives**: They are not completely correlated: $I(E_{m_i}, E_{m_j})$ is not maximal.
3. **Imperfect knowledge**: None of them is an oracle, they do not predict the evaluation data perfectly $P(D|\mathcal{S}_{m_i}) \neq 1$.

Thus if the efficiency (Section 2.7.2) of $E_{m_j}$ and $E_{m_k}$ are similar, and

$$I(E_{m_i}, E_{m_j}) > I(E_{m_i}, E_{m_k}) \tag{2.13}$$

then we hypothesise that there is a combination method with which, combining $E_{m_i}$ with $E_{m_k}$ is more efficient than using $E_{m_i} + E_{m_j}$, as they convey more complementary information which can be combined. The question of how this combination is realised depends on all the parameters in $\mathcal{S}_{m_i}$ and $\mathcal{S}_{m_k}$ and the combination method itself. In this work we explore mid-fusion combination as it allows us to study the information from different modalities separately as well as combined, and it makes it straightforward to compare individual embeddings.

### 2.7.6 A Utility Based Model of Embedding Independence

In this section we introduce a toy model based on probabilistic games, which serves as a theoretical backing for Mutual Information minimisation. As it is just a toy model, it is not a fundamental part of the framework of this thesis. However, it provides an interesting perspective on learning multi-modal semantic representations based on information-theory, which could be generalised in the future.

Before we create our own model of multi-modal fusion, we introduce Kelly's framework of betting in a game through a noisy binary channel [Kelly jr, 1956], [Cover and Thomas, 2012, p. 162].

**Rate of Growth** Let us consider a repeatable game, where in each round a gambler can bet some amount of their wealth (including the whole) on either of two outcomes. After each round the gambler wins the double of their bet if they guessed right, and loses it otherwise. If $p$ is the probability of error and $q$ is the probability of a right guess, how much would they bet? Let $V_0$ be the starting capital, $V_N$ is the capital after $N$ bets. If they bet their entire capital each time, this in fact, would maximise the expected value of their capital $\langle V_N \rangle$, which in this case would be given by

$$\langle V_N \rangle = (2q)^N V_0 \tag{2.14}$$

This would be little comfort, however, since if they continued indefinitely ($N \to \infty$), they would be broke with probability one. Let us, instead, assume that the gambler bets a fraction $l$ of their capital each time. Then

$$V_N = (1+l)^W (1-l)^L V_0 \tag{2.15}$$

where $W$ and $L$ are the number of wins and losses in the $N$ bets. Then the doubling factor or rate of growth of the gambler's capital $G$ is[8]

$$\begin{aligned} G &= \lim_{N\to\infty} \left[ \frac{W}{N} \log(1+l) + \frac{L}{N} \log(1-l) \right] \\ &= q \log(1+l) + p \log(1-l) \text{ with probability one} \end{aligned} \tag{2.16}$$

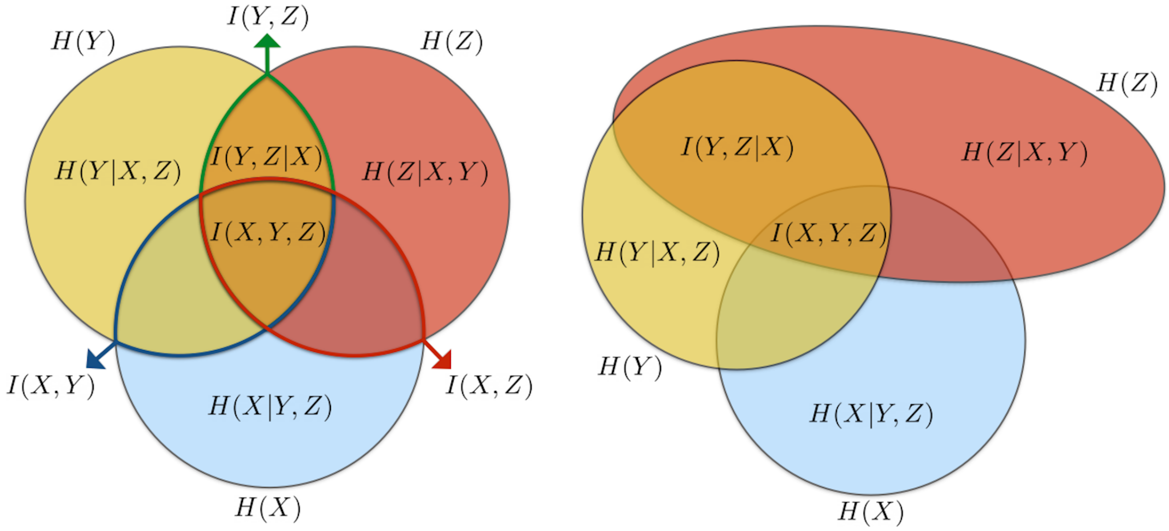
We want to maximise this gain. Since it is logarithmic, we can take its derivative at the point of zero, and we get

$$G_{max} = 1 + p \log p + q \log q = 1 - H(X) \tag{2.17}$$

which is 1 minus the Shannon entropy, where $X$ is a random variable which can take the value of $p$ or $q$. The model has been generalised by Kelly for more than two outcomes in [Kelly jr, 1956].

**Gain of Multi-Modal Fusion** Now, let us imagine learning concepts in a language from data as such a game. Figure 2.4 illustrates the model the following way. Winning corresponds to learning a semantic model of target concepts $T$ which highly correlates with human semantic judgement. The noisy channel corresponds to the dataset $D$ via which our models can learn embedding representations $E_{m_1}, E_{m_2}$. In this game we are interested in maximising our gain, by combining two modalities the most efficient way. Let $X$ denote a perfect "ground-truth" semantic representation, which maximally correlates

[8] Here, log denotes $\log_2$.



(a) Low inter-modality dependence, *independently* from $X$: $I(Y,Z|X)$.

(b) High inter-modality dependence, *independently* from $X$: $I(Y,Z|X)$.

Figure 2.5: Three Random Variables $X, Y$ and $Z$. Here $X$ represents a "ground-truth" variable, a perfect semantic representation. $Y$ and $Z$ are two random variables, corresponding to embeddings of two modalities $E_{m_1}, E_{m_2}$.

with human judgement on our task. For the sake of readability let $Y := E_{m_1}, Z := E_{m_2}$. Then the maximal rate of growths for each model and for the ground-truth are:

$$\begin{aligned} G_Y &= 1 - H(X|Y) \\ G_Z &= 1 - H(X|Z) \\ G_0 &= 1 - H(X) \end{aligned} \tag{2.18}$$

The rate of growth or gain with the combination of $Y$ and $Z$ is

$$G_{YZ} = 1 - H(X|Y,Z) \tag{2.19}$$

We are interested in maximising the rate of growth after we combine the information from both modalities. Let us maximise the following difference:

$$\Delta G_{YZ} = G_{YZ} - G_0 \tag{2.20}$$

Thus, the following theorem holds:

**Theorem 1.** *$\Delta G_{YZ} = \Delta G_Y + \Delta G_Z - I(X,Y,Z)$.*

*Proof.* From Equations 2.18, 2.19 and 2.20:

$$\begin{aligned} \Delta G_Y &= H(X) - H(X|Y) = I(X,Y) \\ \Delta G_Z &= H(X) - H(X|Z) = I(X,Z) \\ \Delta G_{YZ} &= H(X) - H(X|Y,Z) \end{aligned} \tag{2.21}$$

(This is also Kelly's result for the general case, with more than two outcomes to bet on, with independent transmitted symbols with fair odds. Fair odds means that the odds paid on the occurrence of the $s$'th transmitted symbol is proportional to the probability that the transmitted symbol is the $s$'th one [Kelly jr, 1956].)

Furthermore, we apply the I-Diagram in Figure 2.5, a geometrical representation of the relationship among the information measures. It is analogous to the Venn Diagram in set theory, which makes several information-theoretical proofs easier [Yeung, 1991].

Therefore,

$$\Delta G_{YZ} = \Delta G_Y + \Delta G_Z - I(X,Y,Z) \text{ (see Figure 2.5a)} \tag{2.22}$$

□

Furthermore, the following inequality holds:

**Theorem 2.** $I(X,Y,Z) \leq I(Y,Z)$. *Mutual Information is an upper bound to minimise, in order to maximise the rate of growth after multi-modal fusion.*

*Proof.* $\Delta G_Y$ and $\Delta G_Z$ are given because the individual embeddings have already been trained. Therefore, from Theorem 1 it follows that we need to *minimise* $I(X,Y,Z)$ in order to maximise $\Delta G_{YZ}$.

Furthermore, using the I-Diagram in Figure 2.5a:

$$I(X,Y,Z) \leq I(Y,Z) \tag{2.23}$$

□

Let us notice that if $I(Y,Z)$ is high, the reason might be independent from $X$. Therefore, $I(X,Y,Z)$ can be small while $I(Y,Z|X)$ is high, as it is illustrated in Figure 2.5b. In practice, however, this would mean that two embeddings $E_{m_1}, E_{m_2}$ are correlated in some way which is irrelevant to learning semantic representations. For example two corpora may have similar number of documents, or written in the same verse etc. If this spurious correlation is too high, minimising $I(Y,Z)$ may not be a good approximation. Our investigation of the datasets we use did not reveal such spurious correlations. Therefore, we treat $I(X,Y,Z)$ being very close to $I(Y,Z)$.

Maximising the gain from multi-modal embedding combination serves as a framework for analysing efficient multi-modal fusion. An exciting future extension of this model would be to generalise it further for odds which are not fair, based on [Kelly jr, 1956]. In Section 3.2.4 we will introduce empirical MI estimation methods, which we will apply in experiments presented in Chapter 6.

## 2.8 Summary: Comprehensive and Interpretable Word Semantic Analysis

In this chapter we reviewed the philosophical and theoretical background of word semantics and motivated researching distributional word semantic models as a proxy for statistical analysis of concepts. After reviewing the literature on textual distributional semantics, visual embeddings and multi-modal approaches, we proposed a new type of embedding in between linguistic and visual modalities, based on small data. Furthermore, we introduced a general framework and formalism for investigating multi-modal semantic embedding models. Lastly, we presented a framework for treating modalities as partial observers of meaning based on information-theory.

To tackle inconsistencies and the lack of systematic comparisons in multi-modal literature, we proposed extending the analyses of previous work with an **interpretable** analysis framework of three pillars:

1. *Performance testing*: **Black-Box** testing – How representations of different modalities **perform** on evaluation tasks? We extended previous work with:

    (a) *Comprehensive analysis* of models across data sources, machine learning models and modalities.
    (b) *New Modality* based on small data and in between low level visual information and high level linguistic, symbolic data.
    (c) *Efficiency* analysis controlling for data size, data distribution and model size.

2. *Qualitative / Quantitative structural analysis*: **Transparency** testing – **How** representations of different modalities **differ**? An analysis of model-concept structures captured by modalities.

3. *Independence analysis*: **Transparency** testing: **How much** representations **differ**?

We postulated that none of these pillars are alone sufficient for an interpretable semantic embedding analysis, however, when combined, they can offer a fuller picture on what and how our models capture. We need a (1.) *comprehensive performance testing* combined with *efficiency* metrics as a goal. Within this context we can make transparency analysis involving (2.) zooming into the *structural properties* of embeddings and (3.) quantifying the optimal *information gain* from multi-modal fusion.

Within this proof-of-concept framework we showcase that structured small data can be an efficient alternative to expensive big data and models, when the resources are scarce.

# Chapter 3

# Methodology of Data Selection and Proposal for Interpretable Evaluation

In this chapter we introduce the training and evaluation datasets which form the basis of this study. Understanding how each training data and evaluation sets have been created is crucial for interpreting the results. Using the notation from Section 2.6, Section 3.1 describes image, text and structured corpora $D_V$, $D_L$, $D_S$ used as training data. Section 3.2 gives an overview of the evaluation data and methodology. Finally, we summarise the roadmap of the scheme of our three pillar analysis in Section 3.3.

## 3.1 Training Data Matters

One of the main objectives of this thesis is to analyse the data sources that are being used during model training. Recalling our notation of semantic embedding models of modality $m$ (with output embedding $E_m$):

$$\mathcal{S}_m = \langle T, O_m, D_m, X_m, A_m, d_m \rangle \tag{3.1}$$

The dataset $D_m$ comprising observable items and target elements is an essential parameter. Analysing them, therefore, is the basis for all three contributions. In our I. **comprehensive analysis** we aim to overcome the often inconsistent or hard to compare results in previous work. Introducing a new mapping $X_S$ from a structured data source as well as analysing the properties of the data is in the centre of our study of a II. **new type of semantic embedding model $\mathcal{S}_S$**. Lastly, getting more familiar with the training data is imperative if we want to create III. transparent and **interpretable** semantic models.

Section 3.1.1 gives a summary of the properties of image datasets $D_V$ which are used throughout the thesis for visual models $\mathcal{S}_V$. Section 3.1.2 introduces text corpora $D_L$ for linguistic semantic embedding models $\mathcal{S}_L$. Let us highlight that Visual Genome is included in both categories, since it is used both as an image dataset $D_V$ as well as a structured text corpus $D_S$ of $\mathcal{S}_S$ after extracting annotation from its structured annotations.

### 3.1.1 Image Data

This section introduces the details of processing image data and image datasets which deliver observable context $O_V$ in visual semantic embedding models $\mathcal{S}_V$.

| | **Google** | **Bing** | **Flickr** | **ImageNet** | **Visual Genome** |
|---|---|---|---|---|---|
| Type | Search engine | Search engine | Photo sharing | Image database | Image database |
| Annotation | Automatic | Automatic | Human | Human | Human |
| Coverage | Unlimited | Unlimited | Unlimited | Limited | Limited |
| Sorted | Yes | Yes | Yes | No | No |
| Tag specificity | Unknown | Unknown | Loose | Specific | Dense |

Table 3.1: Sources of image data.

**Processing Image Data** We used MMFeat toolkit[1] (based on Caffe[2]) to obtain image representations for three different convolutional network architectures: AlexNet [Krizhevsky et al., 2012], GoogLeNet [Szegedy et al., 2015] and VGGNet [Simonyan and Zisserman, 2014], and our own toolkit, EmbEval[3] for ResNet [He et al., 2016] and AlexNet based on Pytorch-torchvision[4]. Image representations are turned into an overall word-level visual representation by taking the mean of the relevant image representations. All four networks are trained to maximize the multinomial logistic regression objective using mini-batch gradient descent with momentum:

$$-\sum_{i=1}^{D}\sum_{k=1}^{K}\mathbf{1}\{y^{(i)}=k\}\log\frac{\exp(\theta^{(k)\top}x^{(i)})}{\sum_{j=1}^{K}\exp(\theta^{(j)\top}x^{(i)})} \tag{3.2}$$

where $\mathbf{1}\{\cdot\}$ is the indicator function, $x^{(i)}$ and $y^{(i)}$ are the input and output, respectively. $D$ is the number of training examples and $K$ is the number of classes. The networks are trained on the ImageNet classification task and we transfer layers from the pre-trained network.

As we use CNN models pre-trained on ImageNet the other datasets do not serve as CNN training data. However, all CNN networks work as a mapping from our $O_V$ images to a vector space. The vector representations are obtained by running a feed-forward step in the network and extracting the last layer as the representation of the image. We use the last fully connected layer from AlexNet and VGGNet (both 4096 dimensional vectors), and the last pooling layer from GoogLeNet (1024 dimensions) and ResNet (512 dimension). We have multiple image results for a word, hence this method has to be a vector aggregation, such as element-wise maximum, mean or median (studied in Section 4.1). The learning algorithm and the aggregation method together constitutes the mapping function $A_V$ in $\mathcal{S}_V$.

**Image Datasets** Previous systematic studies of parameters for text-based distributional methods have found that the source corpus has a large impact on representational quality [Sahlgren and Lenci, 2016, Kiela and Clark, 2014]. The same is likely to hold in the case of visual representations. Various sources of image data have been used in multi-modal semantics, but there have not been many comparisons: [Bergsma and Goebel, 2011] compare Google and Flickr, and [Kiela and Bottou, 2014] compare ImageNet [Deng et al., 2009] and the ESP Game dataset [von Ahn and Dabbish, 2004], but most works use

[1] https://github.com/douwekiela/mmfeat

[2] https://caffe.berkeleyvision.org/

[3] https://github.com/anitavero/embeval

[4] https://pytorch.org/docs/stable/torchvision/index.html

a single data source. In this work, one of our objectives is to asses the quality of various sources of image data $D_V$.

We selected the presented datasets because they are all standard in Computer Vision or NLP while they all differ in at least one of the following properties:

- *Type*: search engines; photo sharing social networks or hand crafted image datasets.
- *Annotation*: Automatic by an algorithm or annotated by humans.
- *Coverage*: Unlimited – crowd sourced on the internet or a prepared dataset of limited size.
- *Sorted*: Whether there is a relevance score assigned to each image that indicates how descriptive it is of a word (e.g., search engine order).
- *Tag specificity*: Whether the annotation of images are: specific of objects / scenes in the image; loose – related to the image on a higher semantic level or from a personal annotator's angle; dense – detailed labels of objects and relationships *within* an image.

Table 3.1 provides an overview of the data sources. Descriptions of each dataset follow:

**Google Images** Google's image search[5] results have been found to be comparable to hand-crafted image datasets [Fergus et al., 2005].

**Bing Images** An alternative image search engine is Bing Images[6]. It uses different underlying technology from Google Images, but offers the same functionality as an image search engine.

**Flickr** Although [Bergsma and Goebel, 2011] have found that Google Images works better in one experiment, the photo sharing service Flickr[7] is an interesting data source because its images are tagged by human annotators.

**ImageNet** ImageNet [Deng et al., 2009] is a large ontology of images developed for a variety of Computer Vision applications. It serves as a benchmarking standard for various image processing and Computer Vision tasks. ImageNet is constructed along the same hierarchical structure as WordNet [Miller, 1995], by attaching images to the corresponding synset (synonym set).

**Visual Genome** Visual Genome [Krishna et al., 2016] is a human annotated dataset which contains images with bounding box annotations around objects and relations among many other types of information, such as scene and region descriptions, object attributes, semantic relationships between image regions and objects, and Visual Question Answering (VQA) pairs. The objects, attributes, relationships, and noun phrases in region descriptions, and VQA pairs are also canonicalised to WordNet [Miller, 1995] synsets.

All of the dataset properties can be relevant, however, it is not immediately obvious whether any of the above sources are superior over the other. While search engines provide full data coverage for virtually any vocabularies of various languages, they fall

[5] `https://images.google.com/`

[6] `https://www.bing.com/images`

[7] `https://www.flickr.com`

behind in tag specificity, as the search word is in an associative relationship with the images, not a hand-crafted label. Search engines and Flickr all come with a relevance order, which can be useful for image based meaning representations. However, in case of search engines we rely too much on black-box algorithms and automatic annotation. Hand-crafted datasets, while certainly fall behind in size and thus coverage, contain more carefully collected human annotation, which are usually more specific and detailed. In both ImageNet and VisualGenome, annotations are aligned with WordNet, which is a standard knowledge base.

Figure 3.1 contains image samples from all datasets which serve as observable contexts $O_V$, that are mapped to vectors by a feed-forward step in a CNN. All networks are pre-trained on ImageNet, thus our models do not differ in this regard. While there is less difference for the more specific concept of *elephant*, results for *animal* are more diverse across sources. Visual Genome (Figure 3.1a) includes several bounding boxes with dense annotations, whereas the others are ordered by relevance. Flickr tends to include more personal photos, such as pets in Figure 3.1d. Google and Bing have more versatile results (Figure 3.1b, 3.1c). In order to see clearer how each properties affect model performance, we propose measuring the effect of image source choice and discuss its effectiveness regarding the costs of dataset creation.

### 3.1.2 Text Corpora

Linguistic modes $\mathcal{S}_L$ are naturally trained on text corpora $D_L$. Structured embeddings $\mathcal{S}_S$ are also trained on text, however, the main difference from traditional text corpora is that these are ordered in a specific structure instead of free text, e.g., a graph of expressions, hence the distinct notation $D_S$. We used different versions of Wikipedia and Common Crawl datasets as $D_L$ training data. $D_S$ consists of Visual Genome Scene Graphs. All these are described in the following.

**Wikipedia** Wikipedia[8] is a widely used corpus in NLP applications. It is a crowd-sourced encyclopaedia, which covers various common sense and scientific concepts. Its topic structure has been directly exploited in Explicit Semantic Analysis [Gabrilovich et al., 2007]. It has been used as a general training corpus for its wide topic coverage, and long history of crowd-sourced quality control. In this work we use versions, trained on 2013 and 2020 Wikipedia dumps, as baseline models.

**FastText** In Section 4.3 we use more recent pretrained word embeddings from the FastText framework[9]. These models use the traditional CBOW model, with versions extended with subword information [Mikolov et al., 2018]. The following training datasets were used:

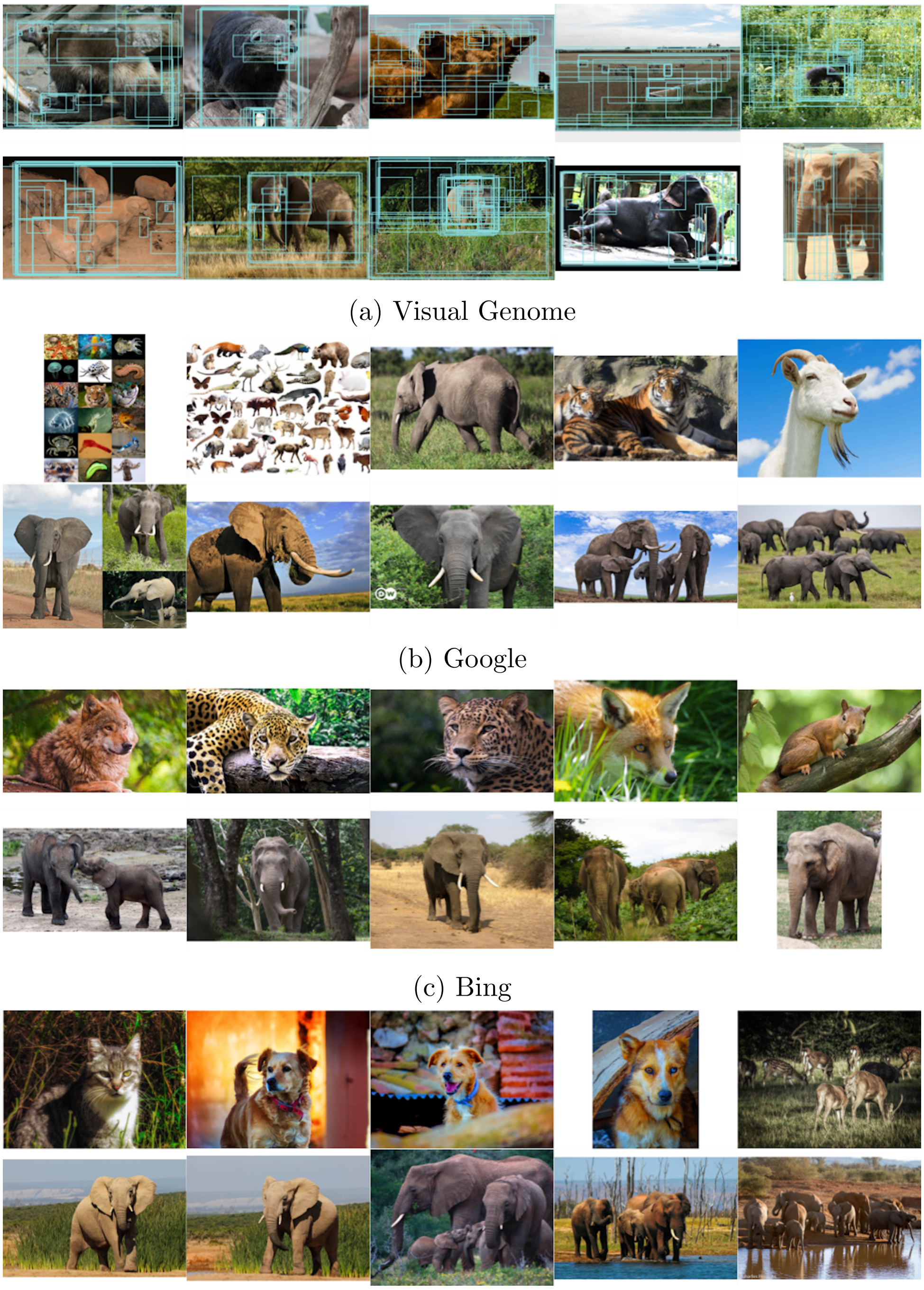
1. *wiki-news-300d-1M*: 1 million word vectors trained on Wikipedia 2017, UMBC webbase corpus and statmt.org news dataset (16B tokens).

2. *wiki-news-300d-1M-subword*: 1 million word vectors trained with subword infomation on Wikipedia 2017, UMBC webbase corpus and statmt.org news dataset (16B tokens).

3. *crawl-300d-2M*: 2 million word vectors trained on Common Crawl (600B tokens).

[8] `https://www.wikipedia.org/`

[9] `https://fasttext.cc/docs/en/english-vectors.html`



(a) Visual Genome

(b) Google

(c) Bing

(d) Flickr

Figure 3.1: Example images for *animal* and *elephant* from the various data sources used as observable contexts $O_V$. While there is less difference for the more specific concept of *elephant*, results for *animal* are more diverse across sources. Visual Genome includes several bounding boxes with dense annotations, whereas the others are ordered by relevance.

4. *crawl-300d-2M-subword*: 2 million word vectors trained with subword information on Common Crawl (600B tokens).

**Visual Genome Scene Graph** In this work we do not only use Visual Genome [Krishna et al., 2016] as an image dataset, but we exploit its dense and structured human annotation as well, as a text corpus. The Visual Genome dataset contains complete set of descriptions and QAs for each image based on multiple image regions, and a formalized representation of the components of an image. It consists of seven main components:

| | |
|---|---|
| Total region descriptions | 4,297,502 |
| Total image object instances | 1,366,673 |
| Unique image objects | 75,729 |
| Total object-object relationship instances | 1,531,448 |
| Unique relationships | 40,480 |
| Total attribute-object instances | 1,670,182 |
| Unique attributes | 40,513 |
| Total Scene Graphs | 108,249 |
| Total Region Graphs | 3,788,715 |
| Total Question Answers | 1,773,258 |

Table 3.2: Visual Genome annotation statistics.

region descriptions, objects, attributes, relationships, region graphs, scene graphs, and question answer pairs. Figure 3.2 shows examples of each component for one image. Although, it falls behind the above mentioned text corpora in terms of size, its highly structured nature can convey semantic information in itself. This dataset is special in terms of its "modality". It includes dense textual annotation of image objects and scenes which people normally do not write about. Therefore, even though the annotation consists of character series, it conveys some high level visual common-sense knowledge. Besides its relevance for research, this type of annotation collection methodology could benefit data acquisition of low-resource languages, where there is no abundance (or there is an absence) of corpora. Applying tools where speakers can point out visually grounded meaning of their language could be a highly efficient way of documenting and analysing these languages. Moreover, automatic Scene Graph Generation algorithms [Xu et al., 2020] can further boost the efficiency of such methods. Some statistics[10] on the size of different annotation types are summarized in Table 3.2. Preliminary studies on a new embedding type based on this dataset is discussed in Section 4.2. The model is thoroughly studied in Section 4.3 and in Chapters 5 and 6.

We decided to use Wikipedia and corpora from the FastText system, because they are all standard in the literature and are also easily and openly available. While we used pretrained models in our studies, we also trained our own SGNS model on various subsets of Wikipedia for quantity and distribution control experiments. Experimenting with even bigger datasets would be a potential improvement. However, given our resources and the number of experiments planned, this was a sensible data size limit. Visual Genome is a unique data source for its structured annotations. We chose it to investigate the potentials of such a dataset for multi-modal semantics.

## 3.2 From Intrinsic Evaluation to Interpretable Model Anatomy

In this section we discuss the used evaluation datasets, metrics and analysis methodology, which we applied to implement our three-pillar transparent testing of multi-modal

[10] `https://visualgenome.org/data_analysis/statistics`

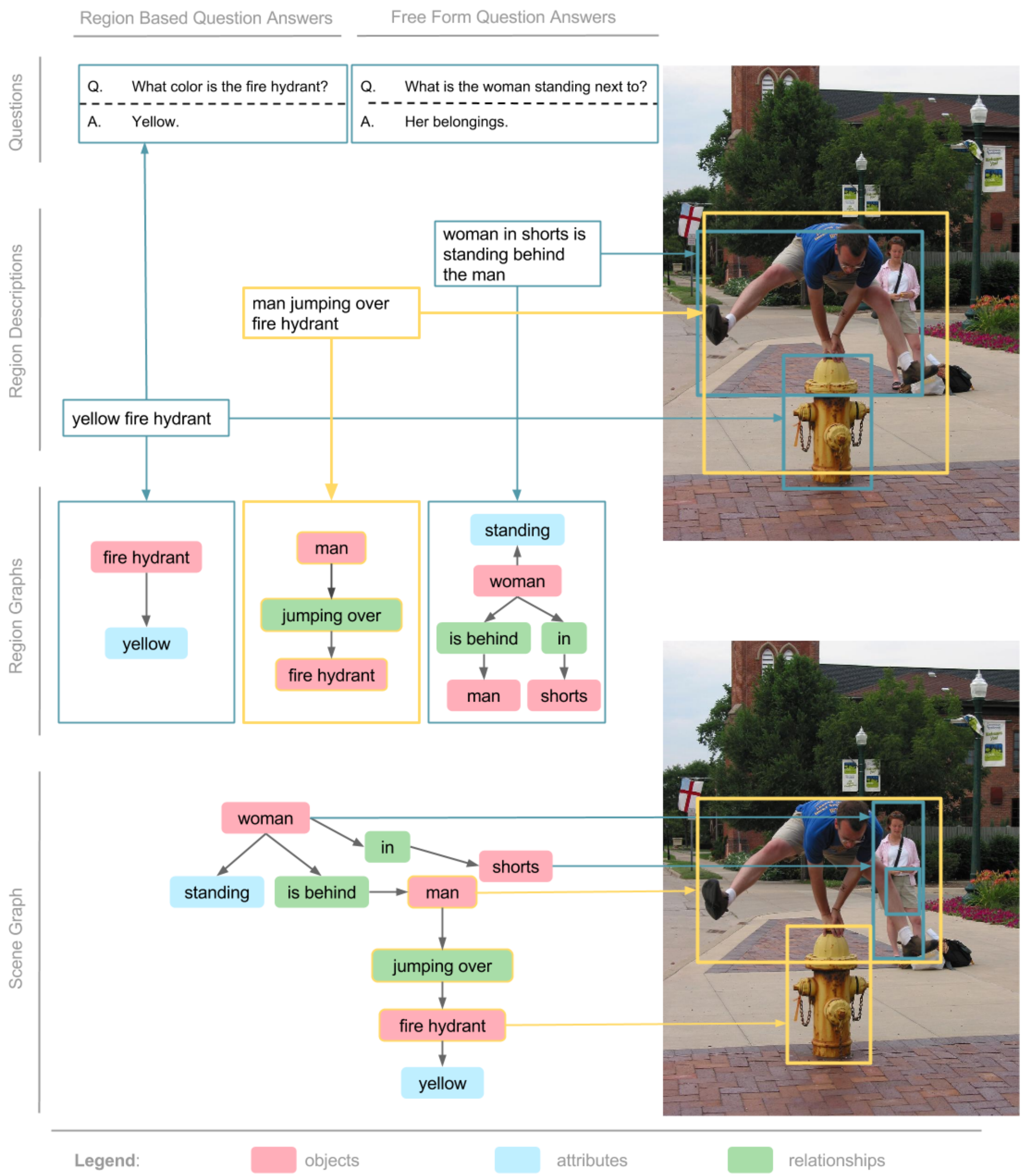


Figure 3.2: A representation of the Visual Genome dataset. Each image contains region descriptions that describe a localized portion of the image. There are two types of question answer pairs (QAs): free form QAs and region-based QAs. Each region is converted to a region graph representation of objects, attributes, and pairwise relationships. Finally, each of these region graphs are combined to form a scene graph with all the objects grounded to the image. [Krishna et al., 2016]

embeddings, laid out in Section 2.7.2. Section 3.2.1 describes the tools for 1 *Performance testing.* Section 3.2.2 describes analysis on brain data as embedding analysis. Section 3.2.3 introduces cluster analysis as 2 *Qualitative / Quantitative structural analysis.* Finally Section 3.2.4. introduces empirical Mutual Information estimation methods for 3 *Independence analysis.*

### 3.2.1 Behavioural Tasks

Most multi-modal word embedding work evaluate on semantic similarity and relatedness tasks in the hope of gathering information about the intrinsic behaviour of abstract semantic representations. However, the ambiguous notion of similarity and the low inter-annotator agreement make it difficult to draw robust conclusions on the differences between models [Batchkarov et al., 2016]. As a first black-box step, we will also evaluate on these standard datasets. Unlike previous work, however, we first aim to create an extensive study of comparing several semantic models $\mathcal{S}_m$ with varying parameters of $T, O_m, D_m, X_m, A_m, d_m, E_m$. Then we gradually move towards more in-depth transparency analysis.

We briefly describe the standard evaluation datasets and metrics we use in our experiments:

**MEN** The MEN data set [Bruni et al., 2014] consists of 3,000 word pairs, randomly selected from words that occur at least 700 times in the freely available ukWaC and Wackypedia corpora combined and at least 50 times (as tags) in the opensourced subset of the ESP game dataset.[11] Pairs were sampled so that they represent a balanced range of relatedness levels according to a text-based semantic score. Each pair was randomly matched with a comparison pair and rated in this setting (as either more or less related than the comparison point) by an annotator on Amazon Mechanical Turk. This binary comparison task is both more natural for an individual annotator, and also permits seamless integration of the supervision from many annotators. The downside is that this way, there is no well-defined inter-subject agreement. In total, each pair was rated against 50 comparison pairs, thus obtaining a final score on a 50-point scale, although the Turkers' choices were binary.

**SimLex-999** SimLex-999 [Hill et al., 2015] is a dataset structurally similar to MEN, including 999 word pairs for intrinsic semantic evaluation. Its objective is, however, to measure how well models capture *similarity*, rather than *relatedness* or association. The scores in SimLex-999 therefore differ from other well-known evaluation datasets such as MEN. For example, "coast" and "shore" would have high score in both MEN and SimLex. On the other hand, "cloth" and "closet" would have low score in SimLex but high score in MEN, since they have different materials, function etc., even though they are very much related. This task is challenging for computational models to replicate because, in order to perform well, they must learn to capture similarity independently of relatedness/association. These two relationships between words show up in different contextual features. Similarity is inferred from similar co-occurrences with other words. Similarity or relatedness is then captured by the type of co-occurrence / window size [Kilgarriff and Yallop, 2000]. In addition SimLex includes concreteness Part-Of-Speech and association scores from the University of South Florida (USF) Free Association Norms [Nelson et al., 2004].

[11] `https://staff.fnwi.uva.nl/e.bruni/MEN`

**SimVerb-3500** SimVerb-3500 [Gerz et al., 2016] is an evaluation resource that provides human ratings for the similarity of 3,500 verb pairs. It covers all normed verb types from the USF Free Association database, providing at least three examples for every VerbNet [Schuler, 2005] class. Verb pairs are rated on a scale 0-10, for example: “to reply” / “to respond” - 9.79; “to participate” / “to join” - 5.64; “to stay” / “to leave” - 0.17. We included this dataset in Section 4.2, where predicate - object relationships are in focus, to test how it affects verb representations in particular.

**Evaluation metric** Model performance is assessed through the Spearman $\rho_s$ rank correlation between the embedding similarity scores for a given pair of words, together with human judgements in each evaluation datasets. Pearson correlation has also been considered, however, humans find it much harder to attach a numerical score to a pairwise comparison like “cat”−“dog”, rather than having to judge whether that comparison is more similar than “cat”−“television”. Furthermore, Pearson correlation coefficient should also be avoided because even if humans give numerical scores as similarity ratings, these are unlikely to be normally distributed.

Embedding similarity scores are computed using the cosine distance of the two word vectors, $\vec{w1}, \vec{w2}$ of a word pair, $w1, w2$.

$$\text{Cosine}(\vec{w1}, \vec{w2}) = \frac{\vec{w1} \cdot \vec{w2}}{\|\vec{w1}\|\|\vec{w2}\|} \tag{3.3}$$

$$= \frac{\vec{w1} \cdot \vec{w2}}{\sqrt{\sum_i w1_i^2}\sqrt{\sum_i w2_i^2}} \tag{3.4}$$

The dot product in the numerator is calculating numerical overlap between the word vectors, and dividing by the respective lengths provides a length normalisation which leads to the cosine of the angle between the vectors. Normalisation is important because we would not want two word vectors to score highly for similarity simply because those words were frequent in the corpus. The cosine measure is commonly used in studies of distributional semantics, however, we could use any other vector space metric [Clark, 2015]. It is difficult to reach a conclusion from the literature regarding which similarity measure is best; we use cosine distance here because it has become standard in NLP. Future work could involve revisiting these standard metrics because they may behave differently depending on the task and the source/modality of training data.

### 3.2.2 Brain Imaging as Embedding Analysis

Evaluating on brain imaging data has been introduced as NLP evaluation tasks on various occasions [Mitchell et al., 2008, Anderson et al., 2016] (Section 2.1.2). In some cases visually grounded models have been included in the evaluation [Davis et al., 2019, Anderson et al., 2017, Bulat et al., 2017]. The measured impact of multi-modal information, however, varies across studies, thus in this work we included a broader analysis on these tasks as well. We aim to use correlation studies with brain data as a type of black-box analysis, which is substantially different from behavioural tasks and as such can shed new light on differences between our Semantic Embedding models of different modalities. The findings in cognitive neuro-science (Section 2.1.2) on multi-modal human brain activities while performing semantic tasks, further motivates us to include brain data in our studies.

We evaluate on two brain image datasets which were collected while participants viewed 60 concrete nouns with line drawings [Mitchell et al., 2008, Sudre et al., 2012]. One dataset was collected using fMRI (Functional Magnetic Resonance Imaging) and one with MEG (Magnetoencephalography). Each dataset has 9 participants, but the participant sets are disjoint, thus there are 18 unique participants in total. Though the stimuli is shared across the two experiments, MEG and fMRI are very different recording modalities and thus the data are not redundant [Xu et al., 2016].

**fMRI dataset** fMRI measures the change in blood oxygen levels in the brain, which varies according to the amount of work being done by a particular brain area. In this fMRI dataset collected by Mitchell et al. [Mitchell et al., 2008] participants were presented with line drawings and noun labels of 60 concrete nouns from 12 semantic categories: animals, body parts, buildings, building parts, clothing, furniture, insects, kitchen items, tools, vegetables, vehicles and man-made objects. The experimental task was to think about the properties of the noun concept they were shown - the set of 60 concepts was presented in a random order six times to each participant. Each concept was presented for 3 seconds, with seven second gaps between presentations.

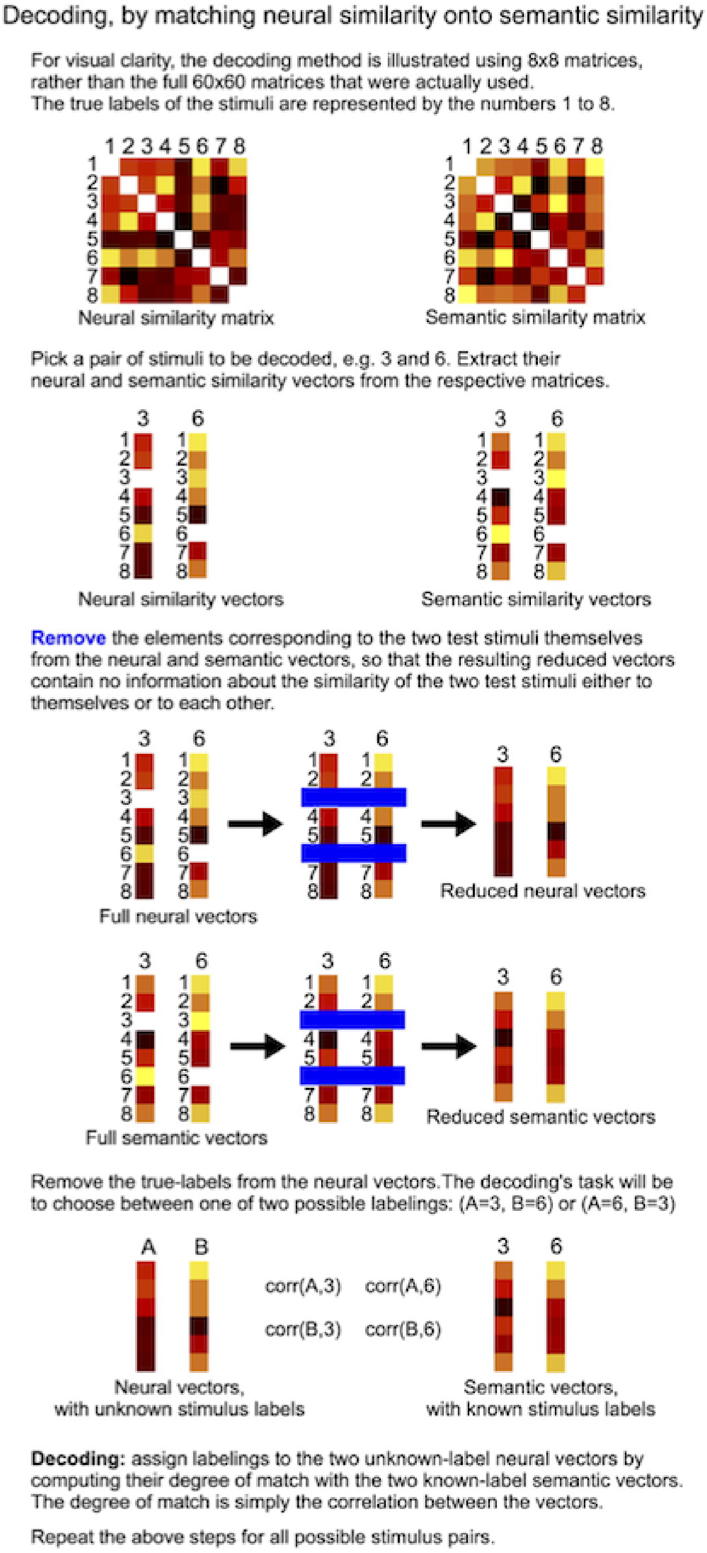
**MEG dataset** This experiment involved the same task as the previous one but using MEG machine, a large helmet with 306 sensors that measure aspects of the magnetic fields at different locations in the brain. A MEG brain image is the time signals recorded from each of these sensors. Each of the words was presented 20 times (in random order) for a total of 1200 brain images.

Both brain image data have been preprocessed by the BrainBench Test Suit [Xu et al., 2016]. They used "partialling out" process in order to remove low level activity attributable to visual properties from the brain images. They used the methodology from Mitchell et al. to select the most stable brain image features for each of the 18 participants. The stability metric assigns a high score to features that show strong self-correlation over presentations of the same word.

**Two vs. two test** To evaluate on brain data we need to compare representation similarities from brain imaging vectors and meaning representation vectors. This type of evaluation if fundamentally different from the behavioural tasks, as we do not have human similarity score labels for word pairs. We use leave-two-out cross validation, the testing methodology from Mitchell et al. which has become standard for brain imaging evaluation of semantic embeddings. Our implementation is based on BrainBench with modifications so we can perform analysis on individual participants. The evaluation starts from two similarity matrices, a neural and a brain similarity matrix. Columns of this matrices are called similarity codes. Similarity codes $(\vec{s_i}, \vec{s_j})$ and brain activity similarity codes $(\vec{a_i}, \vec{a_j})$ are selected for two nouns. Elements $i.$ and $j.$ from each of the similarity codes are removed, as these entries correspond to the nouns being tested. Figure 3.3 visualises an example of the decoding procedure. Decoding is successful if the sum of Pearson correlations for the correct pairings is greater than the sum of Pearson correlations for the incorrect pairings, resulting in decoding accuracy of 1 for this pair and 0 otherwise. Thus, the expected chance-level decoding accuracy is 50%.

### 3.2.3 How do Models Conceptualise? – Cluster Analysis

As introduced in Section 2.7.3 the second pillar (2) of our analysis is a transparent investigation of the concepts our embedding spaces $E_L, E_V, E_S$ capture. We are interested in how much these model-concepts differ from each other to understand under what circumstances each modalities can complement each other. As mentioned before this *qualitative /*



Decoding, by matching neural similarity onto semantic similarity

For visual clarity, the decoding method is illustrated using 8x8 matrices, rather than the full 60x60 matrices that were actually used. The true labels of the stimuli are represented by the numbers 1 to 8.

Pick a pair of stimuli to be decoded, e.g. 3 and 6. Extract their neural and semantic similarity vectors from the respective matrices.

**Remove** the elements corresponding to the two test stimuli themselves from the neural and semantic vectors, so that the resulting reduced vectors contain no information about the similarity of the two test stimuli either to themselves or to each other.

Remove the true-labels from the neural vectors.The decoding's task will be to choose between one of two possible labelings: (A=3, B=6) or (A=6, B=3)

**Decoding:** assign labelings to the two unknown-label neural vectors by computing their degree of match with the two known-label semantic vectors. The degree of match is simply the correlation between the vectors.

Repeat the above steps for all possible stimulus pairs.

Figure 3.3: Visualisation of leave-two-out cross validation from [Anderson et al., 2016].

*quantitative structural analysis* is meant to be used in the context of previous performance analyses and the third pillar of 3 *independence* analysis, we will detail in Section 3.2.4.

By model-concept, here, we mean some similarity metric based clusters in the embedding spaces, which do not necessarily correspond to the meaning of one word, but rather some higher level or different structure. As a straightforward implementation, we chose to use standard clustering algorithms and metrics, to compare our different embeddings.

In order to grasp how the concept structure of our embedding spaces differ from each other we first searched for ways to quantify their cluster structure. We do not know the ground truth labels of our clusters or even the number of clusters each embedding spaces should be broken into. Therefore, we experiment with three standard clusterization metrics which are designed for the case when a ground truth labelling is not available. Furthermore, we report results for a range of number of clusters.

In Chapter 6 we present the design, implementation and result of our transparency studies. Section 6.2 includes qualitative and quantitative cluster analysis. In Section 6.2.2 we compare our embeddings' cluster structures and visualise the learnt clusterings. In Section 6.2.3 we present supervised visualisations of the embedding spaces alongside an automatic label generation method and compare the results against the clusterization metric scores. As an effective visualisation we use the T-SNE algorithm [Maaten and Hinton, 2008, Wattenberg et al., 2016].

Clustering and T-SNE have been previously used for multi-modal embedding analysis e.g., [Gupta et al., 2019]. In Section 6.2 we report qualitative analyses by investigating the elements of the clusters, as well as reporting further quantitative cluster structure comparison analyses. One of our clustering analyses is based on the pre-defined cluster labels of [Gupta et al., 2019]. They also use Visual Genome, otherwise, their work is fundamentally different from ours as they use different models, they do not exploit the Visual Genome graph structure and evaluate on downstream tasks.

In the following we present all the standard algorithms and metrics used for the clustering studies.

### Clustering Methods and Metrics

We ran the K-means [MacQueen et al., 1967] clusterization algorithm on all three embeddings to see if it can reveal more about the underlying structure of the spaces. We used the k-means++ initialization scheme [Arthur and Vassilvitskii, 2006], which has been implemented in the Scikit-learn package[12]. This initializes the centroids to be (generally) distant from each other, leading to probably better results than random initialization. As a control for consistency of clustering we also present results using Agglomerative Clustering[13]. To measure the rate of clusterization, when the labels are not known, we used three standard metrics implemented in the Scikit-learn package[14]. One drawback of these metrics is that they are generally higher for convex clusters than other concepts of clusters. However, convexity is not always given. They respond poorly to elongated clusters, or manifolds with irregular shapes.

1. **Davies–Bouldin Index** can be calculated by the following formula:

$$\mathrm{DB} = \frac{1}{K}\sum_{i=1}^{K}\max_{j\neq i}\left(\frac{\sigma_i+\sigma_j}{d(c_i,c_j)}\right) \tag{3.5}$$

[12] https://scikit-learn.org/stable/modules/clustering.html#k-means

[13] https://scikit-learn.org/stable/modules/clustering.html#hierarchical-clustering

[14] https://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation

where $\sigma_x$ is the average distance of all elements from the cluster centroid in cluster $C_x$. $d(c_i, c_j)$ is the distance between centroids $c_i, c_j$. Since clusters with low intra-cluster distances (high intra-cluster similarity) and high inter-cluster distances (low inter-cluster similarity) will have a low Davies–Bouldin index, the smaller this number is the better the clusterization is considered to be.

The computation of Davies-Bouldin is simpler than that of Silhouette scores. The index is solely based on quantities and features inherent to the dataset as its computation only uses point-wise distances.

2. **Calinski-Harabasz Index** – also known as the Variance Ratio Criterion – can be used to evaluate the model, where a higher Calinski-Harabasz score relates to a model with better defined clusters. The index is the ratio of the sum of between-clusters dispersion and of inter-cluster dispersion for all clusters (where dispersion is defined as the sum of distances squared):

$$\mathrm{CH} = \frac{\mathrm{tr}(B_K)}{\mathrm{tr}(W_K)} \times \frac{N-K}{K-1} \tag{3.6}$$

where $\mathrm{tr}(B_K)$ is the trace of the between group dispersion matrix and $\mathrm{tr}(W_K)$ is the trace of the within-cluster dispersion matrix defined by:

$$W_K = \sum_k \sum_{e \in C_k} (e - c_k)(e - c_k)^T \tag{3.7}$$

$$B_K = \sum_k (c_k - c_E)(c_k - c_E)^T \tag{3.8}$$

with $c_E$ being the centroid of $E$.

The score is higher when clusters are dense and well separated, which relates to a standard concept of a cluster.

3. **Silhouette Coefficient** value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation).

For each data point $e_i$ we define:

$$a(e_i) = \frac{1}{|C_i| - 1} \sum_{j \in C, i \neq j} d(e_i, e_j) \tag{3.9}$$

$$b(e_i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(e_i, e_j) \tag{3.10}$$

We now define a silhouette (value) of one data point $e_i$:

$$\mathrm{S}(e_i) = \frac{b(e_i) - a(e_i)}{\max\{a(e_i), b(e_i)\}}, \text{ if } |C_i| > 1 \tag{3.11}$$

Silhouette Coefficient is also higher when clusters are dense and well separated.

### 3.2.4 Information Gain from Modalities

The third pillar of our analysis is the second transparency study, which aims to uncover **how much** representations **differ**? We formulated it as an *independence analysis* (Pillar 3) of our embeddings $E_L, E_V, E_S$ as multivariate random variables in Section 2.7.5. Applying equation 2.13 to the three modalities (including the same three assumptions), we aim to measure whether

$$I(E_L, E_V) > I(E_L, E_S) \tag{3.12}$$

in which case we hypothesise that there is a combination method with which, combining $E_L$ with $E_S$ is more efficient than using $E_L + E_V$, as they convey more complementary information which can be combined. The experiment design and the results are reported in Section 6.3. We need to estimate the empirical Mutual Information of our vector spaces from data, which is a hard problem. In the following we introduce standard methods and tools we used for this purpose.

#### Empirical Mutual Information Estimation

Since Mutual Information is a special case of divergence (such as $D_{KL}$ in Equation 2.10), divergence estimators can be employed to estimate it. To recall the definition of $D_{KL}$ (Equation 2.12): if $p(x)$ and $q(x)$ are densities then

$$D_{KL}(p||q) = \int_{\mathbb{R}^d} p(x) \log \frac{p(x)}{q(x)} dx. \tag{3.13}$$

The estimators then approximate Equation 2.10:

$$I(X, Y) = D_{KL}(P_{X,Y} || P_X \otimes P_Y) \tag{3.14}$$

In our application, $P_{X,Y}$ is a sample from a multi-modal embedding created by mid-fusion, whereas the marginals are the uni-modal embeddings. To estimate the densities $p(x)$ and $q(x)$, the traditional approach is to use histograms with equally sized bins [Wang et al., 2005]. However, the computational complexity of such methods is exponential in $d$ and the estimation accuracy deteriorates quickly as the dimension increases. Hence, a more robust way of estimating multidimensional Mutual Information is using k-Nearest Neighbor distances ($I_{KNN}$) which bypasses the difficulties associated with partitioning in a high-dimensional space [Wang et al., 2009]. This method estimates a density by computing the average frequency of each point's KNNs in the Euclidean ball centred around the point. This provides a consistent estimate of $D_{KL}(p||q)$. In practice these methods become unreliable in a high-dimensional space due to the sparsity of the data objects.

To overcome this, another approach is to introduce non-linearity using a kernel, when calculating the distances. In this work we use a kernel method called the Hilbert-Schmidt Independence Criterion (HSIC) algorithm [Gretton et al., 2005], because it has been shown to work in practical applications [Jitkrittum et al., 2017].

Consider a reproducing kernel Hilbert space $\mathcal{F}$ of functions from $\mathcal{X}$ to $\mathbb{R}$. To each point $X \in \mathcal{X}$, there corresponds an element $\phi(X) \in \mathcal{F}$ such that $\langle \phi(X), \phi(X') \rangle_{\mathcal{F}} = k(X, X')$, where $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a unique positive definite kernel.

Then the HSIC estimate is given by the following:

$$I_{HSIC}(X, Y) = \|C_{X,Y}\|_{HS}, \tag{3.15}$$

where $\|.\|_{HS}$ is the Hilbert-Schmidt Norm. $C_{X,Y}$ is a cross-covariance operator between $X$ and $Y$:

$$C_{X,Y} = \mathbb{E}_{X,Y}([k_1(\cdot, X) - \mu_X] \otimes [k_2(\cdot, Y) - \mu_Y]) \tag{3.16}$$

where $\mu_X = \mathbb{E}_X[k_1(\cdot, X)]$ and $\mu_Y = \mathbb{E}_Y[k_2(\cdot, Y)]$ are the mean embeddings of $X$ and $Y$ respectively to a Reproducing Kernel Hilbert Space. $k_1$ and $k_2$ are kernels on $X$ and $Y$ respectively. For more details on the theoretical background see [Gretton et al., 2005].

We apply an open source Python implementation of the above algorithms from the Information Theoretical Estimators Toolbox[15] [Szabó, 2014].[16]

## 3.3 Analysis Scheme

Figure 3.4 represents a roadmap of our three pillar analysis.[17] On the top: Pillar 1 *Performance testing*: broad comparison across data sources, ML models and modalities, which will be presented in Chapter 4. Based on this $\mathcal{S}_L, \mathcal{S}_V, \mathcal{S}_S$ are narrowed down to a particular combination of model and data source. In Chapter 5 we change our focus on more in-depth analysis of fewer models based on the findings in the previous blanket studies. Here, we restrict ourselves to behavioural tests, but we inspect our models in a more fine grained fashion, regarding size and distribution ranges. Following, in the middle: Pillar 2 of *structural cluster analysis* to discover embedding concepts. At the bottom: Pillar 3: *Independence analysis* of embeddings. Chapter 6 includes the two parts of our transparency analysis. Here, we will focus on the structure of each embedding types $E_L$, $E_S$ and $E_V$. Lastly, we measure the information gain $E_S$ and $E_V$ entail when combined with $E_L$.

Narrowing the umbrella studies down to a few model, data and modality combinations is another layer on top of the three pillar analysis framework. However, this layer is not necessary for our proposed evaluation methodology. Performing costly large scale studies with numerous current models would become shortly obsolete. Our aim is rather to provide a general framework with proof-of-concept studies, which can be applied to various models in the future.

[15] `https://bitbucket.org/szzoli/ite`

[16] We would also like to thank Zoltán Szabó for his counsel on the theoretical background.

[17] Icons made by Freepik, Smashicons, Good Ware, Eucalyp and Becris from `https://flaticon.com`/authors/<author name>. Voronoi diagrams were generated using `http://alexbeutel.com/webgl/voronoi.html`.

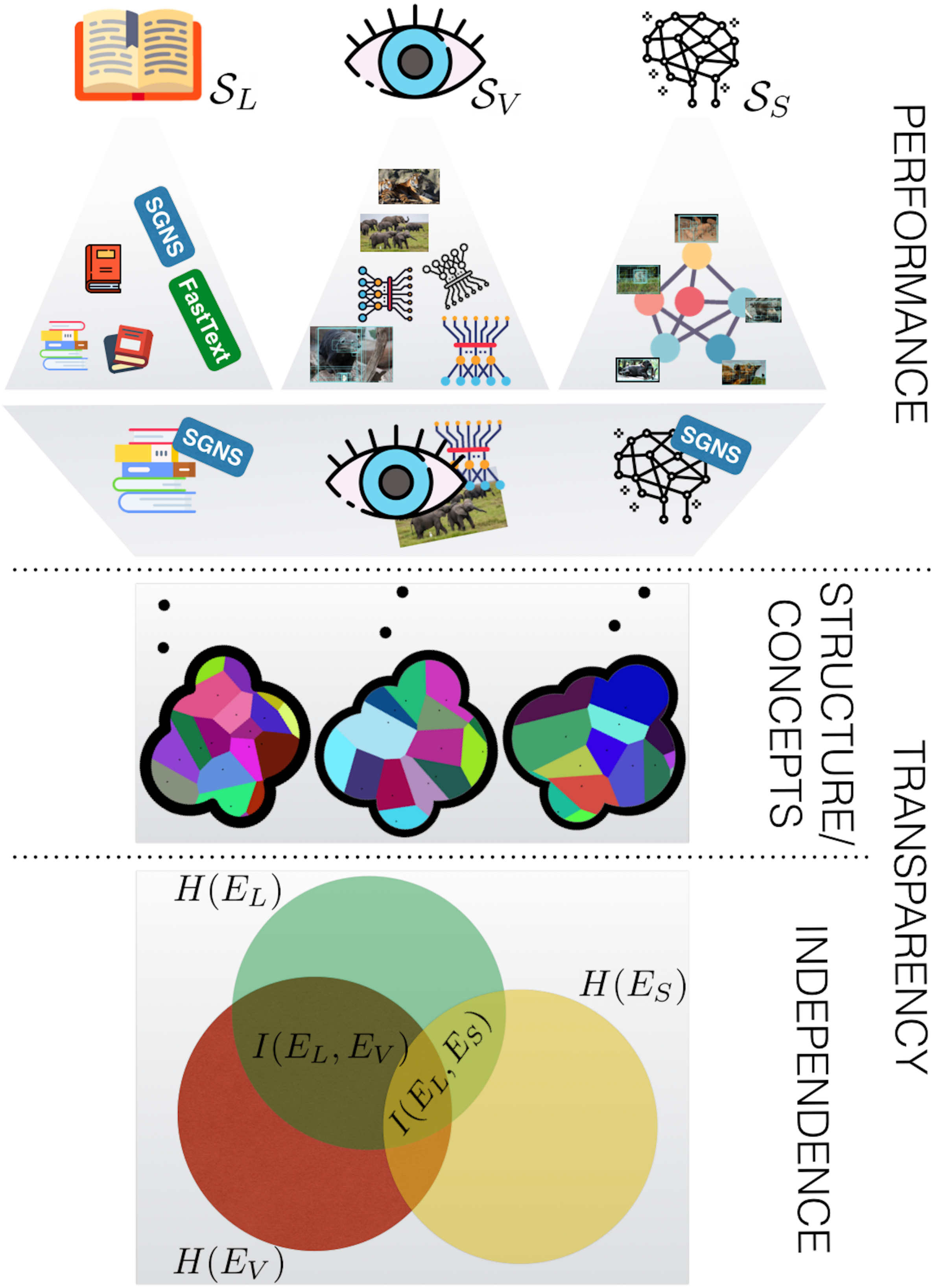


Figure 3.4: Roadmap of analyses. On the top: Pillar 1 *Performance testing*: broad comparison across data sources, ML models and modalities. Based on this $\mathcal{S}_L, \mathcal{S}_V, \mathcal{S}_S$ are narrowed down to a particular combination of model and data source. Following, in the middle: Pillar 2 of *structural cluster analysis* to discover embedding concepts. At the bottom: Pillar 3: *Independence analysis* of embeddings.

# Chapter 4

# Impact of Visual Information in Semantics

This chapter covers experiments which form an implementation of pillar 1 *Performance testing* (Figure 3.4 on top). We cover experiments towards a *comprehensive analysis* of models across data sources, machine learning models and modalities. We introduce the implementation of a *new structured hybrid modality* based on small data and in between low level visual information and high level linguistic, symbolic data. We use evaluations which we refer to as *black-box* testing, for looking at only performance numbers. However, by performing a broad study we aim to offer a more comprehensive analysis of multi-modal studies than in previous work.

The experiments are designed to addresses our research Questions 1, 2 and 3, laid out in Chapter 1.1. To recap and frame them in our Semantic Embedding model framework (Section 2.6):

1. How does the source of images $D_V$ affect the performance of multi-modal semantic representations?

2. Does the number of images have an impact on performance? – Variability of the visual *extraction function* $X_V$.

3. Do previous findings on complementary visual information scale to different types and sizes of linguistic corpora? – Variability of observable context data $O_L, O_V, O_S$ and introducing a new extraction function for structured data $X_S$.

In Section 4.1 we present a systematic study of the performance of state-of-the-art image data sources and CNN architectures, and measure the impact of image quantity (Questions 1 and 2). In Section 4.2 we introduce a new embedding type based on a visually structured, textual data source, the Visual Genome Scene Graphs [Krishna et al., 2016], and show preliminary studies on its performance for "sanity-check". In Section 4.3 we present a broader analysis involving the models from the previous sections, extended with new ones. We tackle Question 3 by comparing several data sources of different sizes and modalities. Section 4.4 involves a study on how pretrained word embedding initialisation affects sequence model performance on textual entailment.

## 4.1 Comparing Visual Models and Data Sources for Semantics

This section focuses on the analysis of $E_L + E_V$ type multi-modal word embeddings with mid-fusion and various Convolutional Neural Network based $E_V$ visual representations. The study explores the following questions regarding semantic similarity and relatedness tasks:

1. How important is the source of images $D_V$? Is there a difference between search engines and manually annotated data sources?

2. How should we aggregate the image representations for a search key into one visual representation? – Post-processing part of the visual *mapping function* $A_V$.

3. Does the number of images obtained for each search key matter? – Variability of the visual *extraction function* $X_V$.

4. Does the choice of the CNN architecture have an impact on the performance of visual and multi-modal models? – ML algorithm part of the visual *mapping function* $A_V$.

To address the first question, we decided to use different search engines and other existing image datasets. For that purpose, we extended Douwe Kiela's MMFeat toolkit[1] with an API for the Flickr search engine. Later on we continued working on a joint project addressing the above questions in multi-modal distributional word semantics. The results have been published in an EMNLP long paper [Kiela et al., 2016].[2] In this project, we systematically compared deep visual representation learning techniques, experimenting with three well-known network architectures, AlexNet, GoogLeNet and VGGNet (see Section 2.3.1). In addition, we explored the various data sources (described in Section 3.1.1) that can be used for retrieving relevant images, showing that images from search engines perform as well as, or better than, those from manually crafted resources such as ImageNet. Furthermore, we explored the optimal number of images and the multi-lingual applicability of multi-modal semantics.

### 4.1.1 Evaluation

We employ *behavioural evaluation tasks* described in detail in Section 3.2.1. In summary, model performance is assessed through the Spearman $\rho_s$ rank correlation between the system's similarity scores for a given pair of words, together with human judgements. We evaluate on two well-known similarity and relatedness judgement datasets: MEN [Bruni et al., 2014] and SimLex-999 [Hill et al., 2015].

In each experiment, we examine performance of the visual representations compared to text-based representations, as well as performance of the multi-modal representation that fuses the two. In this case, we apply mid-level fusion – a popular technique in multi-modal semantics (described earlier) – concatenating the L2-normalized representations. Linguistic representations are 300-dimensional and are obtained by applying skip-gram with negative sampling to a 2013 dump of Wikipedia. Visual vectors based on AlexNet and VGGNet are both 4096-dimensional, GoogLeNet vectors are of 1024 dimensions. The

[1] `https://github.com/douwekiela/mmfeat`

[2] I implemented the Flickr API and all the data collection, experiments and evaluations presented in this thesis.

normalization step that is performed before applying fusion ensures that both modalities contribute equally to the overall multi-modal representation.

We evaluated the different architectures and data sources using either the mean or elementwise maximum method for aggregating image representations into visual ones ($A_V$ post-processing). However, we found no significant difference between these two methods.

### 4.1.2 Results

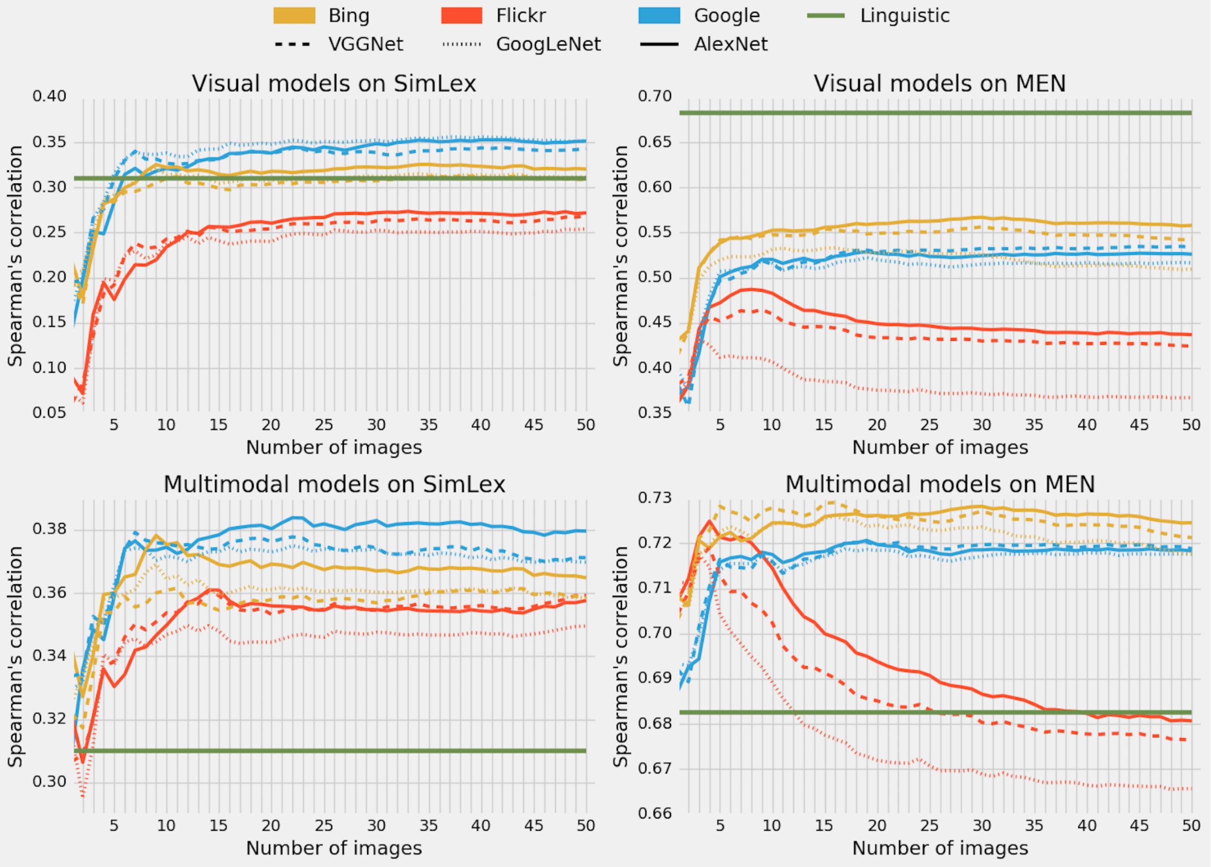


Figure 4.1: The effect of the number of images on representation quality.

We found that multi-modal representation learning yields better performance across the board: for different network architectures, different data sources and different aggregation methods (Figure 4.1).

We examined AlexNet, GoogLeNet and VGGNet, all three winners of the ILSVRC ImageNet classification challenge, and found that they perform very similarly. If efficiency or memory are issues, AlexNet or GoogLeNet are the most suitable architectures. For overall best performance, AlexNet and VGGNet are the best choices.

The choice of data sources has a bigger impact: Google, Bing, Flickr and ImageNet were much better than the ESP Game dataset. Google, Flickr and Bing have the advantage that they have potentially unlimited coverage. Google and Bing are particularly suited to full-coverage experiments, even when these include abstract words [Kiela et al., 2016].

Another question is the number of images we want to use: does performance increase with more images? There is an obvious trade-off here, since downloading and processing images takes time (and may incur financial costs). This experiment only applies to relevance-sorted image search data sources. We found that the number of images has an impact on performance, but that it stabilizes at around 10-20 images, indicating that it is usually not necessary to obtain more than 10 images per word. For Flickr, obtaining more images is detrimental to performance. The effect of the number of images on the performance is shown in Figure 4.1.

### 4.1.3 Conclusion

This work explores some important factors for choosing visual models and data sources for multi-modal semantics. It is important to note that the multi-modal results only apply to the mid-level fusion method of concatenating normalized vectors: although these findings are indicative of performance for other fusion methods, different architectures or data sources may be more suitable for different fusion methods.

Understanding what it is that makes these representations perform so well is another important question. Is it more data or the multi-modal nature of the data which is increasing performance? Building on these preliminary findings, in Section 4.3 we explore a broader range of factors which may shed more light to visual models' behaviour in multi-modal semantics.

## 4.2 Visual Context in the Linguistic Domain

Despite the indisputable success of data driven methods in NLP, humans' ability to generalise after having been exposed to only a small amount of data provides motivation to further explore alternative machine learning methods. An appealing option is to exploit structured prior information combined with multi-modal input. There is a need for more work on applying and automatically acquiring structured prior information that can help us to take a step towards human level and interpretable language generation and understanding.

The second key contribution (II.) of this thesis is the introduction and analysis of a **new modality** (Section 2.5). The study, presented here, aims is to explore the possibilities for learning semantic word representations based on structured and visually grounded prior information. This way we further explore the types of text corpora we use, expanding on Question 3.

We use the Visual Genome (VG) dataset's scene graphs and bounding boxes as structured training data (introduced in Section 3.1.2). Visual Genome images are annotated with region graph representation of objects, attributes, and pairwise relationships. Each of these region graphs are combined to form a scene graph with all the objects grounded to the image (see Figure 3.2).

The main questions this work aims to examine are the following: What is the information coming from (structured) image data? Is it the high level information of visual scene structure which enhances linguistic information or low level visual features matter as well?

### 4.2.1 Scene Graph Context

We introduce a new Semantic Embedding model $\mathcal{S}_S$. There could be many ways to incorporate structured, visually grounded prior information from VG, such as using graph neural networks [Scarselli et al., 2008] as part of the *mapping function* $A_S$. In this work, we implemented a much simpler method in order to see if a small, fast to train model performs well. Instead of developing a new mapping function, we introduce a new *extraction function* $X_S$, which extracts the relevant context information from the scene graphs then feeds it into a simple shallow-network as $A_S$.

Using the scene graph annotations as a corpus, $X_S$ takes as input the whole scene graph dataset $D_S$ and returns "relevant" context items from $O_S$ to each target element from $T$ – that is it returns a mapping from target/context item pairs to numbers in $\mathbb{N}$, representing a *relevance score* of context pairs: $X_S : T \to (O_S(T) \to \mathbb{N})$. In this case this score is a binary number representing whether a context node $o \in O_S$ is in the graph neighbourhood of the surface representations of $t \in T$. The *relevant context* corresponds to a radius in this graph around an object or predicate node. The radius is the number of steps we take starting from a node in a breadth first search manner. The context words are all the node labels within this sub-graph. Algorithm 1 presents the pseudo code for the Scene Graph Context Generation Algorithm. $G$ denotes the scene graph, *rad* is the radius. It returns a word, context pair list $[< t_1, o_1 >, ..., < t_n, o_n >]$. Each node in $G$ has more word labels or "names" (e.g. *elephant* and *animal* can be names of the same object node). We take all the combinations of the given node names of two nodes, which are in each others context. This operation is denoted by the direct product of the two name lists, ×. E.g., if node {*elephant, animal*} is in the neighbourhood of node with label {*sleep*}, then we generate context pairs of: $[\langle elephant, sleep\rangle, \langle animal, sleep\rangle]$.

In this case the *mapping function* $A_S$, is a Skip-gram algorithm [Mikolov et al., 2013b], which maps from context items to a word embedding space $E_S \in R^{|T| \times d_S}, d_S = 300$. Figure 4.2 shows an example for creating contexts for embeddings from Visual Genome Scene Graphs. The context words (orange) used are up to three links from a target node (black).

**Algorithm 1:** Scene Graph Context Generation Algorithm

```
Input: $G$, $rad$
Result: contexts = $[< t_1, o_1 >, ..., < t_n, o_n >]$
for $node \in G$ do
    context_nodes = breadth_first_traverse($node$, $rad$);
    for $cnode \in context\_nodes$ do
        contexts += $[node.names \times cnode.names]$
    end
end
```

Visual Genome scene graphs have been used for word meaning representations [Kuzmenko and Herbelot, 2019, Herbelot, 2020]. They build a truth theoretic model including predicate / entity pairs before feeding it to a skip-gram model. Our method is more relaxed since we directly process the Scene Graphs into contexts of a given size (radius), without any further restriction based on grammatical information. The results are compared in Section 5.2.4.

This model is linguistic in a sense that it only uses text context in the graph neighbourhood, without grounding it to visual features. However, it still uses visual information implicitly, since the graph represents relationships in visual scenes.

Different versions of the above model are compared to the following baselines:

1. *w2v-wikipedia*: A traditional skip-gram trained on a 2013 dump of Wikipedia.

2. *w2v-descriptions*: A skip-gram model trained on the Visual Genome image descriptions.

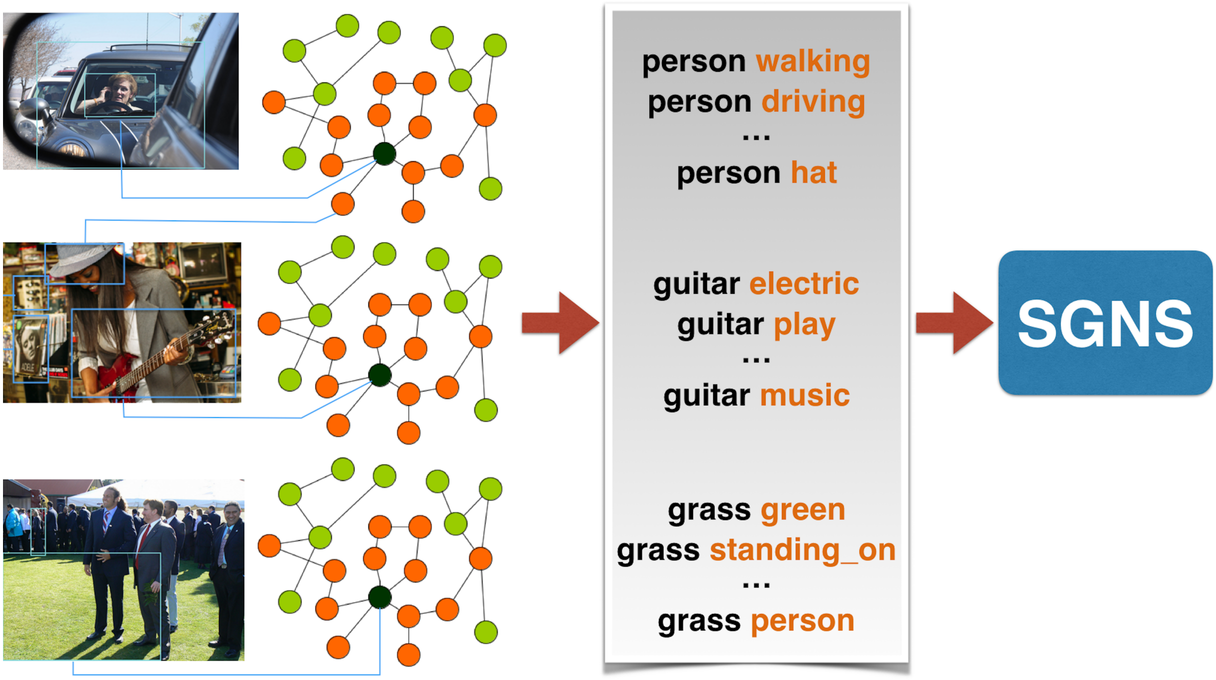


Figure 4.2: Generating contexts for embeddings from Visual Genome Scene Graphs. The context words (orange) used are up to three links from a target node (black). The <target, context word> pairs are then fed to a Skip-gram algorithm. Photos are from `https://visualgenome.org/`

For evaluation we perform the following intrinsic and extrinsic tests:

- *Semantic relatedness/similarity* on the MEN [Bruni et al., 2014] , SimLex [Hill et al., 2015] and SimVerb [Gerz et al., 2016] datasets.
- *Brain data*: Predicting patterns of brain activity associated with the meaning of nouns, making use of two datasets: fMRI (Functional Magnetic Resonance Imaging) [Mitchell et al., 2008] and one with MEG (Magnetoencephalography) [Sudre et al., 2012]. (See in Section 4.3.4)

### 4.2.2 Results

Table 4.1 shows some preliminary results using *Scene Graph context*, that is based on the proximity of words in the Visual Genome Scene Graph. *N* in "*radN*" indicates the number of steps we take around a node in a breadth first search manner. The context words are all the node labels within this radius. Results are shown for both lemmatised and non lemmatised versions of the scene graph corpus. There is no substantial difference after using this preprocessing step (non lemmatised versions even perform slightly better on MEN and SimLex), therefore we do not lemmatise in the following experiments. Using a radius of three, our model outperforms the baseline *w2v-wikipedia* and *w2v-description* baselines on SimLex, but it performs worse on the other datasets.

Further results on behavioural tasks and brain imaging datasets are discussed in Section 4.3.

### 4.2.3 Conclusion

Based on these preliminary results, using structured small-data is a promising area to explore. Despite its size, structured training data can achieve comparable results to

| Lemmatised | Method | MEN | SimLex | SimVerb |
|---|---|---|---|---|
| No | *VG rad3* | 0.433 | **0.274** | 0.008 |
| | *w2v-wikipedia* | **0.680** | 0.238 | **0.149** |
| Yes | *VG rad3* | 0.433 | **0.274** | 0.132 |
| | *w2v-wikipedia* | **0.673** | 0.257 | **0.134** |
| No | *VG rad1* | 0.211 | 0.16 | -0.031 |
| | *w2v-wikipedia* | **0.680** | **0.238** | **0.238** |
| Yes | *VG rad1* | 0.206 | 0.154 | 0.040 |
| | *w2v-wikipedia* | **0.673** | **0.257** | **0.134** |
| Yes | *w2v-description* | 0.427 | 0.289 | 0.127 |

Table 4.1: Pearson correlations of the different versions of the model and the Skip-gram baseline on the MEN, SimLex and SimVerb datasets. $N$ in "*radN*" indicates the number of steps we take around a node in a breadth first search manner. The context words are all the node labels within this radius. Results are shown for both lemmatised and non lemmatised versions of the scene graph corpus.

our big corpus based baseline. Collecting such data by manual labour is expensive, but it is probably worthwhile to explore crowd-sourced, gamified or even (semi–)automatic techniques [Xu et al., 2020] for collecting structured training data. We report on a broader scale analysis of various models including the ones we introduced in this section and in Section 4.1.

## 4.3 Modalities, Sources and Models: a Thorough Analysis

In the previous sections we investigated the impact of visual models and data sources for non-visual evaluations. We compared different convolutional networks for visual embeddings and different image sources. We also experimented with a "small-data" based embedding, using structured information somewhere between the visual and the linguistic domains.

There are two main problems, however, which the multi-modal literature (including the above studies) suffer from:

1. Too small and probably not well formed evaluation datasets [Faruqui et al., 2016].
2. Lack of standardized comparative studies involving many different models.

The first problem is a challenging one due to the cost of data collection. Traditional semantic similarity and relatedness tasks can provide a good starting point to evaluate word semantics, but we certainly need a more thorough analysis if we really want to compare semantic embedding spaces. Recently, the NLP community started evaluating on Brain imaging data as well (see Section 3.2.2), in the hope of learning about the relationship between word embeddings and brain activation of people while thinking of corresponding concepts. These datasets are relatively expensive to create, hence they are not very large. While evaluating on them can provide with interesting insights, we should be cautious when drawing conclusions from these results.

In the following study we use both semantic similarity / relatedness and brain datasets as evaluation. Unlike previous work, however, we try and make a further step towards a more in depth analysis of the results to filter out the potential noise we face in these experiments, coming from different models and small evaluation sets.

As for the second problem, multi-modal models are usually compared to only one linguistic baseline and maybe except for our study in Section 4.1, only one visual source / model combination. Here, we present a broader study involving several different visual and linguistic embeddings in order to get a better picture of the variance we have in performance, tackling our Question 3.

All the experiments have been implemented as part of the EmbEval toolkit (see Section 1.3), including the creation of uni-modal embeddings as well as new mid-fusion techniques (described in Section 4.3.2).

### 4.3.1 Studied Embeddings

In the following we summarise the parameters of the studied Semantic Embedding models, which were described in detail in Chapters 2 and 3.

#### Linguistic Embeddings

To train $\mathcal{S}_L$ models we use pretrained embeddings from the FastText System [Mikolov et al., 2018]. Each model has been trained on different sources $D_L$:

1. *wiki-news-300d-1M*: 1 million word vectors trained on Wikipedia 2017, UMBC webbase corpus and statmt.org news dataset (16B tokens).

2. *wiki-news-300d-1M-subword*: 1 million word vectors trained with subword infomation on Wikipedia 2017, UMBC webbase corpus and statmt.org news dataset (16B tokens).

3. *crawl-300d-2M*: 2 million word vectors trained on Common Crawl (600B tokens).

Furthermore, for comparison with earlier works we also use the same Skip-Gram model, trained on a Wikipedia dump from 2013.

#### Visual Embeddings

Based on the findings in Section 4.1 we test the following datasets and models for $\mathcal{S}_V$:

**Image Source $D_V$** We use Google Images as a source, as it had a stable performance across models, and is widely used. We compare this big data source to visual representations trained on Visual Genome Images. This way we compare a big data source to a smaller, but systematically annotated dataset.

**ML part of $A_V$** For CNN models we use the best and fastest AlexNet model, based on previous findings. Since publishing the results in Section 4.1 a new CNN architecture, called Deep Residual Network (ResNet) [He et al., 2016] appeared, which is the current state-of-the-art in object recognition on images both in terms of classification accuracy and speed. Therefore, in this broader study we included this model as well. We also compare two AlexNet models trained on Visual Genome images *internal* object bounding box images or on the *whole* images, similarly to [Davis et al., 2019].[3]

[3] The training of the models has been done by Christopher Davis. In this paper, I provided supervision with the experiments, help with using MMFeat and helper code for processing Visual Genome.

**Post-processing part of $A_V$** Since our findings in [Kiela et al., 2016] suggest no obvious difference between the two methods, here we only use the mean of image embedding vectors (as opposed to taking the maximum) to create one visual representation for a word.

**Extraction function $X_V$** Furthermore, since after 10-20 images the performance plateaus across the board, in this study we always use 10 images for each word representation.

**Structured Embeddings**

We analyse the $X_S$ version from Section 4.2, when we take three steps around a node in a breadth first search manner.

### 4.3.2 Mid-fusion methods

To create multi-modal embeddings using mid-fusion we applied two methods:

1. *Intersection*: Similarly to previous work a multi-modal embedding is the concatenation of visual and linguistic vectors. Therefore, we only have representations for the intersection of their vocabularies. This is mainly relevant in the case of Visual Genome, where we might not have full coverage (as opposed to Google).

2. *Padding*: In order to have full coverage in every case, in this method if one modality does not cover a word in the vocabulary we just pad the multi-modal vector with as many zeros as the dimensionality of the modality space with the missing vector. This way we have multi-modal embeddings for all the words in the intersection of their vocabularies, and uni-modal vectors, where one of the modalities failed to cover the word.

### 4.3.3 Evaluation Methods

Evaluation of word embeddings on similarity tasks has been shown to be problematic due to 1) the lack of train/development/test splits, 2) the absence of statistical significance, 3) low correlation with downstream performance, 4) the hubness problem and 5) their inability to account for polysemy [Faruqui et al., 2016]. To tackle the first problem we performed three-way cross-validation on MEN and SimLex, leaving out one third of the word pairs randomly. Based on the results – reported in Appendix A – we present correlation figures up to two decimal points. As for the second issue we present a series of detailed evaluation methods in the next chapters, which aim to unearth the reasons behind the behaviour of our models beyond correlation. For correlation scores we report p-values for every correlation score. 4) and 5) are addressed in Chapters 5 and 6. As we discussed in Section 2.1.2, in this work, we view semantic space analysis as a statistical tool for dataset analysis which provides value on its own without downstream applications, therefore 3) is beyond the scope of this thesis.

We cannot directly compare models trained on different data sources, because they have different coverage, but we can look at absolute performance and compare network architectures and modalities. We also present results on the common subset of the evaluation datasets, where all word pairs have images in each of the data sources.

Results on the Brain datasets are analysed averaged over participants for embedding comparison. We present further analyses, where results are averaged over modalities, therefore we can focus more on the variability between participants.

**Concreteness**

Concreteness of words has been studied before in the context of multi-modal semantics and for Brain imaging evaluation. Kiela et al. [Kiela et al., 2014] applied a dispersion metric on the visual domain to filter out words with image results which are noisier than a threshold, based on their metric. They hypothesised that abstract words have higher, whereas concrete words have lower image dispersion. Anderson et al. [Anderson et al., 2017] systematically selected word categories for their Italian dataset based on concreteness.

In this work we developed an automatic concreteness score based on WordNet. The concreteness score of a word is its distance (one minus similarity) from its root hypernyms in the Synset graph.

Since in WordNet we have multiple synsets for one surface form we compare two different techniques to aggregate each sysnets' distances from the root:

1. Taking the *median* of all sysnet's distances for a word.

2. Selecting the synset with the *maximum distance* from the root, so we have the most concrete sense of the word.

Hence, the formula for our WordNet concreteness score is:

$$\text{WNConc(w)} = \text{Agg}_\text{w}[d(s_i, r_i) \mid i \in \{1, \ldots, N_w\}], \tag{4.1}$$

where $\text{Agg}_\text{w}(.)$ is the synset aggregation method, $d(.,.)$ is the WordNet distance, $w$ is a word. $s_i$ are the synsets for $w$ and $r_i$ are the roots of each synset in the WordNet hypernym hierarchy. $N_w$ is the number of synsets for word $w$.

Another question is, how we should combine the concreteness scores for word pairs in the behavioural tasks? We present two methods to do this:

1. Taking the *sum* of the two words' concreteness scores.

2. The *absolute difference* of the two words' concreteness scores.

**Qualitative Analysis on Nouns of the Brain Datasets**

Lastly, we performed qualitative analysis regarding the 60 nouns in the Brain evaluation datasets. Looking at the word concreteness scores did not show any pattern, but this is unsurprising, since this dataset already consists of mainly concrete nouns.

Instead, in this work we included an analysis of the relationship between all studied models in terms of their performance for individual words, averaged over participants (Figures 4.5 and 4.6). Even though this evaluation set is small in terms of vocabulary size, it still can be useful for looking into the nuances we may find regarding individual concepts.

### 4.3.4 Results

The tables in this section show evaluation scores for each task using different versions of evaluation methods. The notation for all tables is the following: Each line corresponds to an embedding. Separator lines divide embeddings by modalities: Linguistic $E_L$, Visual $E_V$, Structured $E_S$ and Multi-modal models $E_L + E_V$ and $E_L + E_S$. *wikinews*, *wikinews_sub* and *crawl* signify FastText vectors trained on the corresponding corpora. *w2v13* is a Skip-Gram model trained on a 2013 Wikipedia dump. Visual Embeddings' names that are trained on Google are in the format of *<image source> <CNN model>*. *VG-internal|external* denotes training on Visual Genome images, either on the internal object images or on the whole images, as it is done in [Davis et al., 2019]. Finally, *VG SceneGraph* stands for the Visual Genome Scene Graph Embeddings from Section 4.2. Multi-modal embeddings have a "+" in their names which separates the two embedding names they are built on.

Red colour indicates best performance, blue means that the multi-modal embedding outperformed the corresponding uni-modal ones. In case of aggregated results for each modality, best performance is signified by bold font.

### Correlations on the Behavioural Tasks

Tables 4.2-4.7 present the standard Spearman's correlation scores of different embeddings on the Semantic Similarity and Relatedness tasks. Tables 4.2, 4.3 and 4.4 present results on the full datasets, Tables 4.6, 4.7 include results on the embeddings' common coverage subsets. Results using *padding* mid-fusion method are shown in Tables 4.2 and 4.3, results applying the *intersection* method are presented in Tables 4.4, 4.5, 4.6, 4.7. Except for the relatively small common subset on SimLex (Table 4.7), *crawl* linguistic embedding outperforms all the other. Multi-modal models outperform uni-modal ones mainly on SimLex, but only in the case of the 2013 Skip-Gram model, which is in line with previous results in Section 4.2. The only multi-modal model outperforming uni-modal ones on MEN is the combination of *w2v13* and *VG Scene Graph*.

Interestingly, using *ResNet* does not provide any performance gain over *AlexNet*, similarly to the other more complicated models in Section 4.1. Both models are fast to run, and *AlexNet* sometimes performs even better, so there is no good reason to use *ResNet* in this task.

Padding multi-modal vectors for bigger coverage does not help, in the case of *w2v13+VG SceneGraph* it even hurts performance. However, this may be due to including more, and perhaps "harder" concept-pairs in the test set than in the smaller intersection set.

The success of combining *w2v13* and *VG Scene Graph* over other visual vectors is interesting. While image embeddings did not help on MEN, this embedding in-between visual and linguistic conveys some complementary information to this linguistic baseline.

Note that in this study we used our EmbEval toolkit for creating multi-modal embeddings, with two different types of mid-fusion methods. In Section 4.1 we used MMFeat, which includes slightly different mid-fusion techniques, therefore, the results are not directly comparable. The main point of this comprehensive study was to reveal patterns across several different sources, architectures and modalities. In the efficiency studies in Chapter 5 and for the transparency analysis in Chapter 6, we also used the EmbEval toolkit.

### Results on Brain Data

Results on the Brain datasets include scores from the *2 vs. 2 test*, described in Section 3.2.2. These experiments have all been run using the *Intersection* mid-fusion technique. This is because padding did not make much of a difference in performance, but it requires much more memory.

In addition to the previous visual models, here, we use the best performing models from [Davis et al., 2019], namely the internal bounding box images, the whole images and the combined image representations of *Visual Genome*. In some cases we created a nested multi-modal model where we combined their initial multi-modal models (denoted by *MM*) with all our linguistic models.

| Modality | Embedding | Spearman | P-value | Coverage |
|---|---|---|---|---|
| $E_L$ | *wikinews* | 0.79 | 0 | 3000 |
| | *wikinews_sub* | 0.80 | 0 | 3000 |
| | *crawl* | **0.85** | 0 | 3000 |
| | *w2v13* | 0.68 | 0 | 3000 |
| $E_V$ | *Google AlexNet* | 0.50 | 0 | 3000 |
| | *Google VGG* | 0.51 | 0 | 3000 |
| | *VG-internal* | 0.37 | 0 | 2784 |
| | *VG-whole* | 0.41 | 0 | 2784 |
| | *Google ResNet-152* | 0.47 | 0 | 3000 |
| $E_S$ | *VG SceneGraph* | 0.42 | 0 | 2574 |
| $E_L + E_V$ | *wikinews+Google AlexNet* | 0.50 | 0 | 3000 |
| | *wikinews+Google VGG* | 0.51 | 0 | 3000 |
| | *wikinews+VG-internal* | 0.36 | 0 | 3000 |
| | *wikinews+VG-whole* | 0.39 | 0 | 3000 |
| | *wikinews+Google ResNet-152* | 0.48 | 0 | 3000 |
| | *wikinews_sub+Google AlexNet* | 0.50 | 0 | 3000 |
| | *wikinews_sub+Google VGG* | 0.51 | 0 | 3000 |
| | *wikinews_sub+VG-internal* | 0.36 | 0 | 3000 |
| | *wikinews_sub+VG-whole* | 0.39 | 0 | 3000 |
| | *wikinews_sub+Google ResNet-152* | 0.47 | 0 | 3000 |
| | *crawl+Google AlexNet* | 0.51 | 0 | 3000 |
| | *crawl+Google VGG* | 0.52 | 0 | 3000 |
| | *crawl+VG-internal* | 0.37 | 0 | 3000 |
| | *crawl+VG-whole* | 0.40 | 0 | 3000 |
| | *crawl+Google ResNet-152* | 0.51 | 0 | 3000 |
| | *w2v13+Google AlexNet* | 0.50 | 0 | 3000 |
| | *w2v13+Google VGG* | 0.51 | 0 | 3000 |
| | *w2v13+VG-internal* | 0.36 | 0 | 3000 |
| | *w2v13+VG-whole* | 0.40 | 0 | 3000 |
| | *w2v13+Google ResNet-152* | 0.48 | 0 | 3000 |
| $E_L + E_S$ | *w2v13+VG SceneGraph* | 0.64 | 0 | 3000 |
| | *crawl+VG SceneGraph* | 0.78 | 0 | 3000 |
| | *wikinews_sub+VG SceneGraph* | 0.37 | 0 | 3000 |
| | *wikinews+VG SceneGraph* | 0.57 | 0 | 3000 |

Table 4.2: Spearman correlation on the MEN dataset. Multi-modal embeddings are created using the Padding technique. The table sections contain linguistic, visual and multi-modal embeddings in this order. Red colour signifies the best performance. Blue would mean that the multi-modal embedding outperformed the corresponding uni-modal ones, which here did not happen.

Tables 4.8, 4.9, 4.10, 4.11 show the scores of each embedding for every participant, and their averages over participants. Multi-modal models on average are clearly bigger winners of this task than of the previous one. In all settings a multi-modal model achieved the highest performance. In all but one case (MEG scores on the common subset vocabularies in Table 4.11) about half of multi-modal models outperformed their corresponding uni-modal ones.

On the full datasets (Tables 4.8 and 4.9) *VG SceneGraph* and *AlexNet* improved the

| Modality | Embedding | Spearman | P-value | Coverage |
|---|---|---|---|---|
| $E_L$ | *wikinews* | 0.45 | 0 | 999 |
| | *wikinews_sub* | 0.44 | 0 | 999 |
| | *crawl* | **0.50** | 0 | 999 |
| | *w2v13* | 0.31 | 0 | 999 |
| $E_V$ | *Google AlexNet* | 0.34 | 0 | 999 |
| | *Google VGG* | 0.34 | 0 | 999 |
| | *VG-internal* | 0.31 | 0 | 103 |
| | *VG-whole* | 0.19 | 0.06 | 103 |
| | *Google ResNet-152* | 0.35 | 0 | 999 |
| $E_S$ | *VG SceneGraph* | 0.26 | 0 | 593 |
| $E_L + E_V$ | *wikinews+Google AlexNet* | 0.34 | 0 | 999 |
| | *wikinews+Google VGG* | 0.34 | 0 | 999 |
| | *wikinews+VG-internal* | 0.31 | 0 | 999 |
| | *wikinews+VG-whole* | 0.31 | 0 | 999 |
| | *wikinews+Google ResNet-152* | 0.35 | 0 | 999 |
| | *wikinews_sub+Google AlexNet* | 0.34 | 0 | 999 |
| | *wikinews_sub+Google VGG* | 0.34 | 0 | 999 |
| | *wikinews_sub+VG-internal* | 0.30 | 0 | 999 |
| | *wikinews_sub+VG-whole* | 0.30 | 0 | 999 |
| | *wikinews_sub+Google ResNet-152* | 0.35 | 0 | 999 |
| | *crawl+Google AlexNet* | 0.34 | 0 | 999 |
| | *crawl+Google VGG* | 0.34 | 0 | 999 |
| | *crawl+VG-internal* | 0.32 | 0 | 999 |
| | *crawl+VG-whole* | 0.32 | 0 | 999 |
| | *crawl+Google ResNet-152* | 0.37 | 0 | 999 |
| | *w2v13+Google AlexNet* | 0.34 | 0 | 999 |
| | *w2v13+Google VGG* | 0.34 | 0 | 999 |
| | *w2v13+VG-internal* | 0.23 | 0 | 999 |
| | *w2v13+VG-whole* | 0.23 | 0 | 999 |
| | *w2v13+Google ResNet-152* | 0.35 | 0 | 999 |
| $E_L + E_S$ | *w2v13+VG SceneGraph* | 0.29 | 0 | 999 |
| | *crawl+VG SceneGraph* | 0.45 | 0 | 999 |
| | *wikinews_sub+VG SceneGraph* | 0.20 | 0 | 999 |
| | *wikinews+VG SceneGraph* | 0.35 | 0 | 999 |

Table 4.3: Spearman correlation on the SimLex dataset. Multi-modal embeddings are created using the Padding technique. The table sections contain linguistic, visual and multi-modal embeddings in this order. Red colour signifies the best performance. Blue would mean that the multi-modal embedding outperformed the corresponding uni-modal ones, which here did not happen.

most on the fMRI and MEG datasets respectively. On the common subset evaluations *ResNet* won the first medal. Interestingly, in all cases the combination with the older *w2v13* linguistic model outperformed the combinations with FastText embeddings.

When it comes to individual participants we see a substantial variance. Tables 4.12, 4.13, 4.14, 4.15 average performances over modalities for each of them. In all settings except for the common subset of fMRI dataset (Table 4.14), multi-modal models achieve a higher average performance than the uni-modal ones in more than 50% of the cases.

| Modality | Embedding | Spearman | P-value | Coverage |
|---|---|---|---|---|
| $E_L$ | *wikinews* | 0.79 | 0 | 3000 |
| | *wikinews_sub* | 0.80 | 0 | 3000 |
| | *crawl* | **0.85** | 0 | 3000 |
| | *w2v13* | 0.68 | 0 | 3000 |
| $E_V$ | *Google AlexNet* | 0.50 | 0 | 3000 |
| | *Google VGG* | 0.51 | 0 | 3000 |
| | *VG-internal* | 0.37 | 0 | 2784 |
| | *VG-whole* | 0.41 | 0 | 2784 |
| | *Google ResNet-152* | 0.47 | 0 | 3000 |
| $E_S$ | *VG SceneGraph* | 0.42 | 0 | 2574 |
| $E_L + E_V$ | *wikinews+Google AlexNet* | 0.50 | 0 | 3000 |
| | *wikinews+Google VGG* | 0.51 | 0 | 3000 |
| | *wikinews+VG-internal* | 0.38 | 0 | 2784 |
| | *wikinews+VG-whole* | 0.41 | 0 | 2784 |
| | *wikinews+Google ResNet-152* | 0.48 | 0 | 3000 |
| | *wikinews_sub+Google AlexNet* | 0.50 | 0 | 3000 |
| | *wikinews_sub+Google VGG* | 0.51 | 0 | 3000 |
| | *wikinews_sub+VG-internal* | 0.37 | 0 | 2784 |
| | *wikinews_sub+VG-whole* | 0.41 | 0 | 2784 |
| | *wikinews_sub+Google ResNet-152* | 0.47 | 0 | 3000 |
| | *crawl+Google AlexNet* | 0.51 | 0 | 3000 |
| | *crawl+Google VGG* | 0.52 | 0 | 3000 |
| | *crawl+VG-internal* | 0.38 | 0 | 2784 |
| | *crawl+VG-whole* | 0.42 | 0 | 2784 |
| | *crawl+Google ResNet-152* | 0.51 | 0 | 3000 |
| | *w2v13+Google AlexNet* | 0.50 | 0 | 3000 |
| | *w2v13+Google VGG* | 0.51 | 0 | 3000 |
| | *w2v13+VG-internal* | 0.38 | 0 | 2784 |
| | *w2v13+VG-whole* | 0.41 | 0 | 2784 |
| | *w2v13+Google ResNet-152* | 0.48 | 0 | 3000 |
| $E_L + E_S$ | *w2v13+VG SceneGraph* | **0.70** | 0 | 2574 |
| | *crawl+VG SceneGraph* | 0.81 | 0 | 2574 |
| | *wikinews_sub+VG SceneGraph* | 0.45 | 0 | 2574 |
| | *wikinews+VG SceneGraph* | 0.65 | 0 | 2574 |

Table 4.4: Spearman correlation on the MEN dataset. Multi-modal embeddings are created using the Intersection technique. The table sections contain linguistic, visual and multi-modal embeddings in this order. Red colour signifies the best performance, blue means that the multi-modal embedding outperformed the corresponding uni-modal ones.

On the common subset of fMRI dataset, for all participants, visual, structured and multi-modal averages are higher than the linguistic ones.

In a recent paper [Pereira et al., 2018] also report high variance between subjects in the way their different systems (linguistic, visual, etc.) encode conceptual information: some are more visually oriented, others more linguistic, etc. Although, their dataset includes abstract concepts as well, which may explain the lesser involvement of visual information.

One important observation is that the standard deviations on the maximally covered fMRI data are moving around 0.1, whereas on MEG data they are around 0.06. On the

| Modality | Embedding | Spearman | P-value | Coverage |
|---|---|---|---|---|
| $E_L$ | *wikinews* | 0.45 | 0 | 999 |
| | *wikinews_sub* | 0.44 | 0 | 999 |
| | *crawl* | **0.50** | 0 | 999 |
| | *w2v13* | 0.31 | 0 | 999 |
| $E_V$ | *Google AlexNet* | 0.34 | 0 | 999 |
| | *Google VGG* | 0.34 | 0 | 999 |
| | *VG-internal* | 0.31 | 0 | 103 |
| | *VG-whole* | 0.19 | 0.06 | 103 |
| | *Google ResNet-152* | 0.35 | 0 | 999 |
| $E_S$ | *VG SceneGraph* | 0.26 | 0 | 593 |
| $E_L + E_V$ | *wikinews+Google AlexNet* | 0.34 | 0 | 999 |
| | *wikinews+Google VGG* | 0.34 | 0 | 999 |
| | *wikinews+VG-internal* | 0.31 | 0 | 103 |
| | *wikinews+VG-whole* | 0.18 | 0.06 | 103 |
| | *wikinews+Google ResNet-152* | 0.35 | 0 | 999 |
| | *wikinews_sub+Google AlexNet* | 0.34 | 0 | 999 |
| | *wikinews_sub+Google VGG* | 0.34 | 0 | 999 |
| | *wikinews_sub+VG-internal* | 0.31 | 0 | 103 |
| | *wikinews_sub+VG-whole* | 0.18 | 0.06 | 103 |
| | *wikinews_sub+Google ResNet-152* | 0.35 | 0 | 999 |
| | *crawl+Google AlexNet* | 0.34 | 0 | 999 |
| | *crawl+Google VGG* | 0.34 | 0 | 999 |
| | *crawl+VG-internal* | 0.31 | 0 | 103 |
| | *crawl+VG-whole* | 0.19 | 0.06 | 103 |
| | *crawl+Google ResNet-152* | 0.37 | 0 | 999 |
| | *w2v13+Google AlexNet* | 0.34 | 0 | 999 |
| | *w2v13+Google VGG* | 0.34 | 0 | 999 |
| | *w2v13+VG-internal* | 0.31 | 0 | 103 |
| | *w2v13+VG-whole* | 0.18 | 0.06 | 103 |
| | *w2v13+Google ResNet-152* | 0.35 | 0 | 999 |
| $E_L + E_S$ | *w2v13+VG SceneGraph* | 0.29 | 0 | 593 |
| | *crawl+VG SceneGraph* | 0.44 | 0 | 593 |
| | *wikinews_sub+VG SceneGraph* | 0.30 | 0 | 593 |
| | *wikinews+VG SceneGraph* | 0.35 | 0 | 593 |

Table 4.5: Spearman correlation on the SimLex dataset. Multi-modal embeddings are created using the Intersection technique. The table sections contain linguistic, visual and multi-modal embeddings in this order. Red colour signifies the best performance. Blue would mean that the multi-modal embedding outperformed the corresponding uni-modal ones, which here did not happen.

common subsets the numbers are around 0.08 and 0.09 respectively. In many cases the difference between models' average performances fall within these error margins. However, in most cases the improvements over uni-modal models go beyond this error.

### Concreteness

Figure 4.3 and 4.4 show Spearman's correlation scores of each of our embeddings on splits of the MEN and SimLex datasets, of size 100. On the $x$ axis we see the index of word pairs

| Modality | Embedding | Spearman | P-value | Coverage |
|---|---|---|---|---|
| $E_L$ | *wikinews* | 0.80 | 0 | 2481 |
| | *wikinews_sub* | 0.80 | 0 | 2481 |
| | *crawl* | **0.84** | 0 | 2481 |
| | *w2v13* | 0.67 | 0 | 2481 |
| $E_V$ | *Google AlexNet* | 0.52 | 0 | 2481 |
| | *Google VGG* | 0.51 | 0 | 2481 |
| | *VG-internal* | 0.38 | 0 | 2481 |
| | *VG-whole* | 0.41 | 0 | 2481 |
| | *Google ResNet-152* | 0.47 | 0 | 2481 |
| $E_S$ | *VG SceneGraph* | 0.44 | 0 | 2481 |
| $E_L + E_V$ | *wikinews+Google AlexNet* | 0.52 | 0 | 2481 |
| | *wikinews+Google VGG* | 0.52 | 0 | 2481 |
| | *wikinews+VG-internal* | 0.38 | 0 | 2481 |
| | *wikinews+VG-whole* | 0.41 | 0 | 2481 |
| | *wikinews+Google ResNet-152* | 0.48 | 0 | 2481 |
| | *wikinews_sub+Google AlexNet* | 0.52 | 0 | 2481 |
| | *wikinews_sub+Google VGG* | 0.51 | 0 | 2481 |
| | *wikinews_sub+VG-internal* | 0.38 | 0 | 2481 |
| | *wikinews_sub+VG-whole* | 0.41 | 0 | 2481 |
| | *wikinews_sub+Google ResNet-152* | 0.47 | 0 | 2481 |
| | *crawl+Google AlexNet* | 0.52 | 0 | 2481 |
| | *crawl+Google VGG* | 0.52 | 0 | 2481 |
| | *crawl+VG-internal* | 0.38 | 0 | 2481 |
| | *crawl+VG-whole* | 0.42 | 0 | 2481 |
| | *crawl+Google ResNet-152* | 0.51 | 0 | 2481 |
| | *w2v13+Google AlexNet* | 0.52 | 0 | 2481 |
| | *w2v13+Google VGG* | 0.52 | 0 | 2481 |
| | *w2v13+VG-internal* | 0.38 | 0 | 2481 |
| | *w2v13+VG-whole* | 0.41 | 0 | 2481 |
| | *w2v13+Google ResNet-152* | 0.49 | 0 | 2481 |
| $E_L + E_S$ | *w2v13+VG SceneGraph* | **0.70** | 0 | 2481 |
| | *crawl+VG SceneGraph* | 0.81 | 0 | 2481 |
| | *wikinews_sub+VG SceneGraph* | 0.46 | 0 | 2481 |
| | *wikinews+VG SceneGraph* | 0.66 | 0 | 2481 |

Table 4.6: Spearman correlation on the common subset of the MEN dataset. Multi-modal embeddings are created using the Intersection technique. The table sections contain linguistic, visual and multi-modal embeddings in this order. Red colour signifies the best performance, blue means that the multi-modal embedding outperformed the corresponding uni-modal ones.

in our respective evaluation sets, ordered by WordNet Concreteness, where concreteness for a word is computed using Equation 4.1.

Dark blue line indicates the WordNet Concreteness score for each word pair, therefore, it is on different axes than all the other lines which represent correlation scores. Figure 4.3 depicts the case when the concreteness of a word pair is the *sum* of the concretenesses of the individual words. In Figure 4.4 this is computed using their *absolute difference*. The two versions of synset aggregation for a word are both presented: *median* distance and

| Modality | Embedding | Spearman | P-value | Coverage |
|---|---|---|---|---|
| $E_L$ | *wikinews* | 0.28 | 0 | 103 |
| | *wikinews_sub* | 0.25 | 0.01 | 103 |
| | *crawl* | 0.37 | 0 | 103 |
| | *w2v13* | 0.11 | 0.25 | 103 |
| $E_V$ | *Google AlexNet* | **0.55** | 0 | 103 |
| | *Google VGG* | 0.53 | 0 | 103 |
| | *VG-internal* | 0.31 | 0 | 103 |
| | *VG-whole* | 0.19 | 0.06 | 103 |
| | *Google ResNet-152* | 0.50 | 0 | 103 |
| $E_S$ | *VG SceneGraph* | 0.30 | 0 | 103 |
| $E_L + E_V$ | *wikinews+Google AlexNet* | **0.55** | 0 | 103 |
| | *wikinews+Google VGG* | 0.53 | 0 | 103 |
| | *wikinews+VG-internal* | 0.31 | 0 | 103 |
| | *wikinews+VG-whole* | 0.18 | 0.06 | 103 |
| | *wikinews+Google ResNet-152* | 0.50 | 0 | 103 |
| | *wikinews_sub+Google AlexNet* | **0.55** | 0 | 103 |
| | *wikinews_sub+Google VGG* | 0.53 | 0 | 103 |
| | *wikinews_sub+VG-internal* | 0.31 | 0 | 103 |
| | *wikinews_sub+VG-whole* | 0.18 | 0.06 | 103 |
| | *wikinews_sub+Google ResNet-152* | 0.50 | 0 | 103 |
| | *crawl+Google AlexNet* | 0.55 | 0 | 103 |
| | *crawl+Google VGG* | 0.52 | 0 | 103 |
| | *crawl+VG-internal* | 0.31 | 0 | 103 |
| | *crawl+VG-whole* | 0.19 | 0.06 | 103 |
| | *crawl+Google ResNet-152* | 0.49 | 0 | 103 |
| | *w2v13+Google AlexNet* | 0.55 | 0 | 103 |
| | *w2v13+Google VGG* | 0.53 | 0 | 103 |
| | *w2v13+VG-internal* | 0.31 | 0 | 103 |
| | *w2v13+VG-whole* | 0.18 | 0.06 | 103 |
| | *w2v13+Google ResNet-152* | 0.49 | 0 | 103 |
| $E_L + E_S$ | *w2v13+VG SceneGraph* | 0.25 | 0.01 | 103 |
| | *crawl+VG SceneGraph* | 0.34 | 0 | 103 |
| | *wikinews_sub+VG SceneGraph* | 0.30 | 0 | 103 |
| | *wikinews+VG SceneGraph* | 0.29 | 0 | 103 |

Table 4.7: Spearman correlation on the common subset of the SimLex dataset. Multi-modal embeddings are created using the Intersection technique. The table sections contain linguistic, visual and multi-modal embeddings in this order. Red colour signifies the best performance. Blue would mean that the multi-modal embedding outperformed the corresponding uni-modal ones, which here did not happen.

*maximum distance* (most concrete) selection.

Perhaps because of the size of the datasets, we can see a tendency in the scores on MEN but way less on SimLex. When word pairs are ordered by the *sum* concreteness, we see a slightly upward trend as the concreteness score increases, especially in the *median* synset aggregation case. In the *absolute difference* concreteness ordering there is a steep growth for the first 5-10 splits, then the increase plummets.

Since we have a lot of embeddings, we use colour codes to separate embeddings by

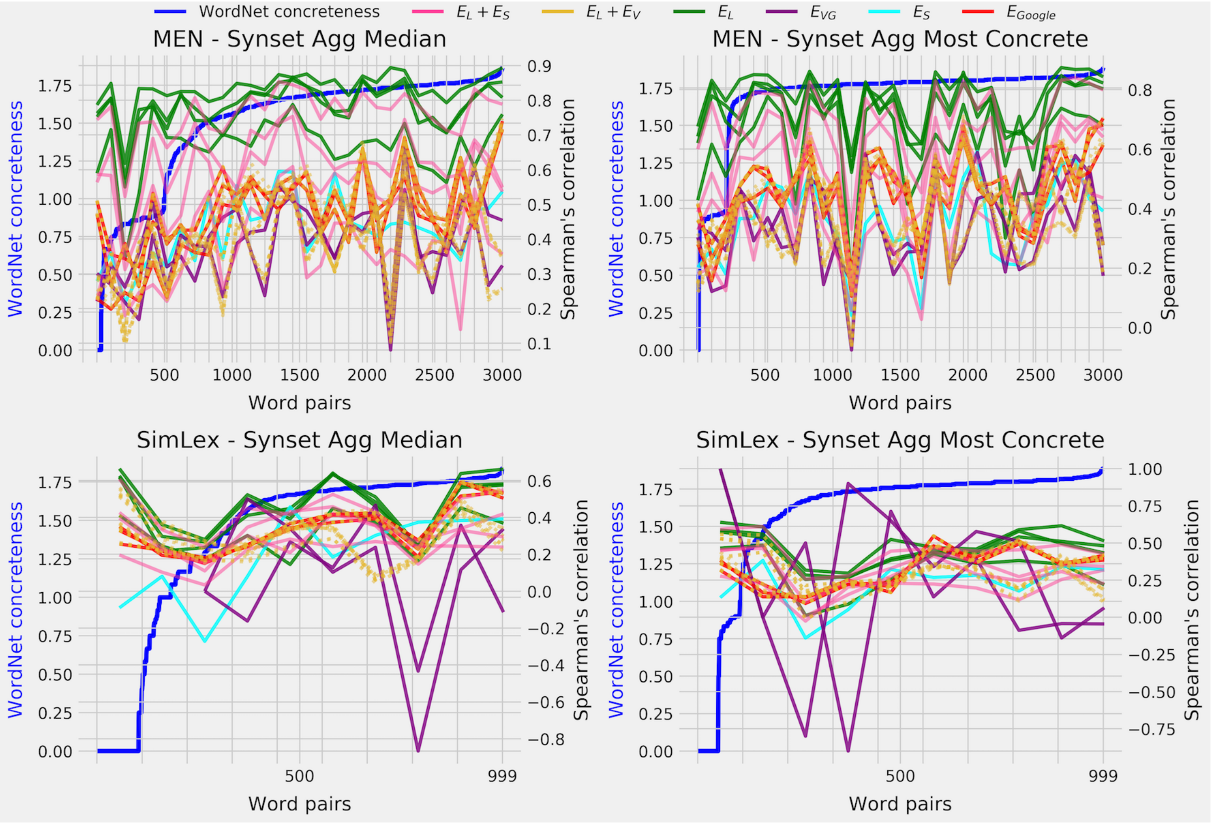


Figure 4.3: Spearman's correlation on the full Semantic Similarity dataset splits, ordered by the *sum* of WordNet concreteness scores of the two words in every word pair. Mid-fusion method: Padding. Axis $x$ shows the index of word pairs, ordered by WordNet Concreteness. There are two plots on top of each other for displaying the trend. Left $y$ axis is the scale for WordNet concreteness score (blue). Right $y$ axis is the scale for Spearman's correlations for all the embeddings $E_{\langle modality \rangle}$.

modality. Furthermore, we distinguish *Visual Genome* images $E_{VG}$ from *Google*, denoted by $E_{Google}$.

An interesting observation is that $E_S$ behaves more like a visual embedding in this experiment. A potential hypothesis is that for such abstract semantic tasks, (as opposed to traditional multi-modal tasks, such as VQA) we may not need low level visual features. Instead, it is rather the co-occurrence statistics, learned on this visually ordered graph structure, which can convey complementary information to a linguistic semantic embedding, trained on a "natural" text corpus. One potential way to test this hypothesis could be to gradually reduce the resolution of images we use for the visual embeddings and see how the performance changes, in what rate it starts to decline in particular. We would expect it to plateau or only decline slowly until a point when the objects are not distinguishable any more. This way we would see how much visual detail we can omit and keep the same gain for these conceptually abstract tasks.

Further results for evaluation on common subsets and *Intersection* type mid-fusion method can be found in Appendix B. They are consistent with the results presented here.

### Qualitative Analysis

Our automatic WordNet concreteness score is not a distinguishing metric for the 60 nouns in the Brain datasets, nevertheless, there can be some pattern when we look at the results for individual words.

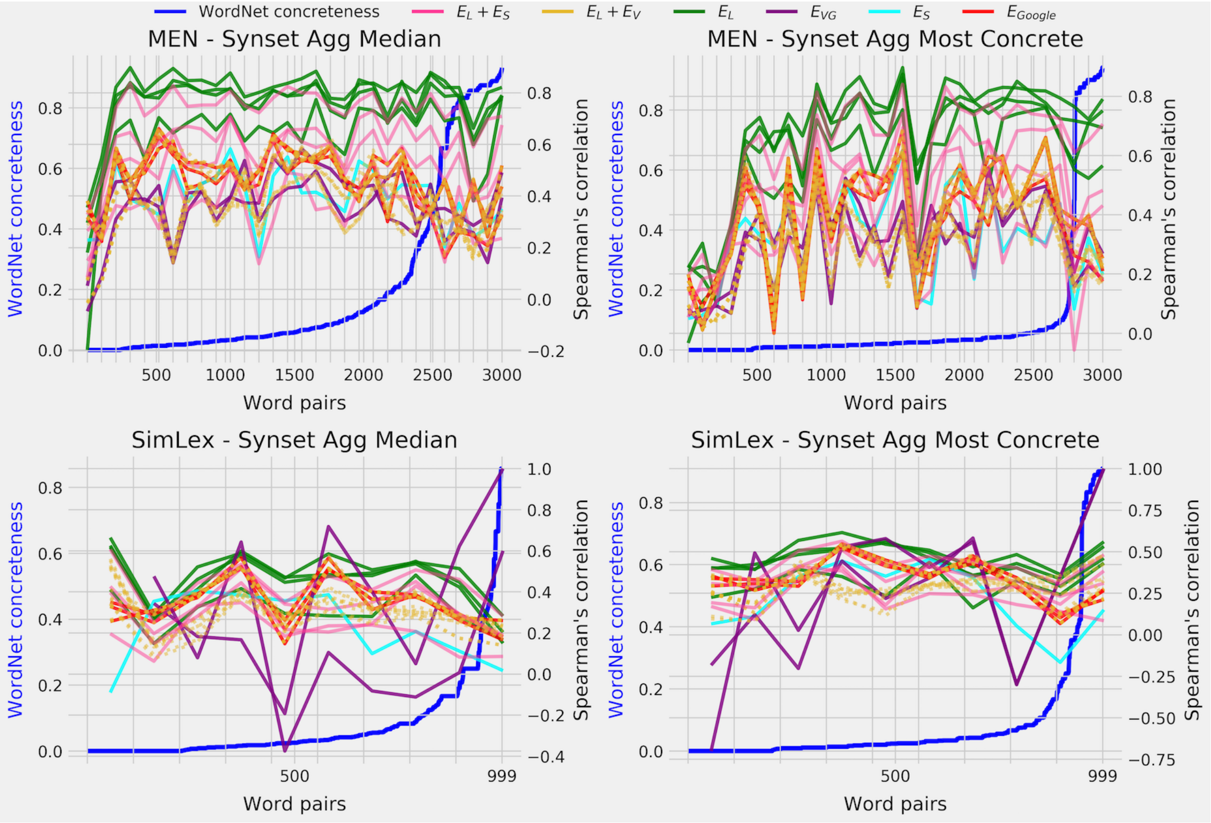


Figure 4.4: Spearman's correlation on the full Semantic Similarity dataset splits, ordered by the *difference* of WordNet concreteness scores of the two words in every word pair. Mid-fusion method: Padding. Axis $x$ shows the index of word pairs, ordered by WordNet Concreteness. There are two plots on top of each other for displaying the trend. Left $y$ axis is the scale for WordNet concreteness score (blue). Right $y$ axis is the scale for Spearman's correlations for all the embeddings $E_{\langle modality \rangle}$.

Figure 4.5 and Figure 4.6 show the number of hits for individual words, averaged over participants. A word gets a hit whenever it was in a word pair with a positive *2 vs. 2 test* score.

Here we order the plot by a *Visual Genome* based embedding on combined image segments. Some words, e.g., *barn*, *airplane* and *spoon* got very high rank in both the fMRI and the MEG dataset. Note that the participant sets of this two datasets are disjoint. It is harder to see such similarity for words with lower hit numbers. Embeddings, trained on different datasets and of different modalities follow a similar trend.

In order to get a better understanding of the type of words which behave differently in these brain imaging experiments we would need more evaluation data.

### 4.3.5 Conclusion

In this study we took a step towards a more detailed analysis on the impact of visual information on high level semantic tasks, with no direct visual input. Furthermore, we investigated two brain imaging evaluation sets, involving two different imaging methods: fMRI (Functional Magnetic Resonance Imaging) and one with MEG (Magnetoencephalography).

The results show that indeed, comparing several different visual and linguistic sources and models on various different evaluation tasks is necessary in order to avoid fooling ourselves with overfitting certain types of evaluation sets. In several occasion, previous

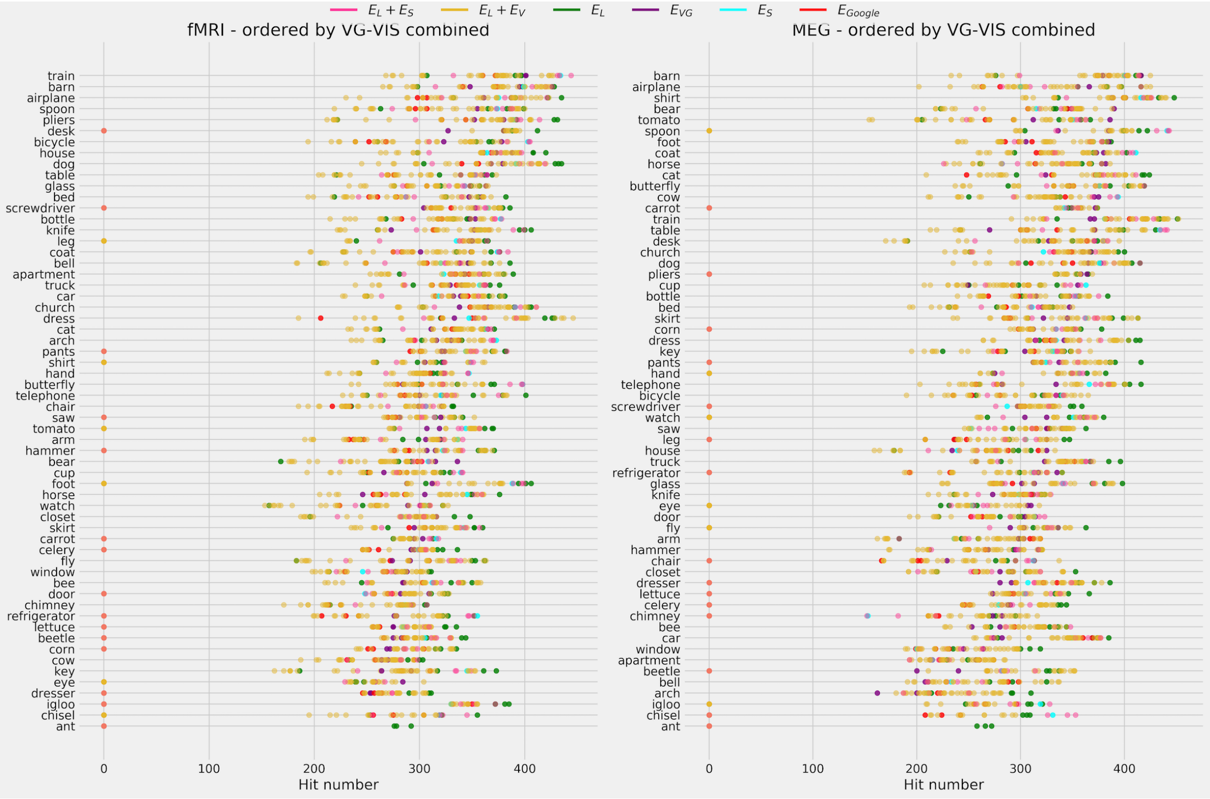


Figure 4.5: Scores on the full the Brain datasets words, ordered by the $E_{VG}$ score. The scores are the number of hits per word, averaged over all participants. Mid-fusion method: Intersection.

literature showed performance gain using multi-modal embeddings of linguistic and visual input. This is indeed the case on certain tasks and using certain embeddings, but not in every case.

In this work we aimed to shed light on the various factors that might play a role. Models behave differently on MEN and SimLex, and the performance gain of multi-modal models, when using linguistic vectors trained on huge textual sources is not well supported on these tasks. Visual information is complementary when our linguistic model has been trained on a smaller corpus, but this effect does not necessarily scale with corpus size.

Multi-modal models achieved a more convincing improvement on the brain imaging data, however these datasets are fairly small, so we would refrain from drawing far-reaching conclusions.

An interesting outcome of this study is that the model trained on the visually structured scene graph of *Visual Genome* achieved a surprising success across the board, despite its small size compared to all the other datasets. This is an interesting model, since it is linguistic in a sense that it is trained on text, but the word contexts are organised in a visually motivated structure. This suggests that images may indeed convey complementary statistical information about the co-occurrence of objects in visual scenes. It is even possible that this information is more important for abstract semantic tasks than lower level visual properties of words. This would be intuitive, since unlike multi-modal tasks with direct visual input, such as Visual Question Answering, in our case we are aiming for abstract meaning representations of concepts. It would make sense if detailed visual information about what a table looks like mattered less when we talk about table as an abstract concept.

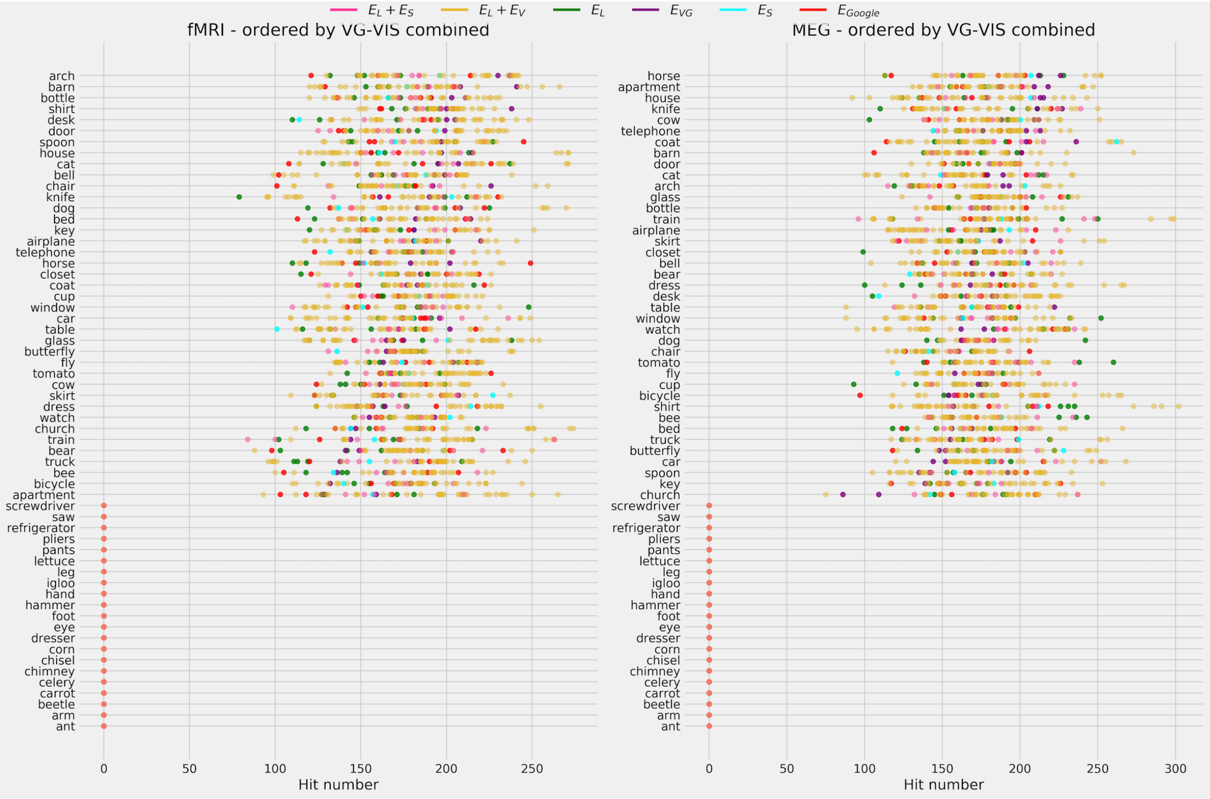


Figure 4.6: Scores on the embeddings' common subset of the Brain datasets words, ordered by the $E_{VG}$ score. The scores are the number of hits per word, averaged over all participants. Mid-fusion method: Intersection.

## 4.4 Model Initialization on a Textual Entailment Task

This section is a brief digression to studying the application possibilities of word embeddings as initialisations on a sentence level task: textual entailment. We evaluate on the Stanford Natural Language Inference (SNLI) corpus [Bowman et al., 2015].

We compared five different neural network models for encoding sentences and four different word embeddings to initialize these models, following the baselines of [Bowman et al., 2015]. The task is a three-way classification, where the input is a sentence pair and the classification labels are *entailment*, *contradiction* and *neutral*. We included words in the multi-modal representation only if they have a visual representation.

On the top of each model there is a three-fold classifier on the concatenated sentence embeddings for the premise and the hypothesis sentences. The five sentence encoding models are the following[4]:

1. **Addition**: Vector addition of word embedding vectors in the sentence.

2. **Addition + translation layer**: The previous model extended with an additional layer that learns another sentence embedding above the fixed word embedding based sentence representation.

3. **Addition + translation layer + full size image embeddings**: The model above, but instead of using dimensionality reduced visual vectors, we use the original

[4] Some of the base code was written in collaboration with Amandla Mabona.

image embeddings and smaller (100 dimensional) linguistic vectors. In the previous models we used PCA to keep the first 300 components out of 4094.

4. **GRU**: Gated Recurrent Unit based recurrent sentence encoding model.

5. **LSTM**: Recurrent sentence encoding model with Long short-term memory units.

All of them were initialized with four different word embeddings:

1. **Linguistic only**: Skip-gram embedding trained on a 2013 Wikipedia dump.

2. **Visual only**: Image embeddings, extracting CNN representations of Google images for the individual words in the sentence.

3. **Multi-modal**: The concatenation of linguistic and visual vectors for each word.

4. **Random**: The initial vector weights are sampled from a normal distribution.

### 4.4.1 Results

The results are shown in Table 4.16. The experiments indicate two phenomena:

1. The translation layer plays an important role in models 1-3. In these cases the simplest model without the translation layer (model 1.) the linguistic initialisation performs the best. After adding the translation layer, however, multi-modal embeddings outperform all the other ones, in case of full size image vectors (model 3.) with a substantial margin in classification accuracy.

2. In case of the more sophisticated recurrent models (4-5.), however, we found that the performance difference across different initialisations vanishes. Even the random initial embeddings do not achieve significantly lower classification accuracy then the other methods.

### 4.4.2 Conclusion

The second finding may suggest that we could create more time efficient models, since we do not necessarily need to spend time on pre-training word embeddings. It also alerts us, however, to the danger of overfitting. Note that we ignored multi-modal representations for words where visual information is missing, which may hurt performance. Although the high performance of random initialisations are more telling. Our findings are in line with Zhang and Bowman's, who found the related phenomenon of high performing random initialized LSTM models [Zhang and Bowman, 2018]. [Yogatama et al., 2019] recently found that transformer type models are overfitting to the quirks of particular datasets. Possible future work could be to gradually increase model complexity as well as performing more ablation studies, in order to better understand the models' capacity.

## 4.5 Conclusion

In this Chapter we demonstrated the effectiveness of image search engines in multi-modal mid-fusion embeddings. We found that around the first 10 image results are sufficient, beyond that the performance plateaus.

We introduced a new visually structured textual embedding based on Visual Genome and showed that it enriches linguistic models trained on smaller corpora, therefore they can be useful for low resource languages.

We found that pretrained word embeddings do not necessarily help sequence model training. However, they can be valuable on their own for discovering concept structures in a data source.

Based on these findings we move on to an in-depth study of our embeddings of different modalities and their combinations. The following chapters showcase the second and third pillars of our methodology, which involve *transparency* analysis (see Section 3.3). We narrow our focus to a few models, as such analyses would be fairly time consuming for all the above combinations of sources, modalities and models. Furthermore, such studies on numerous current models would become shortly obsolete. Our aim is rather to provide a general framework with proof-of-concept studies, which can be applied to various models in the future.

| **Modality** | **Embedding** | **P1** | **P2** | **P3** | **P4** | **P5** | **P6** | **P7** | **P8** | **P9** | **Avg** | **STD** | **Covr.** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $E_L$ | *w2v13* | 0.79 | 0.54 | 0.66 | 0.76 | 0.58 | 0.65 | 0.47 | 0.60 | 0.67 | 0.64 | 0.1 | 45 |
| | *wikinews_sub* | 0.83 | 0.66 | 0.68 | 0.83 | 0.61 | 0.54 | 0.59 | 0.56 | 0.70 | 0.67 | 0.1 | 60 |
| | *wikinews* | 0.83 | 0.68 | 0.63 | 0.81 | 0.64 | 0.54 | 0.56 | 0.48 | 0.65 | 0.65 | 0.11 | 60 |
| | *crawl* | 0.86 | 0.68 | 0.61 | **0.88** | 0.65 | 0.58 | 0.58 | 0.55 | 0.60 | 0.67 | 0.12 | 60 |
| $E_V$ | *Google-VIS whole* | 0.89 | 0.65 | 0.64 | 0.75 | 0.51 | 0.61 | 0.64 | 0.55 | 0.60 | 0.65 | 0.11 | 52 |
| | *Google ResNet-152* | 0.88 | 0.63 | 0.64 | 0.73 | 0.46 | 0.56 | 0.64 | 0.50 | 0.56 | 0.62 | 0.12 | 52 |
| | *VG-VIS internal* | 0.85 | 0.70 | 0.63 | 0.72 | 0.52 | 0.55 | 0.57 | 0.47 | 0.57 | 0.62 | 0.11 | 57 |
| | *Google AlexNet* | 0.89 | 0.61 | 0.66 | 0.72 | 0.48 | 0.63 | 0.62 | 0.54 | 0.66 | 0.65 | 0.11 | 52 |
| | *VG-VIS combined* | 0.85 | 0.71 | 0.65 | 0.76 | 0.55 | 0.57 | 0.60 | 0.44 | 0.66 | 0.64 | 0.11 | 57 |
| $E_S$ | *VG SceneGraph* | 0.83 | 0.68 | 0.57 | 0.77 | 0.59 | 0.63 | 0.58 | 0.59 | 0.64 | 0.65 | 0.09 | 58 |
| $E_L + E_V$ | *VG-MM internal* | 0.88 | 0.66 | 0.66 | 0.78 | 0.59 | 0.64 | 0.67 | 0.47 | 0.65 | 0.67 | 0.11 | 57 |
| | *VG-MM combined* | 0.88 | 0.64 | 0.67 | 0.79 | 0.61 | 0.65 | **0.67** | 0.48 | 0.68 | 0.67 | 0.11 | 57 |
| | *Google-MM whole* | 0.89 | 0.67 | 0.67 | 0.80 | 0.61 | 0.60 | 0.65 | 0.52 | 0.64 | 0.67 | 0.1 | 52 |
| | *wikinews+Google ResNet-152* | 0.88 | 0.63 | 0.64 | 0.75 | 0.47 | 0.56 | 0.63 | 0.49 | 0.55 | 0.62 | 0.12 | 52 |
| | *wikinews+Google AlexNet* | 0.89 | 0.61 | 0.66 | 0.73 | 0.48 | 0.63 | 0.62 | 0.54 | 0.66 | 0.65 | 0.11 | 52 |
| | *wikinews+VG-VIS internal* | 0.84 | 0.69 | **0.66** | 0.80 | **0.65** | 0.55 | **0.62** | 0.47 | **0.66** | **0.66** | 0.11 | 57 |
| | *wikinews+VG-MM internal* | 0.83 | 0.67 | 0.66 | 0.80 | **0.65** | 0.56 | 0.62 | 0.47 | **0.67** | 0.66 | 0.1 | 57 |
| | *wikinews+VG-VIS combined* | 0.84 | 0.68 | **0.66** | 0.80 | **0.65** | 0.56 | **0.63** | 0.47 | **0.67** | **0.66** | 0.11 | 57 |
| | *wikinews+VG-MM combined* | 0.83 | 0.67 | 0.66 | 0.80 | **0.65** | 0.56 | 0.63 | 0.48 | 0.67 | 0.66 | 0.1 | 57 |
| | *wikinews+Google-VIS whole* | 0.85 | **0.72** | **0.70** | 0.79 | **0.68** | 0.50 | 0.60 | 0.55 | 0.65 | **0.67** | 0.11 | 52 |
| | *wikinews+Google-MM whole* | 0.83 | **0.72** | **0.68** | 0.80 | **0.69** | 0.49 | 0.58 | **0.55** | 0.65 | 0.67 | 0.11 | 52 |
| | *wikinews_sub+Google ResNet-152* | 0.88 | 0.63 | 0.63 | 0.74 | 0.46 | 0.56 | 0.63 | 0.49 | 0.55 | 0.62 | 0.12 | 52 |
| | *wikinews_sub+Google AlexNet* | 0.89 | 0.61 | 0.66 | 0.73 | 0.48 | 0.63 | 0.62 | 0.54 | 0.66 | 0.65 | 0.11 | 52 |
| | *wikinews_sub+VG-VIS internal* | **0.87** | 0.68 | 0.67 | 0.78 | 0.59 | **0.57** | 0.57 | 0.50 | 0.61 | 0.65 | 0.11 | 57 |
| | *wikinews_sub+VG-MM internal* | 0.88 | 0.64 | **0.69** | 0.80 | **0.63** | 0.63 | 0.66 | 0.51 | 0.66 | **0.68** | 0.1 | 57 |
| | *wikinews_sub+VG-VIS combined* | **0.87** | 0.70 | 0.67 | 0.81 | 0.60 | **0.58** | **0.62** | 0.48 | 0.67 | 0.67 | 0.11 | 57 |
| | *wikinews_sub+VG-MM combined* | 0.87 | 0.64 | **0.69** | 0.81 | **0.63** | 0.63 | 0.67 | 0.52 | 0.67 | **0.68** | 0.1 | 57 |
| | *wikinews_sub+Google-VIS whole* | 0.89 | **0.67** | 0.66 | 0.77 | 0.52 | 0.58 | 0.64 | 0.55 | 0.62 | 0.66 | 0.11 | 52 |
| | *wikinews_sub+Google-MM whole* | 0.88 | **0.69** | **0.70** | 0.81 | **0.64** | 0.57 | 0.63 | 0.55 | 0.65 | **0.68** | 0.1 | 52 |
| | *crawl+Google ResNet-152* | 0.88 | 0.64 | 0.64 | 0.75 | 0.47 | 0.55 | 0.62 | 0.50 | 0.55 | 0.62 | 0.12 | 52 |
| | *crawl+Google AlexNet* | 0.89 | 0.61 | 0.66 | 0.73 | 0.48 | 0.63 | 0.62 | 0.54 | 0.66 | 0.65 | 0.11 | 52 |
| | *crawl+VG-VIS internal* | **0.87** | 0.68 | 0.62 | 0.86 | **0.67** | **0.60** | **0.62** | 0.53 | **0.61** | 0.67 | 0.11 | 57 |
| | *crawl+VG-MM internal* | 0.87 | 0.67 | 0.62 | 0.86 | **0.67** | 0.60 | 0.62 | 0.53 | 0.61 | 0.67 | 0.11 | 57 |
| | *crawl+VG-VIS combined* | **0.87** | 0.68 | 0.62 | 0.86 | **0.67** | **0.60** | **0.63** | 0.53 | 0.62 | 0.67 | 0.11 | 57 |
| | *crawl+VG-MM combined* | 0.87 | 0.67 | 0.62 | 0.86 | **0.67** | 0.60 | 0.62 | 0.53 | 0.61 | 0.67 | 0.11 | 57 |
| | *crawl+Google-VIS whole* | 0.87 | **0.72** | **0.69** | 0.87 | **0.72** | 0.51 | 0.60 | **0.60** | 0.57 | **0.69** | 0.12 | 52 |
| | *crawl+Google-MM whole* | 0.86 | **0.72** | **0.69** | 0.87 | **0.73** | 0.51 | 0.60 | **0.61** | 0.57 | **0.68** | 0.12 | 52 |
| | *w2v13+Google ResNet-152* | **0.89** | **0.65** | **0.68** | 0.75 | 0.50 | 0.56 | 0.56 | 0.54 | 0.67 | 0.64 | 0.11 | 40 |
| | *w2v13+Google AlexNet* | **0.90** | **0.66** | **0.71** | 0.74 | 0.53 | 0.64 | 0.58 | 0.57 | **0.77** | **0.68** | 0.11 | 40 |
| | *w2v13+VG-VIS internal* | 0.81 | 0.55 | 0.66 | 0.76 | **0.60** | **0.68** | 0.49 | 0.59 | 0.67 | **0.65** | 0.09 | 44 |
| | *w2v13+VG-MM internal* | 0.80 | 0.55 | 0.65 | 0.76 | **0.60** | **0.67** | 0.50 | 0.59 | **0.68** | 0.65 | 0.09 | 44 |
| | *w2v13+VG-VIS combined* | 0.81 | 0.54 | 0.66 | 0.76 | **0.60** | **0.68** | 0.50 | 0.59 | **0.68** | **0.65** | 0.09 | 44 |
| | *w2v13+VG-MM combined* | 0.80 | 0.55 | 0.65 | 0.76 | 0.60 | **0.67** | 0.50 | 0.59 | 0.68 | 0.64 | 0.09 | 44 |
| | *w2v13+Google-VIS whole* | 0.84 | 0.61 | **0.68** | 0.75 | **0.62** | 0.59 | 0.44 | **0.64** | 0.66 | 0.65 | 0.1 | 40 |
| | *w2v13+Google-MM whole* | 0.82 | 0.59 | 0.67 | 0.74 | **0.62** | 0.59 | 0.45 | **0.64** | 0.65 | 0.64 | 0.1 | 40 |
| $E_L + E_S$ | *wikinews+VG SceneGraph* | **0.84** | **0.71** | 0.59 | 0.80 | 0.63 | 0.60 | 0.58 | 0.58 | 0.65 | **0.66** | 0.09 | 58 |
| | *wikinews_sub+VG SceneGraph* | **0.84** | 0.68 | 0.57 | 0.78 | 0.60 | 0.63 | 0.59 | **0.60** | 0.64 | 0.66 | 0.09 | 58 |
| | *crawl+VG SceneGraph* | **0.87** | **0.71** | 0.59 | 0.86 | **0.66** | 0.60 | **0.60** | **0.60** | 0.64 | **0.68** | 0.1 | 58 |
| | *w2v13+VG SceneGraph* | **0.87** | 0.65 | 0.66 | **0.81** | **0.66** | **0.69** | 0.52 | **0.61** | **0.76** | **0.69** | 0.1 | 45 |

Table 4.8: fMRI scores for each participant and embedding. Multi-modal embeddings are created using the Intersection technique. The table sections contain linguistic, visual and multi-modal embeddings in this order. Red colour signifies the best performance, blue means that the multi-modal embedding outperformed the corresponding uni-modal ones.

| Modality | Embedding | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 | Avg | STD | Coverage |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $E_L$ | *w2v13* | 0.65 | 0.64 | 0.58 | 0.64 | 0.74 | 0.65 | 0.75 | 0.56 | 0.69 | 0.66 | 0.06 | 45 |
| | *wikinews_sub* | 0.63 | 0.59 | 0.48 | 0.70 | 0.72 | 0.65 | 0.71 | **0.66** | 0.73 | 0.65 | 0.07 | 60 |
| | *wikinews* | 0.63 | 0.61 | 0.50 | 0.71 | 0.71 | 0.64 | 0.72 | 0.63 | 0.76 | 0.66 | 0.07 | 60 |
| | *crawl* | 0.65 | 0.58 | 0.57 | 0.69 | 0.67 | 0.63 | 0.73 | 0.65 | 0.71 | 0.65 | 0.05 | 60 |
| $E_V$ | *Google-VIS whole* | 0.70 | 0.51 | 0.56 | 0.71 | 0.69 | 0.73 | 0.70 | 0.62 | 0.69 | 0.66 | 0.07 | 52 |
| | *Google ResNet-152* | 0.65 | 0.55 | 0.52 | 0.69 | 0.70 | 0.68 | 0.63 | 0.61 | 0.66 | 0.63 | 0.06 | 52 |
| | *VG-VIS internal* | 0.62 | 0.55 | 0.54 | 0.66 | 0.62 | 0.69 | 0.64 | 0.49 | 0.59 | 0.60 | 0.06 | 57 |
| | *Google AlexNet* | 0.66 | 0.52 | 0.57 | 0.66 | 0.69 | 0.71 | 0.69 | 0.57 | 0.65 | 0.63 | 0.06 | 52 |
| | *VG-VIS combined* | 0.69 | 0.60 | 0.55 | 0.68 | 0.70 | **0.76** | 0.68 | 0.56 | 0.69 | 0.66 | 0.07 | 57 |
| $E_S$ | *VG SceneGraph* | 0.63 | 0.60 | 0.55 | 0.65 | 0.70 | 0.62 | 0.67 | 0.50 | 0.73 | 0.63 | 0.07 | 58 |
| $E_L + E_V$ | *VG-MM internal* | 0.66 | 0.65 | 0.56 | 0.73 | 0.68 | 0.64 | 0.70 | 0.60 | 0.69 | 0.65 | 0.05 | 57 |
| | *VG-MM combined* | 0.68 | 0.67 | 0.56 | 0.74 | 0.71 | 0.69 | 0.72 | 0.62 | 0.72 | 0.68 | 0.05 | 57 |
| | *Google-MM whole* | **0.72** | 0.59 | 0.53 | 0.72 | 0.71 | 0.67 | 0.72 | 0.65 | 0.72 | 0.67 | 0.06 | 52 |
| | *wikinews+Google ResNet-152* | 0.65 | 0.55 | 0.51 | 0.68 | 0.70 | **0.69** | 0.63 | 0.61 | 0.66 | 0.63 | 0.06 | 52 |
| | *wikinews+Google AlexNet* | 0.66 | 0.52 | 0.57 | 0.66 | 0.69 | 0.71 | 0.69 | 0.57 | 0.65 | 0.64 | 0.06 | 52 |
| | *wikinews+VG-VIS internal* | 0.63 | **0.67** | 0.53 | 0.71 | **0.74** | 0.69 | 0.72 | 0.60 | 0.76 | **0.67** | 0.07 | 57 |
| | *wikinews+VG-MM internal* | 0.62 | **0.67** | 0.54 | 0.70 | **0.75** | **0.68** | **0.73** | 0.62 | 0.76 | **0.67** | 0.07 | 57 |
| | *wikinews+VG-VIS combined* | 0.64 | **0.66** | 0.53 | 0.71 | **0.74** | 0.71 | **0.73** | 0.62 | **0.77** | **0.68** | 0.07 | 57 |
| | *wikinews+VG-MM combined* | 0.62 | 0.67 | 0.53 | 0.70 | **0.75** | 0.69 | **0.73** | 0.62 | 0.76 | 0.67 | 0.07 | 57 |
| | *wikinews+Google-VIS whole* | 0.66 | 0.61 | 0.53 | 0.70 | **0.76** | 0.70 | 0.72 | 0.61 | 0.75 | **0.67** | 0.07 | 52 |
| | *wikinews+Google-MM whole* | 0.66 | **0.63** | 0.53 | 0.69 | **0.76** | 0.66 | 0.71 | 0.61 | 0.76 | 0.67 | 0.07 | 52 |
| | *wikinews_sub+Google ResNet-152* | 0.65 | 0.55 | 0.51 | 0.68 | 0.70 | 0.68 | 0.62 | 0.61 | 0.66 | 0.63 | 0.06 | 52 |
| | *wikinews_sub+Google AlexNet* | 0.66 | 0.52 | 0.57 | 0.66 | 0.69 | 0.71 | 0.69 | 0.57 | 0.65 | 0.64 | 0.06 | 52 |
| | *wikinews_sub+VG-VIS internal* | **0.64** | **0.61** | **0.56** | **0.72** | 0.68 | **0.70** | 0.70 | 0.54 | 0.67 | 0.65 | 0.06 | 57 |
| | *wikinews_sub+VG-MM internal* | 0.66 | **0.66** | 0.56 | **0.75** | 0.72 | **0.67** | **0.72** | 0.63 | 0.73 | **0.68** | 0.06 | 57 |
| | *wikinews_sub+VG-VIS combined* | **0.70** | **0.64** | **0.57** | **0.73** | **0.73** | 0.75 | **0.72** | 0.59 | 0.71 | **0.68** | 0.06 | 57 |
| | *wikinews_sub+VG-MM combined* | 0.68 | **0.67** | 0.55 | **0.76** | **0.74** | **0.71** | **0.73** | 0.63 | **0.75** | **0.69** | 0.06 | 57 |
| | *wikinews_sub+Google-VIS whole* | 0.70 | 0.53 | 0.55 | 0.70 | **0.73** | 0.73 | 0.71 | 0.63 | 0.70 | 0.66 | 0.07 | 52 |
| | *wikinews_sub+Google-MM whole* | 0.71 | **0.62** | 0.53 | 0.71 | **0.76** | 0.67 | 0.72 | 0.65 | **0.74** | **0.68** | 0.07 | 52 |
| | *crawl+Google ResNet-152* | 0.65 | 0.56 | 0.51 | 0.68 | **0.71** | 0.68 | 0.63 | 0.62 | 0.66 | 0.63 | 0.06 | 52 |
| | *crawl+Google AlexNet* | **0.67** | 0.52 | 0.57 | 0.66 | 0.69 | 0.71 | 0.69 | 0.57 | 0.65 | 0.64 | 0.06 | 52 |
| | *crawl+VG-VIS internal* | 0.65 | **0.63** | **0.60** | 0.69 | **0.69** | 0.68 | 0.73 | 0.65 | **0.73** | **0.67** | 0.04 | 57 |
| | *crawl+VG-MM internal* | 0.65 | 0.64 | **0.60** | 0.69 | **0.69** | **0.67** | 0.73 | 0.65 | **0.73** | **0.67** | 0.04 | 57 |
| | *crawl+VG-VIS combined* | 0.65 | **0.63** | **0.60** | 0.69 | 0.69 | 0.68 | 0.73 | 0.65 | **0.73** | **0.67** | 0.04 | 57 |
| | *crawl+VG-MM combined* | 0.65 | 0.64 | **0.61** | 0.69 | 0.69 | 0.67 | 0.73 | 0.65 | **0.73** | 0.67 | 0.04 | 57 |
| | *crawl+Google-VIS whole* | 0.67 | **0.62** | 0.56 | 0.68 | **0.77** | 0.67 | 0.73 | 0.62 | **0.75** | **0.67** | 0.06 | 52 |
| | *crawl+Google-MM whole* | 0.67 | **0.63** | 0.57 | 0.68 | **0.77** | 0.66 | 0.73 | 0.62 | **0.75** | 0.67 | 0.06 | 52 |
| | *w2v13+Google ResNet-152* | **0.68** | **0.66** | **0.60** | **0.72** | 0.74 | **0.69** | 0.73 | **0.62** | 0.66 | **0.68** | 0.05 | 40 |
| | *w2v13+Google AlexNet* | **0.69** | 0.59 | **0.67** | **0.75** | **0.77** | **0.74** | 0.73 | **0.62** | 0.69 | **0.69** | 0.06 | 40 |
| | *w2v13+VG-VIS internal* | 0.65 | **0.65** | **0.61** | 0.62 | **0.75** | 0.65 | 0.75 | 0.53 | **0.71** | 0.66 | 0.07 | 44 |
| | *w2v13+VG-MM internal* | 0.64 | 0.64 | **0.60** | 0.62 | **0.75** | 0.65 | 0.74 | 0.54 | **0.70** | 0.66 | 0.06 | 44 |
| | *w2v13+VG-VIS combined* | 0.66 | 0.64 | **0.61** | 0.62 | **0.76** | 0.67 | 0.74 | 0.54 | **0.71** | 0.66 | 0.06 | 44 |
| | *w2v13+VG-MM combined* | 0.64 | 0.64 | **0.60** | 0.63 | **0.75** | 0.65 | 0.74 | 0.54 | 0.70 | 0.66 | 0.06 | 44 |
| | *w2v13+Google-VIS whole* | 0.69 | 0.59 | 0.56 | 0.64 | 0.73 | 0.65 | 0.69 | 0.56 | **0.71** | 0.65 | 0.06 | 40 |
| | *w2v13+Google-MM whole* | 0.68 | 0.59 | 0.54 | 0.63 | 0.71 | 0.63 | 0.70 | 0.56 | 0.71 | 0.64 | 0.06 | 40 |
| $E_L + E_S$ | *wikinews+VG SceneGraph* | 0.62 | 0.61 | 0.55 | 0.69 | **0.73** | 0.62 | 0.71 | 0.55 | 0.75 | 0.65 | 0.07 | 58 |
| | *wikinews_sub+VG SceneGraph* | 0.62 | 0.60 | 0.55 | 0.67 | 0.70 | 0.62 | 0.68 | 0.51 | 0.73 | 0.63 | 0.07 | 58 |
| | *crawl+VG SceneGraph* | **0.66** | **0.63** | 0.55 | 0.69 | **0.72** | 0.62 | **0.74** | 0.62 | **0.76** | **0.67** | 0.06 | 58 |
| | *w2v13+VG SceneGraph* | **0.69** | 0.63 | 0.56 | **0.68** | **0.81** | **0.68** | **0.78** | **0.59** | 0.72 | **0.68** | 0.08 | 45 |

Table 4.9: MEG scores for each participant and embedding. Multi-modal embeddings are created using the Intersection technique. The table sections contain linguistic, visual and multi-modal embeddings in this order. Red colour signifies the best performance, blue means that the multi-modal embedding outperformed the corresponding uni-modal ones.

| Modality | Embedding | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 | Avg | STD | Coverage |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $E_L$ | *w2v13* | 0.41 | 0.41 | 0.33 | 0.46 | 0.44 | 0.49 | 0.43 | 0.51 | 0.55 | 0.45 | 0.06 | 39 |
| | *wikinews_sub* | 0.47 | 0.47 | 0.52 | 0.49 | 0.48 | 0.45 | 0.50 | 0.52 | 0.52 | 0.49 | 0.03 | 39 |
| | *wikinews* | 0.54 | 0.44 | 0.50 | 0.49 | 0.51 | 0.52 | 0.53 | 0.46 | 0.54 | 0.50 | 0.03 | 39 |
| | *crawl* | 0.41 | 0.49 | 0.44 | 0.54 | 0.56 | 0.48 | 0.47 | 0.39 | 0.56 | 0.48 | 0.06 | 39 |
| $E_V$ | *Google-VIS whole* | 0.47 | 0.33 | 0.54 | 0.46 | 0.44 | 0.41 | 0.24 | 0.49 | 0.45 | 0.43 | 0.08 | 39 |
| | *Google ResNet-152* | 0.42 | 0.39 | 0.58 | 0.54 | 0.53 | 0.53 | 0.57 | 0.62 | 0.39 | 0.51 | 0.08 | 39 |
| | *VG-VIS internal* | 0.66 | 0.37 | 0.60 | 0.65 | 0.58 | 0.57 | 0.52 | 0.56 | 0.56 | 0.56 | 0.08 | 39 |
| | *Google AlexNet* | 0.59 | 0.58 | 0.54 | 0.58 | 0.53 | 0.61 | 0.53 | 0.59 | 0.56 | 0.57 | 0.03 | 39 |
| | *VG-VIS combined* | 0.57 | 0.34 | 0.57 | 0.58 | 0.53 | 0.52 | 0.49 | 0.53 | 0.46 | 0.51 | 0.07 | 39 |
| $E_S$ | *VG SceneGraph* | 0.51 | 0.45 | 0.49 | 0.55 | 0.63 | 0.38 | 0.55 | 0.48 | 0.36 | 0.49 | 0.08 | 39 |
| $E_L + E_V$ | *VG-MM internal* | 0.70 | 0.61 | 0.46 | 0.45 | 0.55 | 0.48 | 0.57 | 0.60 | 0.51 | 0.55 | 0.08 | 39 |
| | *VG-MM combined* | 0.48 | 0.43 | 0.45 | 0.51 | 0.47 | **0.66** | 0.44 | 0.58 | 0.56 | 0.51 | 0.07 | 39 |
| | *Google-MM whole* | 0.45 | 0.32 | 0.48 | 0.44 | 0.46 | 0.49 | 0.27 | 0.49 | 0.40 | 0.43 | 0.07 | 39 |
| | *wikinews+Google ResNet-152* | 0.33 | 0.38 | **0.59** | 0.42 | 0.43 | 0.39 | **0.69** | 0.50 | **0.64** | 0.49 | 0.12 | 39 |
| | *wikinews+Google AlexNet* | 0.42 | 0.41 | 0.47 | 0.49 | 0.49 | 0.41 | **0.65** | 0.48 | 0.55 | 0.48 | 0.07 | 39 |
| | *wikinews+VG-VIS internal* | 0.64 | **0.69** | 0.43 | **0.66** | **0.64** | 0.50 | 0.53 | **0.62** | 0.55 | **0.58** | 0.08 | 39 |
| | *wikinews+VG-MM internal* | 0.66 | **0.68** | 0.40 | **0.67** | **0.61** | 0.50 | 0.55 | **0.64** | 0.52 | **0.58** | 0.09 | 39 |
| | *wikinews+VG-VIS combined* | **0.65** | **0.67** | 0.42 | **0.63** | **0.63** | 0.51 | **0.54** | **0.63** | **0.55** | **0.58** | 0.08 | 39 |
| | *wikinews+VG-MM combined* | **0.69** | **0.70** | 0.42 | **0.65** | **0.60** | 0.49 | **0.61** | **0.66** | 0.48 | **0.59** | 0.1 | 39 |
| | *wikinews+Google-VIS whole* | 0.43 | **0.46** | 0.49 | 0.40 | 0.51 | 0.41 | **0.63** | 0.40 | 0.50 | 0.47 | 0.07 | 39 |
| | *wikinews+Google-MM whole* | 0.42 | **0.48** | 0.47 | 0.41 | **0.53** | 0.41 | **0.63** | 0.41 | 0.49 | 0.47 | 0.07 | 39 |
| | *wikinews_sub+Google ResNet-152* | 0.33 | 0.39 | **0.59** | 0.42 | 0.43 | 0.39 | **0.69** | 0.50 | **0.65** | 0.49 | 0.12 | 39 |
| | *wikinews_sub+Google AlexNet* | 0.42 | 0.41 | 0.47 | 0.49 | 0.48 | 0.41 | **0.65** | 0.48 | 0.55 | 0.48 | 0.07 | 39 |
| | *wikinews_sub+VG-VIS internal* | 0.60 | **0.57** | 0.52 | **0.68** | 0.54 | 0.57 | 0.43 | **0.58** | 0.52 | 0.55 | 0.06 | 39 |
| | *wikinews_sub+VG-MM internal* | 0.68 | 0.57 | 0.40 | **0.75** | **0.57** | **0.51** | 0.43 | **0.64** | 0.46 | **0.56** | 0.11 | 39 |
| | *wikinews_sub+VG-VIS combined* | 0.57 | **0.53** | 0.50 | **0.65** | 0.53 | **0.55** | 0.46 | **0.65** | 0.49 | **0.55** | 0.06 | 39 |
| | *wikinews_sub+VG-MM combined* | **0.69** | **0.59** | 0.40 | **0.70** | **0.59** | 0.50 | 0.50 | **0.65** | 0.42 | **0.56** | 0.1 | 39 |
| | *wikinews_sub+Google-VIS whole* | 0.43 | 0.42 | 0.50 | 0.38 | 0.45 | 0.38 | **0.58** | 0.43 | **0.59** | 0.46 | 0.08 | 39 |
| | *wikinews_sub+Google-MM whole* | 0.38 | 0.47 | 0.46 | 0.38 | 0.47 | 0.42 | **0.66** | 0.44 | **0.57** | 0.47 | 0.09 | 39 |
| | *crawl+Google ResNet-152* | 0.33 | 0.38 | **0.59** | 0.42 | 0.44 | 0.39 | **0.69** | 0.50 | **0.63** | 0.48 | 0.12 | 39 |
| | *crawl+Google AlexNet* | 0.42 | 0.41 | 0.47 | 0.49 | 0.49 | 0.40 | **0.65** | 0.48 | 0.55 | 0.48 | 0.07 | 39 |
| | *crawl+VG-VIS internal* | 0.63 | **0.62** | 0.41 | **0.67** | **0.59** | 0.56 | 0.50 | **0.59** | **0.60** | **0.57** | 0.07 | 39 |
| | *crawl+VG-MM internal* | 0.60 | 0.60 | 0.39 | **0.68** | **0.57** | **0.55** | 0.51 | **0.61** | **0.59** | **0.57** | 0.08 | 39 |
| | *crawl+VG-VIS combined* | **0.64** | **0.62** | 0.41 | **0.67** | **0.58** | **0.55** | **0.50** | **0.59** | **0.60** | **0.57** | 0.07 | 39 |
| | *crawl+VG-MM combined* | **0.60** | **0.62** | 0.43 | **0.65** | **0.58** | 0.59 | **0.55** | **0.62** | 0.54 | **0.58** | 0.06 | 39 |
| | *crawl+Google-VIS whole* | **0.48** | 0.49 | 0.52 | 0.46 | 0.55 | 0.36 | **0.65** | 0.39 | 0.52 | **0.49** | 0.08 | 39 |
| | *crawl+Google-MM whole* | **0.48** | 0.49 | **0.50** | 0.46 | 0.55 | 0.35 | **0.65** | 0.38 | 0.51 | **0.49** | 0.08 | 39 |
| | *w2v13+Google ResNet-152* | **0.87** | **0.67** | **0.69** | **0.73** | **0.60** | **0.61** | 0.56 | 0.58 | **0.69** | **0.67** | 0.09 | 39 |
| | *w2v13+Google AlexNet* | **0.74** | **0.64** | **0.60** | **0.67** | **0.58** | 0.54 | **0.62** | **0.61** | **0.71** | **0.63** | 0.06 | 39 |
| | *w2v13+VG-VIS internal* | 0.35 | **0.47** | 0.28 | 0.38 | 0.55 | 0.51 | **0.54** | 0.48 | 0.40 | 0.44 | 0.09 | 39 |
| | *w2v13+VG-MM internal* | 0.37 | 0.50 | 0.35 | **0.48** | 0.39 | 0.47 | **0.59** | 0.45 | 0.47 | 0.45 | 0.07 | 39 |
| | *w2v13+VG-VIS combined* | 0.36 | **0.47** | 0.29 | 0.37 | **0.55** | 0.50 | **0.54** | 0.47 | 0.41 | 0.44 | 0.08 | 39 |
| | *w2v13+VG-MM combined* | 0.35 | **0.51** | 0.39 | **0.54** | 0.42 | 0.39 | **0.57** | 0.45 | 0.43 | 0.45 | 0.07 | 39 |
| | *w2v13+Google-VIS whole* | **0.76** | **0.57** | **0.61** | **0.66** | **0.61** | **0.53** | **0.49** | **0.60** | **0.69** | **0.61** | 0.08 | 39 |
| | *w2v13+Google-MM whole* | **0.75** | **0.56** | **0.60** | **0.67** | **0.61** | **0.53** | **0.48** | **0.61** | **0.68** | **0.61** | 0.08 | 39 |
| $E_L + E_S$ | *wikinews+VG SceneGraph* | 0.48 | **0.51** | 0.47 | 0.49 | 0.50 | 0.50 | **0.59** | 0.32 | 0.48 | 0.48 | 0.07 | 39 |
| | *wikinews_sub+VG SceneGraph* | **0.54** | **0.55** | 0.49 | 0.50 | 0.40 | **0.51** | **0.58** | 0.45 | **0.59** | **0.51** | 0.06 | 39 |
| | *crawl+VG SceneGraph* | **0.56** | **0.59** | 0.43 | 0.47 | 0.43 | 0.47 | **0.59** | 0.38 | 0.49 | 0.49 | 0.07 | 39 |
| | *w2v13+VG SceneGraph* | **0.59** | 0.49 | **0.63** | 0.48 | 0.45 | **0.60** | 0.42 | **0.53** | **0.67** | **0.54** | 0.08 | 39 |

Table 4.10: fMRI scores for each participant and embedding on the common subset of vocabularies. Multi-modal embeddings are created using the Intersection technique. The table sections contain linguistic, visual and multi-modal embeddings in this order. Red colour signifies the best performance, blue means that the multi-modal embedding outperformed the corresponding uni-modal ones.

| Modality | Embedding | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 | Avg | STD | Coverage |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $E_L$ | *w2v13* | 0.56 | 0.52 | 0.36 | 0.46 | 0.34 | 0.55 | 0.50 | 0.44 | 0.51 | 0.47 | 0.07 | 39 |
| | *wikinews_sub* | 0.55 | 0.64 | 0.40 | 0.52 | 0.59 | 0.42 | 0.62 | 0.46 | 0.53 | 0.53 | 0.08 | 39 |
| | *wikinews* | 0.59 | 0.62 | 0.38 | 0.39 | 0.64 | 0.42 | 0.62 | 0.37 | 0.65 | 0.52 | 0.12 | 39 |
| | *crawl* | 0.50 | 0.45 | **0.66** | 0.41 | 0.60 | 0.40 | 0.55 | 0.54 | 0.45 | 0.51 | 0.08 | 39 |
| $E_V$ | *Google-VIS whole* | 0.49 | 0.52 | 0.56 | 0.35 | 0.44 | 0.64 | 0.45 | 0.65 | 0.52 | 0.51 | 0.09 | 39 |
| | *Google ResNet-152* | 0.56 | 0.55 | 0.65 | 0.24 | 0.38 | 0.60 | 0.50 | 0.60 | 0.45 | 0.50 | 0.12 | 39 |
| | *VG-VIS internal* | 0.46 | 0.54 | 0.51 | 0.53 | 0.66 | 0.46 | 0.49 | 0.52 | 0.65 | 0.54 | 0.07 | 39 |
| | *Google AlexNet* | 0.35 | 0.52 | 0.54 | 0.45 | 0.52 | 0.52 | 0.53 | 0.51 | 0.50 | 0.49 | 0.05 | 39 |
| | *VG-VIS combined* | 0.33 | 0.44 | 0.49 | 0.62 | 0.68 | 0.46 | 0.49 | 0.47 | 0.54 | 0.50 | 0.1 | 39 |
| $E_S$ | *VG SceneGraph* | 0.48 | 0.60 | 0.49 | 0.49 | 0.53 | 0.54 | 0.59 | 0.39 | 0.49 | 0.51 | 0.06 | 39 |
| $E_L + E_V$ | *VG-MM internal* | 0.50 | 0.22 | 0.50 | 0.55 | 0.39 | 0.54 | 0.50 | 0.54 | 0.52 | 0.47 | 0.1 | 39 |
| | *VG-MM combined* | 0.29 | 0.36 | 0.54 | 0.45 | 0.40 | 0.38 | 0.35 | 0.39 | 0.50 | 0.41 | 0.07 | 39 |
| | *Google-MM whole* | 0.39 | 0.51 | 0.48 | 0.29 | 0.40 | 0.54 | 0.46 | 0.65 | 0.57 | 0.48 | 0.1 | 39 |
| | *wikinews+Google ResNet-152* | 0.46 | 0.44 | 0.57 | **0.43** | 0.64 | 0.53 | 0.35 | **0.61** | 0.45 | 0.50 | 0.09 | 39 |
| | *wikinews+Google AlexNet* | 0.52 | 0.36 | 0.42 | **0.46** | 0.48 | **0.57** | 0.31 | **0.63** | 0.58 | 0.48 | 0.1 | 39 |
| | *wikinews+VG-VIS internal* | 0.39 | 0.34 | 0.46 | **0.58** | 0.49 | **0.50** | 0.52 | **0.57** | **0.67** | 0.50 | 0.09 | 39 |
| | *wikinews+VG-MM internal* | 0.41 | 0.33 | 0.49 | **0.59** | 0.47 | 0.51 | 0.41 | **0.57** | 0.64 | 0.49 | 0.09 | 39 |
| | *wikinews+VG-VIS combined* | 0.39 | 0.34 | 0.46 | 0.58 | 0.48 | **0.49** | 0.51 | **0.57** | **0.66** | 0.50 | 0.09 | 39 |
| | *wikinews+VG-MM combined* | 0.42 | 0.38 | 0.51 | **0.59** | 0.49 | **0.53** | 0.42 | **0.59** | 0.65 | 0.51 | 0.09 | 39 |
| | *wikinews+Google-VIS whole* | 0.42 | 0.38 | 0.48 | 0.31 | 0.50 | **0.66** | 0.55 | 0.59 | 0.45 | 0.48 | 0.1 | 39 |
| | *wikinews+Google-MM whole* | 0.41 | 0.38 | 0.47 | 0.31 | 0.48 | **0.66** | 0.56 | 0.59 | 0.42 | 0.48 | 0.1 | 39 |
| | *wikinews_sub+Google ResNet-152* | 0.46 | 0.44 | 0.57 | 0.43 | **0.64** | 0.53 | 0.35 | **0.61** | 0.45 | 0.50 | 0.09 | 39 |
| | *wikinews_sub+Google AlexNet* | 0.52 | 0.35 | 0.42 | 0.46 | 0.48 | **0.57** | 0.31 | **0.63** | **0.58** | 0.48 | 0.1 | 39 |
| | *wikinews_sub+VG-VIS internal* | 0.54 | 0.40 | 0.40 | 0.47 | 0.52 | 0.43 | 0.59 | **0.54** | 0.56 | 0.49 | 0.07 | 39 |
| | *wikinews_sub+VG-MM internal* | 0.51 | 0.43 | 0.44 | 0.51 | 0.48 | **0.55** | 0.53 | **0.55** | **0.66** | 0.52 | 0.06 | 39 |
| | *wikinews_sub+VG-VIS combined* | 0.52 | 0.40 | 0.40 | 0.50 | 0.49 | 0.45 | 0.58 | **0.53** | **0.57** | 0.49 | 0.06 | 39 |
| | *wikinews_sub+VG-MM combined* | 0.50 | 0.48 | 0.50 | **0.56** | 0.48 | **0.56** | 0.52 | **0.57** | **0.65** | 0.53 | 0.05 | 39 |
| | *wikinews_sub+Google-VIS whole* | 0.48 | 0.39 | 0.46 | 0.40 | 0.58 | 0.60 | 0.42 | 0.63 | 0.48 | 0.49 | 0.08 | 39 |
| | *wikinews_sub+Google-MM whole* | 0.44 | 0.37 | 0.45 | 0.39 | 0.53 | **0.61** | 0.46 | 0.57 | 0.44 | 0.47 | 0.08 | 39 |
| | *crawl+Google ResNet-152* | 0.47 | 0.44 | 0.57 | **0.43** | **0.63** | 0.53 | 0.36 | **0.61** | 0.45 | 0.50 | 0.09 | 39 |
| | *crawl+Google AlexNet* | **0.52** | 0.35 | 0.42 | **0.46** | 0.48 | **0.57** | 0.31 | **0.63** | **0.58** | 0.48 | 0.1 | 39 |
| | *crawl+VG-VIS internal* | 0.43 | 0.35 | 0.46 | 0.50 | 0.50 | 0.45 | 0.54 | 0.50 | 0.60 | 0.48 | 0.07 | 39 |
| | *crawl+VG-MM internal* | 0.45 | 0.34 | 0.47 | 0.49 | 0.51 | 0.40 | 0.48 | 0.49 | **0.61** | 0.47 | 0.07 | 39 |
| | *crawl+VG-VIS combined* | 0.42 | 0.35 | 0.46 | 0.50 | 0.49 | 0.45 | 0.54 | 0.50 | **0.60** | 0.48 | 0.07 | 39 |
| | *crawl+VG-MM combined* | 0.49 | 0.44 | 0.50 | **0.50** | 0.54 | **0.42** | 0.47 | 0.50 | **0.65** | 0.50 | 0.06 | 39 |
| | *crawl+Google-VIS whole* | **0.57** | 0.42 | 0.43 | 0.32 | 0.48 | **0.67** | **0.60** | 0.60 | **0.55** | **0.52** | 0.11 | 39 |
| | *crawl+Google-MM whole* | **0.57** | 0.42 | 0.43 | 0.31 | 0.47 | **0.67** | **0.61** | 0.60 | 0.55 | **0.52** | 0.11 | 39 |
| | *w2v13+Google ResNet-152* | **0.67** | **0.61** | 0.62 | **0.65** | **0.73** | **0.73** | **0.67** | 0.58 | **0.70** | **0.66** | 0.05 | 39 |
| | *w2v13+Google AlexNet* | 0.53 | 0.48 | **0.57** | **0.65** | **0.61** | **0.68** | **0.59** | 0.50 | **0.57** | **0.58** | 0.06 | 39 |
| | *w2v13+VG-VIS internal* | 0.52 | 0.46 | 0.51 | **0.56** | 0.45 | 0.54 | 0.48 | **0.68** | 0.53 | 0.53 | 0.07 | 39 |
| | *w2v13+VG-MM internal* | 0.55 | 0.42 | **0.53** | **0.56** | **0.43** | 0.49 | 0.47 | **0.67** | 0.42 | **0.50** | 0.08 | 39 |
| | *w2v13+VG-VIS combined* | 0.52 | 0.46 | **0.51** | 0.56 | 0.44 | 0.54 | 0.48 | **0.68** | 0.52 | **0.52** | 0.07 | 39 |
| | *w2v13+VG-MM combined* | 0.55 | 0.44 | 0.49 | **0.55** | **0.54** | **0.56** | **0.56** | **0.68** | 0.47 | **0.54** | 0.06 | 39 |
| | *w2v13+Google-VIS whole* | **0.66** | **0.59** | **0.59** | **0.55** | **0.68** | **0.74** | **0.66** | 0.47 | **0.72** | **0.63** | 0.08 | 39 |
| | *w2v13+Google-MM whole* | **0.66** | **0.59** | **0.59** | **0.52** | **0.66** | **0.72** | **0.66** | 0.45 | **0.72** | **0.62** | 0.08 | 39 |
| $E_L + E_S$ | *wikinews+VG SceneGraph* | 0.56 | **0.65** | **0.53** | **0.56** | 0.45 | 0.33 | **0.64** | **0.52** | 0.41 | 0.52 | 0.1 | 39 |
| | *wikinews_sub+VG SceneGraph* | **0.62** | **0.67** | **0.62** | **0.56** | 0.49 | 0.35 | 0.60 | 0.44 | 0.46 | **0.54** | 0.1 | 39 |
| | *crawl+VG SceneGraph* | **0.69** | 0.52 | 0.52 | 0.48 | 0.42 | **0.55** | **0.67** | **0.66** | 0.45 | **0.55** | 0.09 | 39 |
| | *w2v13+VG SceneGraph* | 0.52 | 0.49 | **0.54** | **0.54** | 0.40 | 0.55 | 0.52 | **0.51** | 0.50 | 0.51 | 0.04 | 39 |

Table 4.11: MEG scores for each participant and embedding on the common subset of vocabularies. Multi-modal embeddings are created using the Intersection technique. The table sections contain linguistic, visual and multi-modal embeddings in this order. Red colour signifies the best performance, blue means that the multi-modal embedding outperformed the corresponding uni-modal ones.

| Modality | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 |
|---|---|---|---|---|---|---|---|---|---|
| $E_L$ | 0.83 | 0.64 | 0.65 | **0.82** | 0.62 | 0.58 | 0.55 | 0.55 | 0.65 |
| $E_V$ | **0.87** | 0.66 | 0.65 | 0.74 | 0.51 | 0.58 | **0.61** | 0.50 | 0.61 |
| $E_S$ | 0.83 | 0.68 | 0.57 | 0.77 | 0.59 | **0.63** | 0.58 | 0.59 | 0.64 |
| $E_L + E_V$ | 0.86 | 0.65 | **0.66** | 0.79 | 0.60 | 0.59 | 0.60 | 0.54 | 0.64 |
| $E_L + E_S$ | 0.86 | **0.69** | 0.60 | 0.81 | **0.64** | 0.63 | 0.57 | **0.60** | **0.67** |

Table 4.12: fMRI scores averaged over each modality. Bold signifies the highest average performance for each participant.

| Modality | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 |
|---|---|---|---|---|---|---|---|---|---|
| $E_L$ | 0.64 | 0.60 | 0.53 | 0.69 | 0.71 | 0.64 | 0.73 | **0.63** | 0.72 |
| $E_V$ | **0.66** | 0.54 | 0.55 | 0.68 | 0.68 | **0.72** | 0.67 | 0.57 | 0.66 |
| $E_S$ | 0.63 | 0.60 | 0.55 | 0.65 | 0.70 | 0.62 | 0.67 | 0.50 | 0.73 |
| $E_L + E_V$ | 0.66 | 0.62 | **0.56** | **0.69** | 0.73 | 0.68 | 0.71 | 0.60 | 0.71 |
| $E_L + E_S$ | 0.65 | **0.62** | 0.55 | 0.68 | **0.74** | 0.64 | **0.73** | 0.57 | **0.74** |

Table 4.13: MEG scores averaged over each modality. Bold signifies the highest average performance for each participant.

| Modality | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 |
|---|---|---|---|---|---|---|---|---|---|
| $E_L$ | 0.46 | 0.45 | 0.45 | 0.50 | 0.50 | 0.48 | 0.48 | 0.47 | 0.54 |
| $E_V$ | 0.54 | 0.40 | **0.57** | **0.56** | 0.52 | **0.53** | 0.47 | **0.56** | 0.48 |
| $E_S$ | 0.51 | 0.45 | 0.49 | 0.55 | **0.63** | 0.38 | 0.55 | 0.48 | 0.36 |
| $E_L + E_V$ | 0.53 | 0.53 | 0.47 | 0.55 | 0.53 | 0.48 | **0.56** | 0.54 | 0.54 |
| $E_L + E_S$ | **0.54** | **0.53** | 0.50 | 0.48 | 0.45 | 0.52 | 0.54 | 0.42 | **0.56** |

Table 4.14: fMRI scores averaged over each modality on the common subset of vocabularies. Bold signifies the highest average performance for each participant.

| Modality | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 |
|---|---|---|---|---|---|---|---|---|---|
| $E_L$ | 0.55 | 0.56 | 0.45 | 0.44 | **0.55** | 0.45 | 0.57 | 0.45 | 0.53 |
| $E_V$ | 0.44 | 0.51 | 0.55 | 0.44 | 0.54 | 0.54 | 0.49 | 0.55 | 0.53 |
| $E_S$ | 0.48 | **0.60** | 0.49 | 0.49 | 0.53 | 0.54 | 0.59 | 0.39 | 0.49 |
| $E_L + E_V$ | 0.49 | 0.41 | 0.49 | 0.48 | 0.52 | **0.55** | 0.49 | **0.57** | **0.56** |
| $E_L + E_S$ | **0.60** | 0.58 | **0.55** | **0.53** | 0.44 | 0.45 | **0.61** | 0.53 | 0.45 |

Table 4.15: MEG scores averaged over each modality on the common subset of vocabularies. Bold signifies the highest average performance for each participant.

| **Architecture** | **Embedding** | **Accuracy** (%) |
|---|---|---|
| Add | *linguistic only* | **77.54** |
| | *visual only* | 72.70 |
| | *multi-modal* | 76.56 |
| | *random* | 69.87 |
| Add+Translation | *linguistic only* | 81.21 |
| | *visual only* | 79.75 |
| | *multi-modal* | **81.81** |
| | *random* | 78.33 |
| Add+Translation+FullVis | *linguistic only* | 79.85 |
| | *visual only* | 79.11 |
| | *multi-modal* | **81.29** |
| | *random* | 78.79 |
| GRU | *linguistic only* | **79.77** |
| | *visual only* | 77.34 |
| | *multi-modal* | 79.48 |
| | *random* | 79.25 |
| LSTM | *linguistic only* | **79.80** |
| | *visual only* | 78.22 |
| | *multi-modal* | 79.61 |
| | *random* | 76.16 |

Table 4.16: Classification accuracy of the different architectures and embedding initialisations.

# Chapter 5

# Effects of Data Size and Distribution

This chapter shifts the focus towards a more in-depth analysis of some selected model, data source and modality combination based on the results of the previous chapter. Our main metric is still *performance* accuracy, thus this analysis forms the last part of pillar 1.

We aim our attention at studying model *efficiency* regarding size and performance. In this study we dig deeper into the effect of the training data size and distribution. The presented experiments address the following questions:

- Does visual data bolster performance only because we add more data or does it convey complementary *quality* information compared to a higher *quantity* of text? (Question 4)

- Can we achieve comparable performance using small-data if it comes from the right data distribution? (Question 4a)

We perform different experiments in order to test the effect of data size and data distribution on semantic similarity and relatedness tasks. We will compare linguistic, visual and structured embeddings, based on various criteria.

## 5.1 Counting in the "Effort"

The work presented here is related to a recently published information theoretical probing framework based on minimal description length (MDL) [Voita and Titov, 2020] i.e. the minimum number of bits needed to transmit the labels knowing the representations. Our idea is to count in the "effort" of data collection and quantity into the performance of our multi-modal word meaning representations.

Unlike Voita et al., instead of testing on supervised tasks, we focus on unsupervised evaluation. We do not train a multi-layered perception for probing. This is relevant because this way we avoid distorting our results by a network functioning as supervised fine tuning. In Section 4.4 we found that a shallow neural network and a deep LSTM, both with randomly initialised input word vectors, perform on par with an input of pretrained word embeddings on a Textual Entailment task (SNLI). Zhang and Bowman found the related phenomenon of high performing random initialized LSTM models [Zhang and Bowman, 2018]. This is in line with current findings considering the recent transformer type models which are shown to be far from solving general tasks (e.g., document question answering). Rather, these models are overfitting to the quirks of particular datasets

[Yogatama et al., 2019]. Motivated by these results, in this work we decided to focus on diving into unsupervised representation learning.

In unsupervised representation learning we are learning $P(x)$ instead of $P(y|x)$, where $x$ is the input data, $y$ is the corresponding label determined by the supervised evaluation task. Hence, our approach is more related to Voita et al.'s MDL framework with "online" code where the code length is simply calculated by the entropy of the training data.

We pursue measuring how hard it is to achieve a high performing representation with small data. In the previous chapter we controlled for image quantity for $D_V$ (Section 4.1) and the context size (radius) of $D_S$ (Section 4.2). In this chapter we focus on controlling for text data size and distribution $D_L$. Our question is: What is the corpus size where visual information is helpful? We count in the "effort" by discussing performance in the context of data and model size. In the following, we describe our implementation of controlling for data quantity and word frequency distribution.

## 5.2 Experiments

Here, we summarise the notation and specify the models used in the following experiments, based on our previous findings in Chapter 4.

$E_L \in \mathbb{R}^{|T| \times d_L}$: *Linguistic Embedding.* Here, we present results using Skip-Gram with Negative Sampling (SGNS) [Mikolov et al., 2013a, Mikolov et al., 2013b] trained on a 2020 English Wikipedia dump. Due to its simplicity, it is suitable for running a wide range of experiments.

$E_V \in \mathbb{R}^{|T| \times d_V}$: *Visual Embedding.* We ran a feedforward step of *ResNet-152* [He et al., 2016] on Google Images. We apply mean aggregation on the first 10 image results which has been found on of the best performing in Section 4.1.

$E_S \in \mathbb{R}^{|T| \times d_S}$: *Structured Embedding.* We use our in-between visual and linguistic embedding, trained on the visually structured text of Visual Genome Scene Graphs (Section 4.2).

In the following we show results according to $e_1, \ldots, e_l$ samples from the linguistic training corpus $D_L$. $T = |V \cap V_{task}| \approx |V_{task}|, V_{task} \subset V$, where $V$ is the vocabulary of the text corpus and $V_{task}$ is the vocabulary of the evaluation tasks.

### 5.2.1 Control for Data Quantity

We perform experiments where we restrict the training data size of $E_L$. Similarly to Sahlgren et al [Sahlgren and Lenci, 2016], we sample the corpora randomly to subsets with increasing number of tokens: $e_1, \ldots, e_N$.

### 5.2.2 Control for Frequency Ranges

In the second phase we can test how models, trained on different word frequency ranges, interact with the other types of embeddings. Similarly to [Sahlgren and Lenci, 2016] we split the vocabulary into three equally large parts; HIGH, MEDIUM and LOW range. This way we generate samples for $E_L, E_V$ and $E_S$ for the different frequency ranges in the text corpus.

### 5.2.3 Expected Results

These experiments will potentially shed light to patterns across modalities and sources. One interesting result will be to see whether $E_V$ and $E_S$ embeddings contribute more if there is smaller amount of text data for $E_L$. If this is the case, the experiments where we control for word frequencies can reveal whether $E_V$ and $E_S$ contribute differently for words with different data distributions, or whether the effect is more due to data quantity. Similar questions can be answered in the reverse direction when we perform experiments where we control for image data size and distributional properties, such as image resolution or dispersion of image sets.

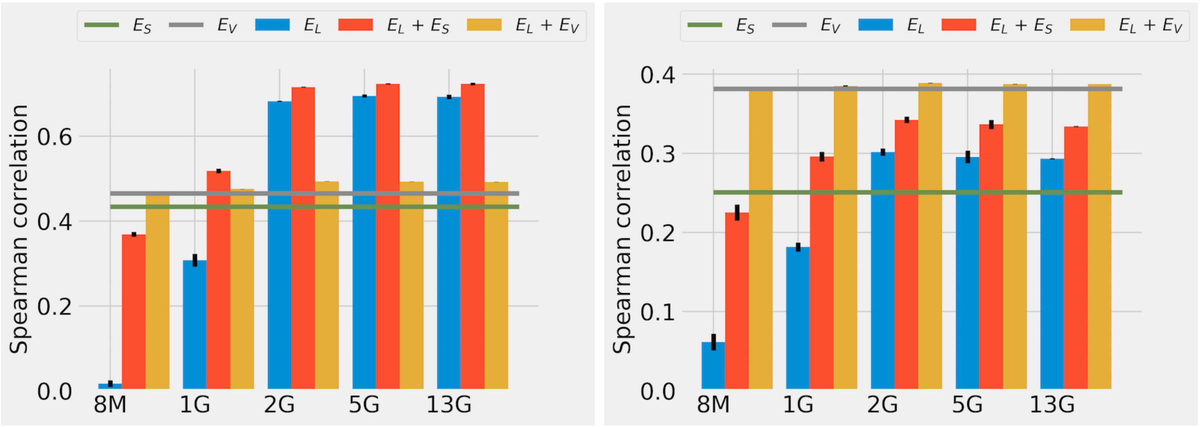
### 5.2.4 Results

Figure 5.1 shows the effect of $E_L$ corpus size on the performance of uni-modal $E_L$ and the combined $E_L + E_S$ and $E_L + E_V$ on the embeddings' common coverage subsets of MEN (Figure 5.1a) and SimLex (Figure 5.1b). The common coverage is 73% on MEN and 56% on SimLex. $E_S$ and $E_V$ are constant since only $E_L$'s training data is varied. Results on the full datasets are presented in Figure 5.2. Axis $x$ represents the size of the training corpus (in the number of tokens). Error bars indicate variance after three runs of random down-sampling of the data. Table 5.1 gives an account of the amount of training data each model requires. The last line shows the size after compression by Lempel-Ziv coding (LZ77). Since ImageNet images are already in jpg format, LZ77 was not able to achieve any further compression.

The first striking result is that $E_S$ alone, with ∼9M tokens, outperforms $E_L$, with ∼1G tokens, on both evaluation tasks. Secondly, when combined with linguistic data, $E_S$ greatly outperforms $E_V$ on MEN and underperforms it on SimLex, however, their difference becomes marginal as text data increases. Importantly, $E_S$ achieves this result with orders of magnitude less data than required by $E_V$ (Table 5.1). Moreover, ResNet-152 with ∼6.8G parameters outputs a 1.7 times bigger model (4.8MB) than SGNS, used for $E_L$ and $E_S$ (2.8MB), consisting of 151,200 parameters. A summary of model sizes is included in Table 5.2 for the common subset of their vocabularies of 1203 words.

Figure 5.2c and 5.2d report the effect of word frequency on performance on the same tasks. Similarly to [Sahlgren and Lenci, 2016] we split the vocabulary into three equally large parts; HIGH, MEDIUM and LOW range. On MEN we see a slight performance gain of the baseline $E_L$ model on medium range frequency words, whereas on SimLex, low frequency words dominate the performance within the whole data (MIXED). On SimLex visual information helps more with HIGH frequency words. This could be due to narrowing down the meaning of ambiguous words. Checking this hypothesis would be an interesting future analysis.

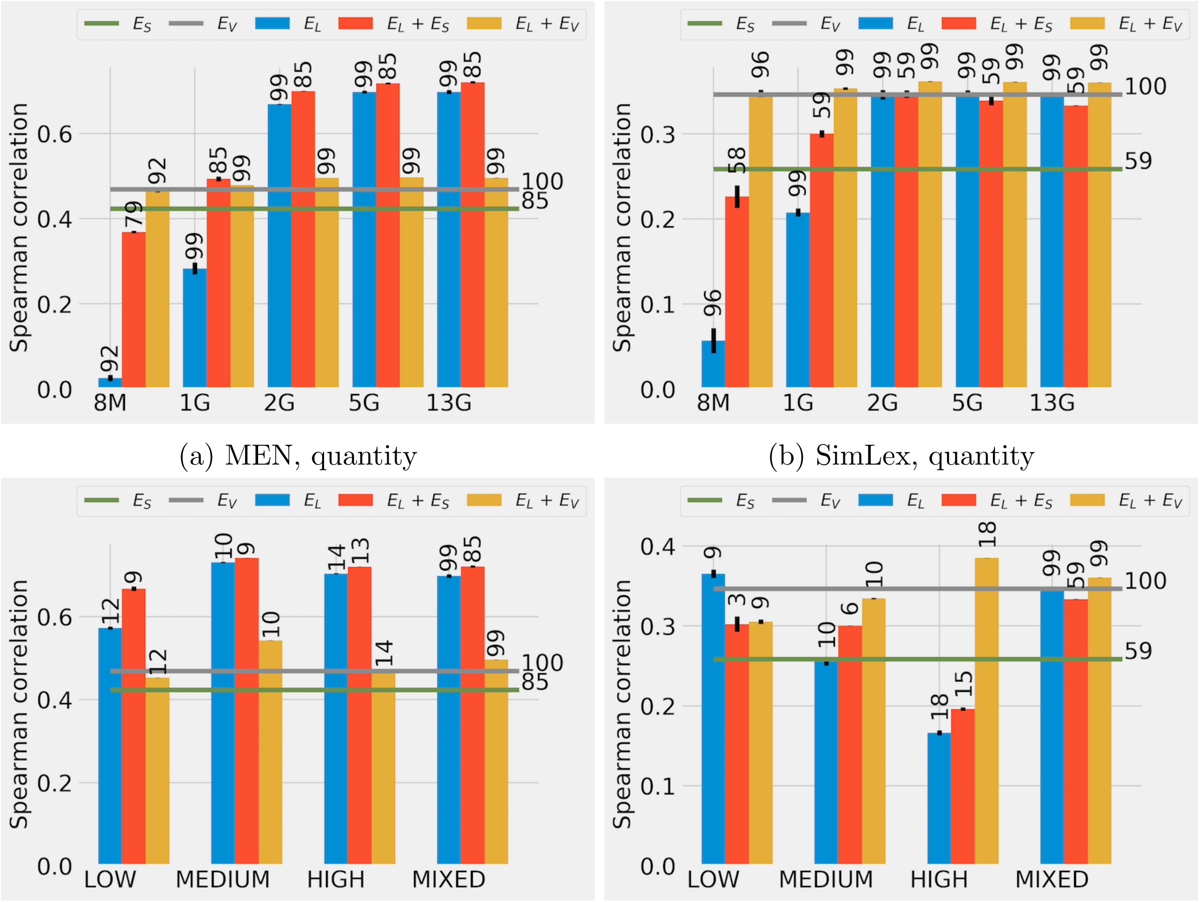
$E_S$ performs similarly to the FastText VG description model of [Herbelot, 2020] on SimLex. The increase of $E_L$ performance is in line with [Sahlgren and Lenci, 2016] until 2G tokens (they stopped at 1G), after which it plateaus. The best Spearman correlation of [Kuzmenko and Herbelot, 2019] using relations on MEN is 0.5499, with almost third the coverage (847) of ours on the common subset: $E_S$ achieves 0.44 with a coverage of 2481. Their *word2vec* model is consistent with results reported by [Sahlgren and Lenci, 2016] and our *word2vec* based $E_L$ model with similar amount of data.



(a) MEN

(b) SimLex

Figure 5.1: Effect of $E_L$ training corpus (token) quantity on performance on the common coverage subsets of evaluation pairs (73% on MEN, 56% on SimLex). $E_S$ and $E_V$ are constant since only $E_L$'s training data is varied.



(a) MEN, quantity

(b) SimLex, quantity

(c) MEN, frequency

(d) SimLex, frequency

Figure 5.2: Effect of $E_L$ training corpus quantity and word frequency on performance. Numbers on top of the bars and on the lines indicate the coverage of evaluation dataset pairs (where both words are in the embedding vocabulary) in percentages. $E_S$ and $E_V$ are constant since only $E_L$'s training data is varied.

|  | $E_L$ | $E_S$ | $E_V$ |
|---|---|---|---|
| Model | SGNS | SGNS | ResNet-152 |
| Training data | Wikipedia 2020 | Visual Genome annotations | ImageNet + Google Images |
| Size in units | 13G tokens | ∼9M tokens | ∼1.28M + 15,770 images (jpg) |
| Storage size | 14GB | ∼1.8GB | ∼140GB |
| Compressed size | ∼5GB | ∼0.2GB | ∼140GB |

Table 5.1: Training data sizes.

|  | $E_L$ | $E_S$ | $E_V$ |
|---|---|---|---|
| Model | SGNS | SGNS | ResNet-152 |
| Number of model parameters | 151,200 | 151,200 | 6.8G |
| Embedding size | 2.8MB | 2.8MB | 4.8MB |

Table 5.2: Model sizes on the common subset of vocabularies ($|V_{common}| = 1203$).

## 5.3 Conclusion

Overall, we conclude that our structured visuo-linguistic embedding contributes to a linguistic model in a much more economic way than the image based ones. We saw that when the linguistic sources are limited, visual or structured information can greatly improve on semantic similarity and relatedness predictions. As the volume of our text corpus increases, both its usefulness plateaus as well as the performance gain using other modalities shrinks, however, in most cases some improvement remains. These findings suggest that in certain cases one can save valuable training time and storage space by balancing the trade-off between training on different modalities or acquiring more text data.

Our structured embedding trained on Visual Genome Scene Graph requires orders of magnitude less data than either of the other two modalities, still contributing substantially to the meaning representation. This may be due to the amount of human effort had been made while creating the dataset. Applying automatically generated scenes graphs [Xu et al., 2020] would mitigate this problem. This would serve as a highly effective tool with important applications for low resource languages. Our findings support the intuition of "no free lunch" when it comes to effort, but depending on the tasks in hand and the available resources it can be crucial to optimise the types of resources we use. Here we only focused on data and model size. Including processing time and costs would be an important future extension of efficiency analysis.

Exactly how $E_S$ contributes to the linguistic $E_L$ representation cannot be interpreted based solely on performance metrics. Therefore, we investigate the interpretation of our representations and the type of information they convey in the next Chapter.

# Chapter 6

# Informativeness of Semantic Spaces

In this chapter, we introduce the third key contribution of this thesis (Chapter 1.1), presenting proof-of-concept studies of interpretable *Transparency analysis*. We present experiments demonstrating pillars 2 *Qualitative / Quantitative structural analysis* and 3 *Independence analysis*.

We aim to take the systematic studies in Chapter 4 and 5 a step further, and perform quantitative and qualitative comparison of embedding space structures. We showcase an implementation in the framework of modalities as partial observers of meaning, introduced in Section 2.7.

Section 6.1 introduces our two hypotheses. In Section 6.2 we tackle Question 5: *Can we move beyond performance evaluation? Are there any emergent concepts in embeddings? Can we quantify the difference between the concept structures of semantic spaces?* We hypothesise that each embedding space represents clusters of word representations which can be interpreted as each embeddings' own "idea" of concepts in the world. They can "disagree" depending on the data distributions of the specific modality and data source they were trained on. By zooming into our embeddings' structure we aim to find out how much their models of concepts differ from each other if they differ at all. We are looking for quantitative ways of measuring the difference between embedding spaces to complement the qualitative analysis.

Section 6.3 addresses Question 6: *Can we quantify the difference between semantic spaces, based on the useful information they contribute to the meaning representation?* We apply an information-theoretical framework laid out in Section 2.7.5 to estimate Mutual Information of two semantic spaces using methods described in Section 3.2.4.

Finally, Section 6.4 investigates the results in the context of distributional properties of the linguistic and structured data sources, $D_L$ and $D_S$.

Our main contribution is a proof-of-concept framework for quantifying the information different data sources, models and modalities bring into multi-modal word representations. It can easily be applied to various more data, model or modality types beyond the ones showcased in this study. These set of methods can help us looking under the hood of accuracy numbers on evaluation tasks and understanding better how these different concept models interact with each other when they are combined in multi-modal models of word meaning.

## 6.1 Hypotheses

Within our generalised embedding framework (Section 2.6) we use the same models as in Section 5.2. We propose investigating the structure of the learnt embedding spaces

$E_L, E_V, E_S$. This aspires to qualitatively compare embedding spaces according to various metrics. These metrics aim to capture the distributional properties of vector spaces. Furthermore, we put the results in the context of analysing the training data distributions.

Based on our previous findings we form the following hypotheses:

I. $E_V$ can be complementary to $E_L$ when the training corpus size is small. It is not clear whether in this case $E_V$ comes from a different and complementary distribution or the performance gain is only relative to the size of the additional data. In this case, we would achieve the same result with training on the same amount of additional text.

II. Due to the manufactured way of collecting data for $E_S$, it is possible that this dataset comes from a substantially different distribution than our linguistic data. Therefore, it can provide useful information and can facilitate learning from small data.

## 6.2 Qualitative Analysis of Semantic Spaces

As described in Section 3.2.3, in order to grasp how the concept structure of our embedding spaces differ from each other we first searched for ways to quantify their cluster structure. We do not know the ground truth labels of our clusters or even the number of clusters each embedding spaces should be broken into. Therefore, in Section 6.2.1 we present the results of experiments with three clusterization metrics which are designed for the case when a ground truth labelling is not available. Furthermore, we report results for a range of number of clusters.

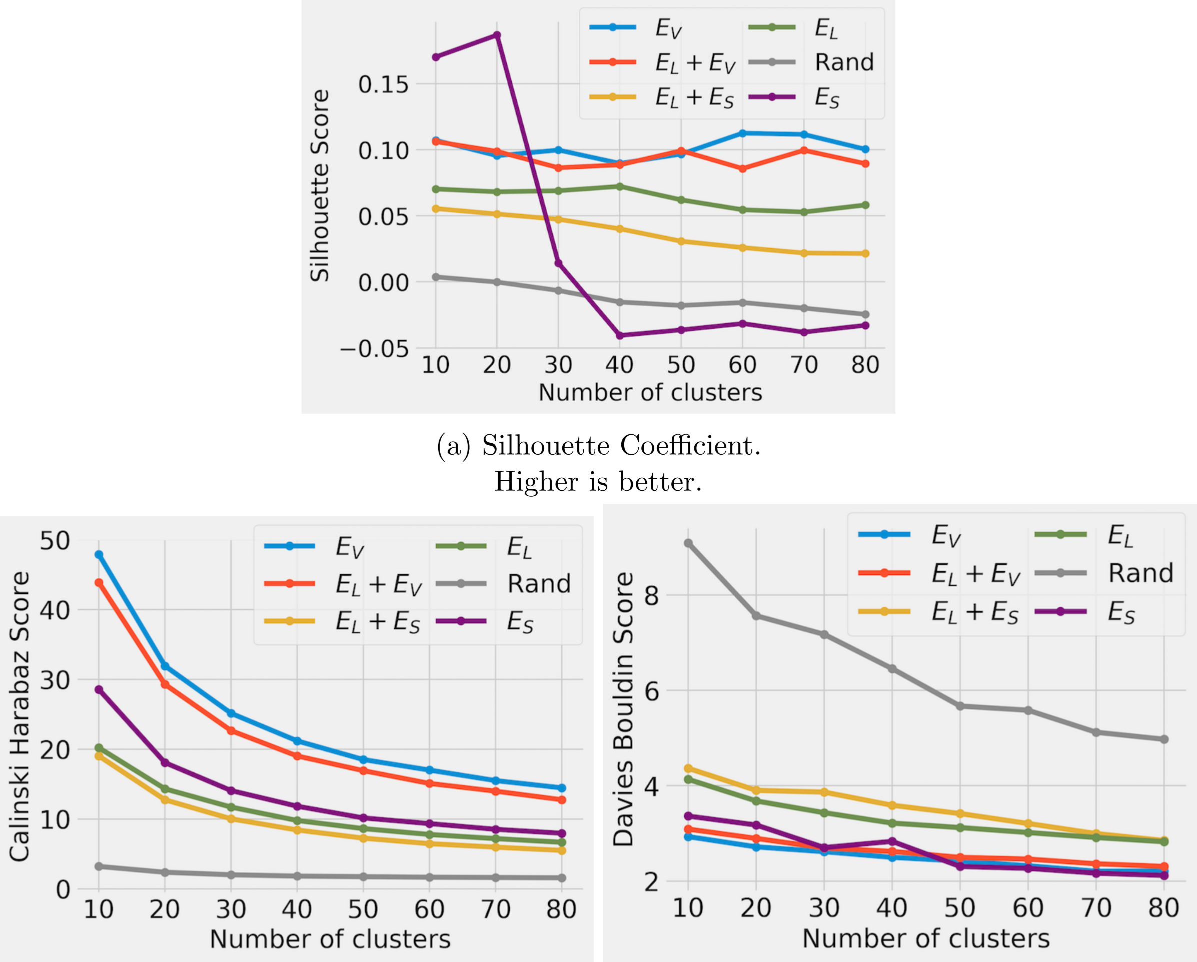
Following the desire of interpreting how our different models conceptualise, in Sections 6.2.2 and 6.2.3 we zoom into our embedding spaces even further. In Section 6.2.2 we compare our embeddings' cluster structures and visualise the learnt clusterings. In Section 6.2.3 we present supervised visualisations of the embedding spaces alongside an automatic label generation method and compare the results against the clusterization metric scores.

### 6.2.1 Cluster Structure Results

Clustering metrics results are presented for increasing numbers of clusters, using K-means clustering in Figure 6.1 (See the definition of metrics in Section 3.2.3). We compare the common subset of our embedding vocabularies, resulting in 1204 words. Calinski-Harabasz Index and Davies-Bouldin Index score results (Figure 6.1c and 6.1b) are fairly consistent with each other, while we see a different pattern on Silhouette Coefficient in Figure 6.1a. This is unsurprising since the first two are based on node and centroid distances, whereas the latter calculates distances solely between nodes in the space.

In Davies-Bouldin Index (Figure 6.1c) all models significantly outperform the baseline Random embedding $E_R \in \mathbb{R}^{|V_{common}| \times 300}$. All models achieve similar scores with the visual, the structured and linguistic-visual multi-modal models performing the best. This index represents the ratio between intra-cluster distances from the centroids and inter-cluster distances of centroids.

Calinski-Harabasz Index scores (Figure 6.1b) show a similar tendency among the models, having $E_V$ and $E_L + E_V$ as best performing across the number of clusters, while all models overcome the Random baseline. As the number of clusters grow the results converge to a lower (worse) score. This score can be interpreted as a measurement of how well



(a) Silhouette Coefficient.
Higher is better.

(b) Calinski-Harabasz Index.
Higher is better.

(c) Davies-Bouldin Index.
Lower is better.

Figure 6.1: Clustering metrics for increasing number of K-means clusters.

defined the clusters are in terms of the ratio between inter- and intra-cluster dispersions, therefore a higher score means better defined clusters.

Silhouette Coefficient measures pairwise distances of data points within their own clusters and between each point's distance to data points in other clusters. It gives a ratio of cluster cohesion and separation. In 6.1a we see a similar tendency across models (having $E_V$ as the best) as before with the exception of the structured model $E_S$. It outperforms all models up to ∼20 clusters then drops below the Random baseline by 40. Furthermore, all the other models do not converge as in the previous two cases. This suggests that $E_S$ has much more cohesive structure of ∼20 clusters, but becomes in-cohesive if we try and break it into more clusters. This phenomenon might be related to the statistical properties of the Visual Genome dataset $E_S$ is trained on. In the original paper [Krishna et al., 2016] the authors report results on clustering region descriptions. They found that on average, each image contains descriptions from 17 different clusters, the image with the most diverse descriptions contains descriptions from 26 clusters. Unlike our model, they clustered averaged pertained word representations of region descriptions, therefore, their results are not directly comparable to ours. Nevertheless, we think this can indicate why this dramatic drop occurs at around 20 clusters in our experiments.

### 6.2.2 Inspecting the Clusters

In the following we inspect the individual clusters in all three embeddings after clustering them for 20 clusters. We also look at $E_S$ after clustering it for 40 clusters, where the drop in Silhouette Coefficient happens.

#### Size Distribution and Visualisation

In Figure 6.2 we present the distribution of cluster sizes (number of cluster members) for each cases. Firstly, we observe that $E_L$ and $E_V$ cluster sizes move between 10 and ~100, whereas in both cases $E_S$ cluster size distribution ranges between 1 and ~400. In the $E_S$ 20 clusters case (Figure 6.2a) most clusters range between 10 and 117, there are two one-element clusters and one with size 444. Clustering it to 40 clusters (Figure 6.2b) we get three one-element clusters and two salient clusters of sizes 148 and 310.

To check the consistency of clustering, in Figure 6.3 we present similar histograms after clustering the embeddings using Agglomerative Clustering. We see a very similar pattern in cluster size distribution as with K-means in all three embeddings. $E_S$ has a saliently big cluster of 351 elements.

The red line shows the average frequencies of words (AF) in each cluster in the corresponding textual dataset (Visual Genome Scene Graphs for $E_S$ and Wikipedia2020 for $E_L$.) In the visual case the notion of word frequency is not applicable. We were mainly interested in whether the saliently big clusters in $E_S$ are due to an artefact of word frequencies. Whereas in the case of 20 K-means clusters we only see a slight drop of AF, in the 40 cluster case the two biggest clusters have relatively low numbers, although there are other low AF clusters among the smaller ones as well (Figure 6.2). After Agglomerative Clustering (Figure 6.3) we observe a more substantial drop in AF for the two biggest clusters. In $E_L$ we see no such patterns, but the cluster sizes are less varied there.

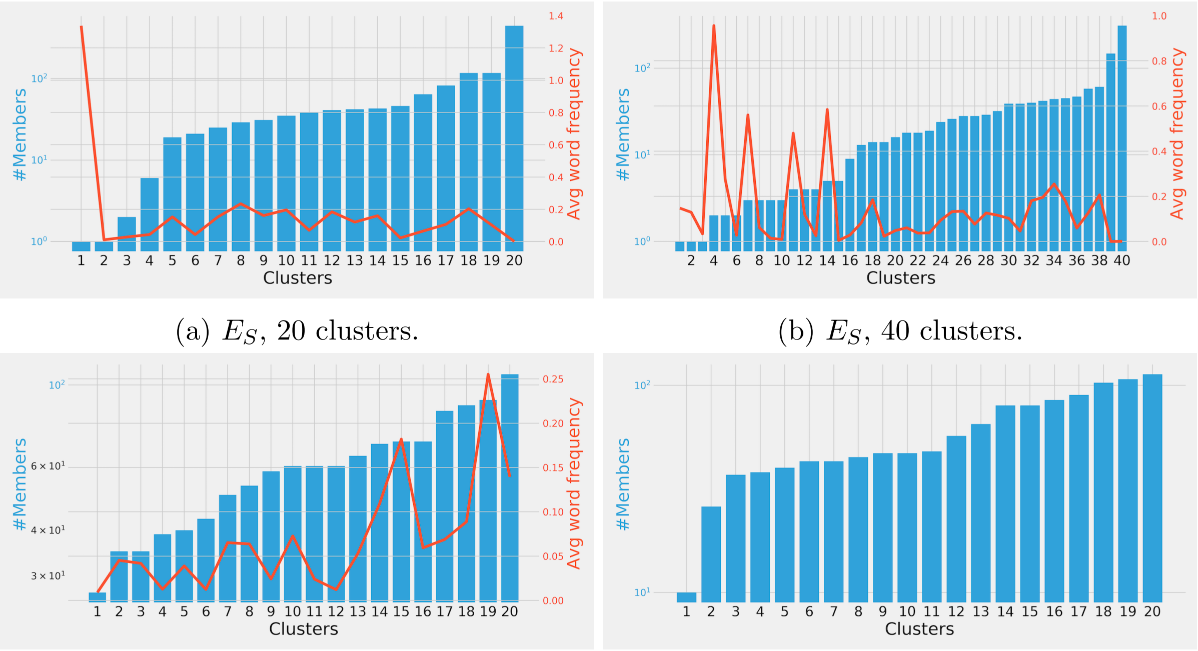
As an effective visualisation we use the T-SNE algorithm [Maaten and Hinton, 2008, Wattenberg et al., 2016] to zoom further into the structure of our embedding spaces. We applied Tensorboard[1] for the projections as well as their implementation of T-SNE. Following the guidelines in [Wattenberg et al., 2016] we tried different perplexity settings (running it multiple times). In most cases we did not find too much difference between the results on our data, but following the suggested range of $5 - 50$, we present results for perplexity = 30 or indicate otherwise. Figures 6.5-6.8 and D.10 contain T-SNE visualisations of the clusterings. The salience of the biggest $E_S$ K-means clusters is visible in all cases (Figure 6.5, 6.8, D.10). Based on the average frequency results, we think, that the reason for this huge separable cluster is at least partially that it includes more low frequency words. The breakdown of cluster cohesion is visible in the 40 cluster cases. In general, the clusters are fairly separated in all projections.

#### Cluster Similarities

Next, we looked into the individual clusters in each embeddings. Each row in Tables 6.2-6.4 contains the members of example clusters for the corresponding embedding. (See tables including all clusters in Appendix D.) Rows are ordered by the number of cluster members in increasing order. Words in column "Members" are ordered by their distance from the cluster centroid in increasing order. (In Tables of $E_S$ clusters in Figures D.2 and D.4 we shortened the biggest cluster, indicated by three dots, for better readability.)

We labelled each clusters post-factum in two ways:

[1] `https://www.tensorflow.org/tensorboard`



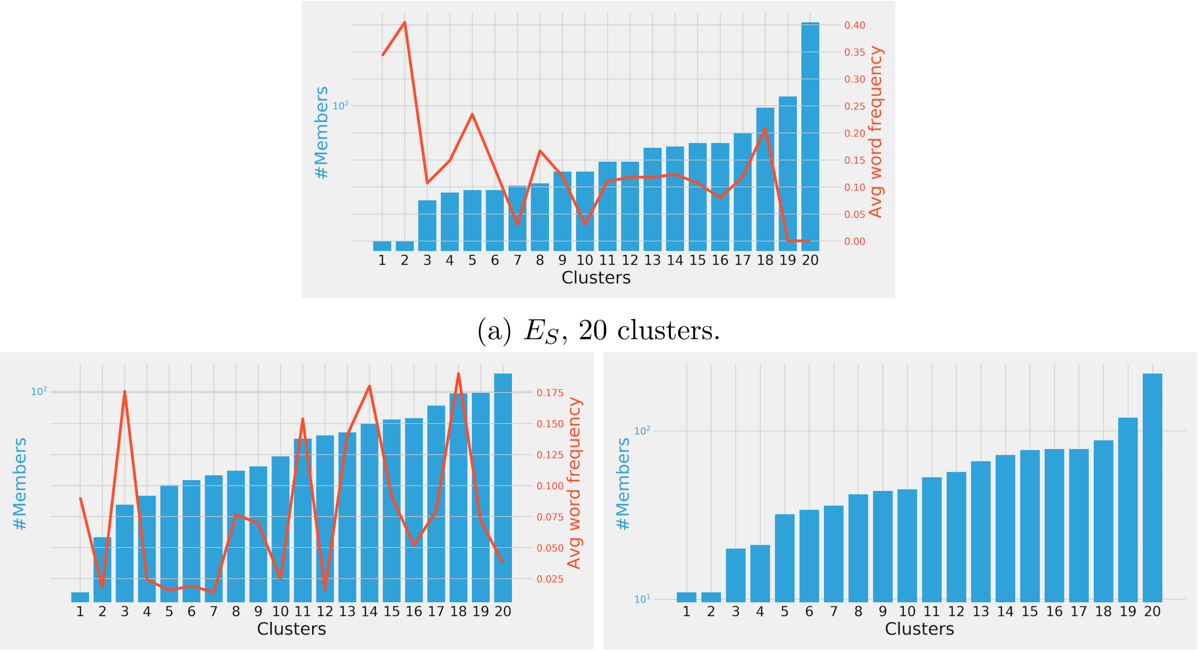
(a) $E_S$, 20 clusters.

(b) $E_S$, 40 clusters.

(c) $E_L$, 20 clusters.

(d) $E_V$, 20 clusters.

Figure 6.2: K-means Cluster size distributions. Y axis shows the number of cluster member in log scale. Red line shows the average frequencies of words in each cluster in the corresponding textual dataset.



(a) $E_S$, 20 clusters.

(b) $E_L$, 20 clusters.

(c) $E_V$, 20 clusters.

Figure 6.3: Agglomerative Cluster size distributions. Y axis shows the number of cluster member in log scale. Red line shows the average frequencies of words in each cluster in the corresponding textual dataset.

```
1. Cluster = [’apple’, ’pizza’]

2. closures(’apple’) = [
    Synset(’edible fruit.n.01’), Synset(’pome.n.01’),
    Synset(’fruit.n.01’), Synset(’produce.n.01’),
    Synset(’reproductive structure.n.01’), Synset(’food.n.02’),
    Synset(’apple tree.n.01’), Synset(’fruit tree.n.01’),
    Synset(’angiospermous tree.n.01’), Synset(’apple.n.01’),
    Synset(’apple.n.02’)]

3. closures(’pizza’) = [
    Synset(’dish.n.02’), Synset(’nutriment.n.01’),
    Synset(’food.n.01’), Synset(’pizza.n.01’)]

4. list of synset names in decreasing frequency order = [
    ’food’, ’nutriment’, ’pizza’, ’dish’, ’apple’, ’pome’,
    ’fruit’, ’apple  tree’, ’edible fruit’, ’fruit tree’,
    ’produce’, ’angiospermous  tree’, ’reproductive structure’]

5. labels = [’food’, ’nutriment’, ’pizza’]
```

Figure 6.4: WordNet label generation example.

1. **WordNet label** was generated by querying the synset closure up to a depth of 3 in the hypernym hierarchy for each words in the cluster. Then we took each synset name in the closure lists and created a set from each of them (by removing duplicates). Next, we concatenated all the sets (corresponding to one word) into one list. The generated cluster label is the first three most common lemmas in this list. An example is shown on Figure 6.4. This can be considered as a form of “crowd-sourced” annotation, as it relies on a dataset created by human linguistic experts.

2. **Own label** is our annotation (without looking at the WordNet labels). “Misc” stands for Miscellaneous, where we could not find an appropriate concept to describe the cluster.

Our own annotations and the WordNet labels are fairly consistent with each other, often use the same words or synonyms e.g., “drink”-“beverage”. One interesting exception is the fifth row in Table 6.4 of the image based clusters which we interpreted as female visual stereotypes, whereas the WordNet label is: “person, organism, casual agent”. We find our interpretation supported by previous work on the bias of Google Images [Kay et al., 2015], however, with the disclaimer of coherence being “in the eye of the beholder” [Bender et al., 2021]. WordNet labels can be sometimes more generic than our annotation. This may be because we exploit WordNet which was created by multiple experts as opposed to our own annotations.

In general, the Wikipedia based $E_L$ has more clusters with abstract topics, such as verbs, activities and communication. $E_S$ has more concrete clusters e.g., train, vehicles, building structures, containers or furnishing. Whereas the image based $E_V$ includes more clusters related to the outdoors, such as “travel”, “transportation”, “landscape” and “vacation”, and on appearance, such as “colours & materials”. These differences may not be surprising regarding each data source, but we would highlight the fact that these statistics are on the exact same vocabulary. Therefore, the difference between these data sources is not simply that they include different vocabularies, but that they "understand" the same words differently. This is the type of information we think is important to be conscious about when building on any data source or modality.

There are also some concepts that all three embeddings capture consistently, such as "food", "colours", "plants", "animals" and "body parts". Different embeddings differ, however, in the number of clusters they have related to similar concepts and of course their exact content differs to various extents.

In order to capture how similar the clusters are across the different embeddings, we measured the pairwise Jaccard similarity coefficient between each two embeddings. The Jaccard similarity coefficient between two clusters A, B is defined as

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}. \tag{6.1}$$

Note that, $0 \leq J(A, B) \leq 1$.

We calculated Jaccard similarity scores between each pair of clusters which represent concepts. Cluster maps of similarities are presented in Figures 6.9, 6.10 and 6.11. These are heat maps of Jaccard similarities, where the rows and columns of the matrix have been clustered for better visibility. Each row and column is labelled with their respective WordNet cluster label.

We observe that "food", "plants", "animals", "body parts" and "travel / vehicle" related clusters are distinctly more similar between each pair of embeddings than the other clusters. Beyond this, $E_S$ and $E_L$ have similar cluster related to "visual property", "clothing", "structures / buildings" and a "food" related $E_S$ cluster is close to a "container" cluster in $E_L$. $E_S$ and $E_V$ contain more similar travel related clusters: "travel, change, object" – "physical_entity, body_of_water, thing" and a pair of containers / instruments: "artifact, whole, instrumentality" – "instrumentality, container, substance". $E_L$ and $E_V$ have similar clusters on "structure / area" and an $E_L$ "artifact, whole, instrumentality" cluster is close to "food, beverage, produce" in $E_V$.

Similar cluster maps are presented for Agglomerative Clustering in Appendix D, Figures D.7–D.9. Figures D.1–D.6 include heat maps, where clusters are ordered by size. We did not find any pattern in similarities based on size.

We also compared K-means and Agglomerative clusters of the same modalities in Figures 6.12–6.15. We found the cluster structures fairly similar, the most similar clusters are food, body parts, animals, plants, vehicles and visual property related.

In order to quantify how similar each pair of cluster structures are, in Table 6.1 we summarise the number of cluster pairs with Jaccard similarities above thresholds of [0.2, 0.3, 0.4, 0.5, 0.6, 0.7]. In case of K-means, even though $E_L$ and $E_V$ have 9 cluster pairs with $0.3 < J(.,.) < 0.479$, $E_S$ has 12 clusters with $E_L$ and 8 with $E_V$ above a similarity of 0.2. With Agglomerative Clustering this relative closeness of $E_S$ and $E_L$ disappears, while the other two pairs show similar patterns to K-means. K-means and Agglomerative clusterings are fairly similar, with $E_V$ sharing the most similar cluster structure.

Figure 6.14 includes a heat map of K-means vs. Agglomerative $E_S$ clusters ordered by size. Here, we can see that the two saliently biggest clusters are relatively similar, reaching 0.65 Jaccard similarity. Their labels also share the words "person"and "change", which indicates that there is more meaningful coherence in those sizeable clusters than merely including low frequency words. Note that this coherence is hard to see with the naked eye because of the number of words to review.

**K-means**

| | >0.2 | >0.3 | >0.4 | Max |
|---|---|---|---|---|
| $E_S$-$E_L$ | 12 | 1 | 0 | 0.358 |
| $E_S$-$E_V$ | 8 | 2 | 0 | 0.363 |
| $E_L$-$E_V$ | 9 | 5 | 4 | 0.479 |

**Agglomerative**

| | >0.2 | >0.3 | >0.4 | Max |
|---|---|---|---|---|
| $E_S$-$E_L$ | 6 | 1 | 0 | 0.347 |
| $E_S$-$E_V$ | 9 | 4 | 2 | 0.5 |
| $E_L$-$E_V$ | 9 | 5 | 2 | 0.467 |

**K-means – Agglomerative**

| | >0.2 | >0.3 | >0.4 | >0.5 | >0.6 | >0.7 | Max |
|---|---|---|---|---|---|---|---|
| $E_L$-$E_L$ | 18 | 14 | 9 | 4 | 2 | 1 | 0.79 |
| $E_S$-$E_S$ | 16 | 15 | 13 | 9 | 5 | 1 | 0.729 |
| $E_V$-$E_V$ | 23 | 16 | 13 | 8 | 4 | 0 | 0.644 |

Table 6.1: Number of cluster pairs out of $20^2$ with Jaccard similarities above thresholds of [0.2, ..., 0.7]. Last column shows the maximum similarity.

| WordNet label | Own label | Members |
|---|---|---|
| food<br>nutriment<br>foodstuff | food | butter, cheese, bread, chicken, soup, sauce, dessert, beef, salad, meat, cake, steak, tomato, potato, pizza, flour, milk, meal, vinegar, bacon, pie, cooking, sushi, sandwich, breakfast, burger, menu |
| vascular plant<br>plant organ<br>plant part | plants | flower, flowers, tree, blossom, dandelion, foliage, fruit, weed, cactus, lily, bloom, shade, leaf, grass, sunflower, poppy, vine, plant, garden, iris, grow, daisy, oak, bulb, rust, herb, moss, tulip, palm, maple, root, tall, bush, seed, family |
| atmospheric phenomenon<br>physical phenomenon<br>change | weather | rain, snow, fog, weather, mist, drizzle, frost, dew, cold, wet, wind, smoke, sunlight, misty, sunrise, winter, storm, sunset, haze, sunshine, fire, spring, dusk, autumn, heavy, atmosphere, cloud, sunny, burn, flood, desert, sun, hot, ice, tropical |
| artifact<br>covering<br>clothing | clothing / fashion | wig, clothes, dress, shoes, jacket, sweater, skirt, sunglasses, leather, hair, costume, shirt, haircut, cloth, socks, waist, mannequin, collar, jewelry, tattoo, lingerie, beard, blonde, mask, fabric, uniform, necklace, linen, outfit, glove, hat, fashion, blanket, bikini, knitting, swimsuit, crochet, badge, coat, carpet, bracelet, arms, makeup |
| artifact<br>structure<br>whole | classical architecture | tower, building, marble, staircase, fountain, doorway, roof, chapel, steeple, porch, ceiling, mural, glass, wall, brick, statue, stone, arch, monument, dome, window, gravestone, sculpture, aisle, tiles, gate, interior, painted, decoration, concrete, church, graveyard, cathedral, curtain, painting, palace, clock, grave, portrait, choir, architecture, pyramid, memorial, square, castle, skyscraper, museum, cemetery, temple, organ |
| change<br>color<br>visual property | colour / decor | blue, bright, green, pink, black, yellow, dark, white, purple, red, brown, violet, rainbow, colour, orange, sky, rusty, silhouette, grey, diamond, redhead, light, flame, peacock, mirror, color, tiny, shadow, stripes, dull, rose, neon, colorful, crystal, bell, moon, horizon, arrow, silver, ivy, gold, swan, dragon, lantern, star, pearl, horn, ray, fox, globe, planet, bold, belt |

| | | |
|---|---|---|
| body part<br>part<br>artifact | body parts | skin, spine, neck, bone, chest, throat, shoulder, wrist, stomach, ear, jaw, cheek, lips, nose, eyes, eye, limb, toe, belly, skull, abdomen, finger, teeth, elbow, cord, whiskers, knee, thumb, tooth, muscle, ankle, tail, paws, lip, brain, flesh, leg, body, calf, heart, blood, tongue, brow, pain, tear, blade, mouth, liver, gut, arm, marrow, curled, canine, feathers, foot, vein, hip, cancer |
| change<br>act<br>be | verbs | bring, get, come, want, go, keep, take, know, find, say, give, make, understand, put, listen, enjoy, feel, leave, think, learn, imagine, gather, believe, fail, arrange, add, lose, create, way, hear, send, meet, collect, carry, avoid, buy, remain, allow, appear, might, enter, arrive, seem, entertain, break, steal, receive, stop, stand, build, locked, compare, retain, sell, handle, danger, eat, wander, face, unhappy, protect, please, pray, become, walk, expand, travel, plenty, greet, inspect, comfort, huge, possess, dominate, attach, roam, participate, speak, step, drawn, construct, replace, divide, great, living |

Table 6.2: Examples of the 20 clusters in $E_L$. Clusters are ordered by size. See all clusters in Appendix Table D.1

| **WordNet label** | **Own label** | **Members** |
|---|---|---|
| artifact<br>line<br>whole | train | railway, railroad, subway, curve, tunnel, run, shelter, train, station, tram, highway, track, rail, way, engine, stop, gate, bridge, smoke |
| structure<br>area<br>room | room | classroom, hallway, hall, closet, bedroom, room, bathroom, garage, office, cafe, museum, doorway, kitchen, shop, restaurant, store, mannequin, stadium, market, ceiling, corner |
| bird<br>vertebrate<br>person | animals | hummingbird, gull, peacock, hawk, pelican, crow, parrot, seagull, wing, swan, pigeon, owl, goose, flamingo, nest, eagle, tail, bird, silhouette, duck, chest, body, ledge, giraffe, zebra |
| travel<br>wheeled vehicle<br>self-propelled vehicle | vehicles | cab, car, taxi, police, vehicle, automobile, drive, racing, scooter, bike, van, street, road, motorcycle, truck, speak, wagon, bus, parade, drawn, asphalt, cop, parking, bicycle, sidewalk, traffic, driver, carriage, meter |
| plant organ<br>plant<br>vascular plant | plants | bloom, foliage, grave, dead, vine, blossom, ivy, pod, cactus, tree, moss, root, leave, limb, forest, bush, plant, lily, branch, weed, leaf, vein, sunshine, log, fence, flower, sunlight, wood, palm, bench, sun |
| structure<br>artifact<br>whole | building parts | chapel, cottage, steeple, castle, dome, story, cathedral, build, skyscraper, arch, lighthouse, apartment, hut, angel, shed, hotel, monument, window, staircase, home, cabin, house, roof, porch, tower, sculpture, patio, bell, deck, brick, church, cross, clock, step, statue |
| instrumentality<br>container<br>substance | vessel | champagne, tea, beverage, alcohol, honey, milk, pencil, tulip, juice, oil, bakery, ceramic, container, coffee, tin, cup, beer, sunflower, daisy, wine, rose, marble, bowl, sweet, maker, jar, vessel, mug, money, bottle, pumpkin, straw, glass, basket, box, pot, bucket, bunch |
| body part<br>artifact<br>part | pets & body parts | jaw, throat, pupil, cheek, canine, belly, brow, mouth, stomach, tongue, eye, nose, poodle, ear, hamster, lip, fur, tooth, teeth, pet, leg, wool, head, feline, toe, panda, smile, neck, face, beard, puppy, collar, horn, skin, cat, kitty, calf, nail, dog, tag, mother |

| | | |
|---|---|---|
| physical entity<br>body of water<br>thing | water | rapid, village, coast, bay, mist, horizon, canal, skyline, valley, sea, cliff, fog, town, waterfall, stream, water, sunset, pier, harbor, boardwalk, break, ocean, lake, fountain, shore, island, river, wave, splash, city, rock, ship, building, sand, hill, crane, mountain, beach, pond, surf, boat, pool |
| location<br>artifact<br>region | farm<br>animal | dandelion, boundary, grass, wild, deer, stork, field, mud, farm, windmill, garden, landscape, desert, cattle, dirt, area, barn, yard, zoo, ox, path, footprint, garbage, puddle, lawn, cow, sheep, concrete, snow, eat, lamb, goat, stone, cone, trail, rain, day, park, animal, cage, horse, bull, elephant |
| change<br>color<br>visual property | colors | bright, beautiful, big, dirty, small, colorful, grey, long, purple, dark, round, men, tiny, pink, eyes, painted, brown, gold, medium, white, hang, iron, silver, old, black, left, tall, red, safety, large, metal, blue, steel, yellow, leather, hanging, make, walk, green, right, color, bath, pair, washing, sitting, carry |
| food<br>produce<br>solid | food | drizzle, nuts, herb, beef, flour, season, cereal, cherry, breakfast, sugar, steak, bacon, burger, butter, rice, meat, meal, sauce, dinner, pie, raspberry, lunch, sushi, bean, mustard, pepper, seed, salt, soup, cheese, tomato, hot, berry, potato, dessert, strawberry, salad, cardboard, food, bone, lemon, burn, frost, chocolate, bread, turkey, sandwich, spoon, pizza, chicken, shell, candy, peel, cooking, bubble, knife, fruit, fish, donut, cake, apple, ice, banana, orange |

Table 6.3: Examples of the 20 clusters in $E_S$. Clusters are ordered by size. See all clusters in Appendix Table D.2

| **WordNet label** | **Own label** | **Members** |
|---|---|---|
| bird<br>aquatic bird<br>seabird | birds | seagull, gull, goose, duck, pelican, swan, mallard, stork, eagle, flamingo |
| furnishing<br>furniture<br>instrumentality | furnishing | furniture, stand, booth, desk, modern, display, bed, chair, container, door, appliance, drawer, sofa, curtain, couch, bench, crib, frame, box, table, tv, window, computer, cradle, television, mac |
| instrumentality<br>self-propelled vehicle<br>wheeled vehicle | car<br>related | accident, cord, vehicle, auto, automobile, skate, photography, truck, race, arrive, ford, chopper, cab, rally, seat, industrial, smart, mechanic, racing, car, demolition, triumph, construction, motorcycle, machine, taxi, engine, driver, crane, carriage, van, bus, cannon, motor, tank, hockey, wagon, camera |
| vascular plant<br>plant<br>grow | plants | weed, bunch, maple, cancer, iris, poppy, dandelion, leave, flower, rose, foliage, grow, plant, cactus, spring, tulip, ivy, palm, lily, leaf, daisy, tree, root, wheat, wool, raspberry, tobacco, flowers, blossom, butterfly, sunflower, cotton, herb, violet, oak, moss, strawberry, nest, dew, berry, rice, branch, coal |
| person<br>organism<br>causal agent | “female<br>topics” | woman, model, brandy, pink, actress, lady, girl, young, wife, tiny, haircut, blonde, women, girls, hot, mother, hair, portrait, body, makeup, cheek, wig, neck, muscle, chest, lingerie, waist, redhead, child, face, bride, belly, bikini, kid, swimsuit, baby, brow, skirt, dress, short |
| food<br>nutriment<br>substance | food | sushi, meal, sandwich, pie, breakfast, lunch, food, supper, flour, cereal, sweet, dessert, dinner, subway, diet, cake, date, steak, sauce, bread, copper, nuts, bacon, cooking, beef, meat, bakery, knitting, eat, potato, salad, donut, pizza, burger, coffee, soup, bean, cheese, vitamin, fruit, pumpkin, rock, marrow, market, timber |

| | | |
|---|---|---|
| artifact<br>change<br>cover | colours & materials | texture, fabric, cloth, metal, rain, concrete, paper, suds, rough, words, stone, wall, square, dense, leather, quote, wood, frost, mud, noise, text, purple, carpet, blue, tiles, dirt, droplets, red, sand, fog, formula, mist, pattern, handwriting, green, straw, linen, asphalt, stripes, crowd, marble, yellow, black, brown, grey, grass, white |
| body part<br>artifact<br>part | body parts | gut, throat, wrist, burn, ear, thumb, elbow, listen, shoulder, liver, pain, knee, arms, hand, toe, finger, give, tongue, limb, abdomen, jaw, receive, nail, arm, feet, hear, skin, washing, head, ankle, hip, teeth, tear, stomach, brain, foot, lip, mouth, leg, flesh, mask, eyes, nose, skull, eye, socks, lips |
| structure<br>artifact<br>area | room | museum, garage, hall, classroom, kitchen, cellar, interior, office, diner, decoration, exhibition, hotel, ceiling, restaurant, store, bathroom, trial, pub, class, closet, cafe, room, porch, stairs, deck, hospital, living, corridor, aisle, bar, staircase, doorway, hallway, chapel, floor, lab, station, bedroom, gate, elevator, theatre, escalator, tunnel, organ, alley, library, jail, tram |
| travel<br>change<br>object | vacation | island, view, reflection, harbor, nice, side, sea, summer, tropical, pollution, port, aircraft, pier, travel, surfers, journey, sunny, coast, flying, morning, ocean, seashore, horizon, mare, holiday, lake, surf, shore, vacation, bay, airport, cliff, sunlight, air, river, storm, ship, fishing, beach, desert, harbour, puddle, flight, sailing, evening, sunrise, skyline, vessel, lighthouse, dawn, sunset, rocket, mountain, whale, underwater, boat, swimming, swim, plane, dusk, jet, cloud, sky, airplane, ski |
| change<br>abstraction<br>state | festival | theme, wisdom, soul, image, possess, large, confidence, happiness, beautiful, joy, love, ceremony, festival, movement, abundance, dead, depth, celebration, lover, run, demon, blurred, pray, happy, remain, wet, dance, navy, family, carnival, angel, sculpture, ray, dragon, drive, atmosphere, night, shadow, band, god, believe, party, dark, hanging, abstract, show, christmas, monster, devil, jump, lighting, sunshine, warrior, painting, water, aquarium, zombie, concert, haze, crystal, statue, explosion, jazz, jellyfish, wave, bright, rainbow, ice, light, smoke, club, neon, colorful, hole, protest, autumn, rust, reef, flame, fire |
| person<br>organism<br>causal agent | animals | animals, animal, picture, painted, zoo, turkey, curled, goat, companion, pets, canine, pet, prey, relaxed, horse, spirit, tail, dog, chipmunk, squirrel, pigeon, fox, cute, please, sheep, owl, birds, military, giraffe, lion, lamb, bee, insect, hamster, hawk, licking, bird, cat, puppy, feline, terrier, deer, calf, rat, chicken, camel, dragonfly, whiskers, poodle, cow, hound, cattle, lizard, fish, bunny, crow, wolf, tiger, parrot, zebra, cheetah, fur, panda, bull, wasp, ox, hen, frog, crab, snake, boxer, hummingbird, rabbit, elephant, pupil, husky, peacock, spider, pug, ant |

Table 6.4: Examples of the 20 clusters in $E_V$. Clusters are ordered by size. See all clusters in Appendix Table D.3

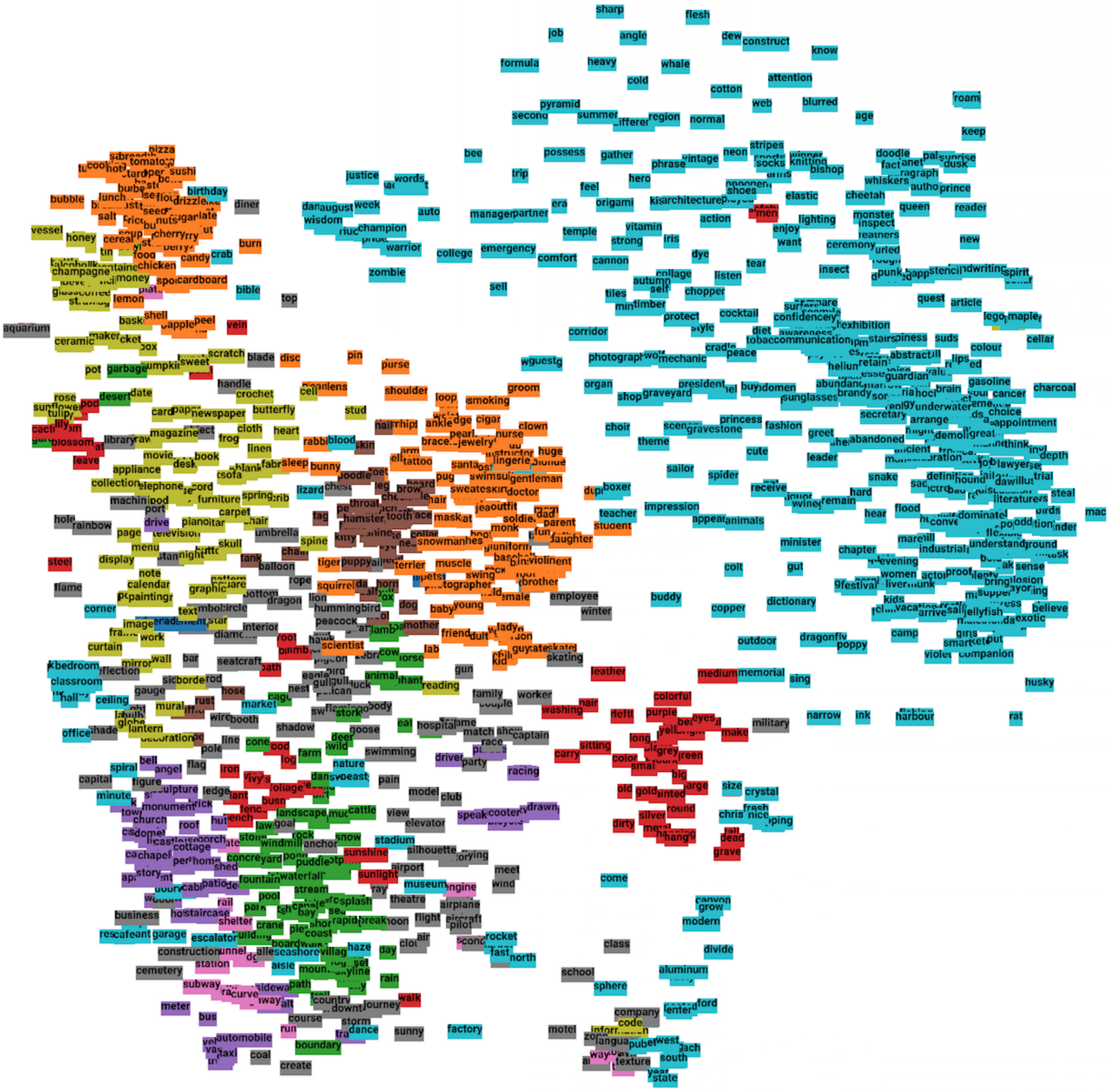


Figure 6.5: T-SNE plot of $E_S$ with 20 cluster labels obtained by K-means clustering.

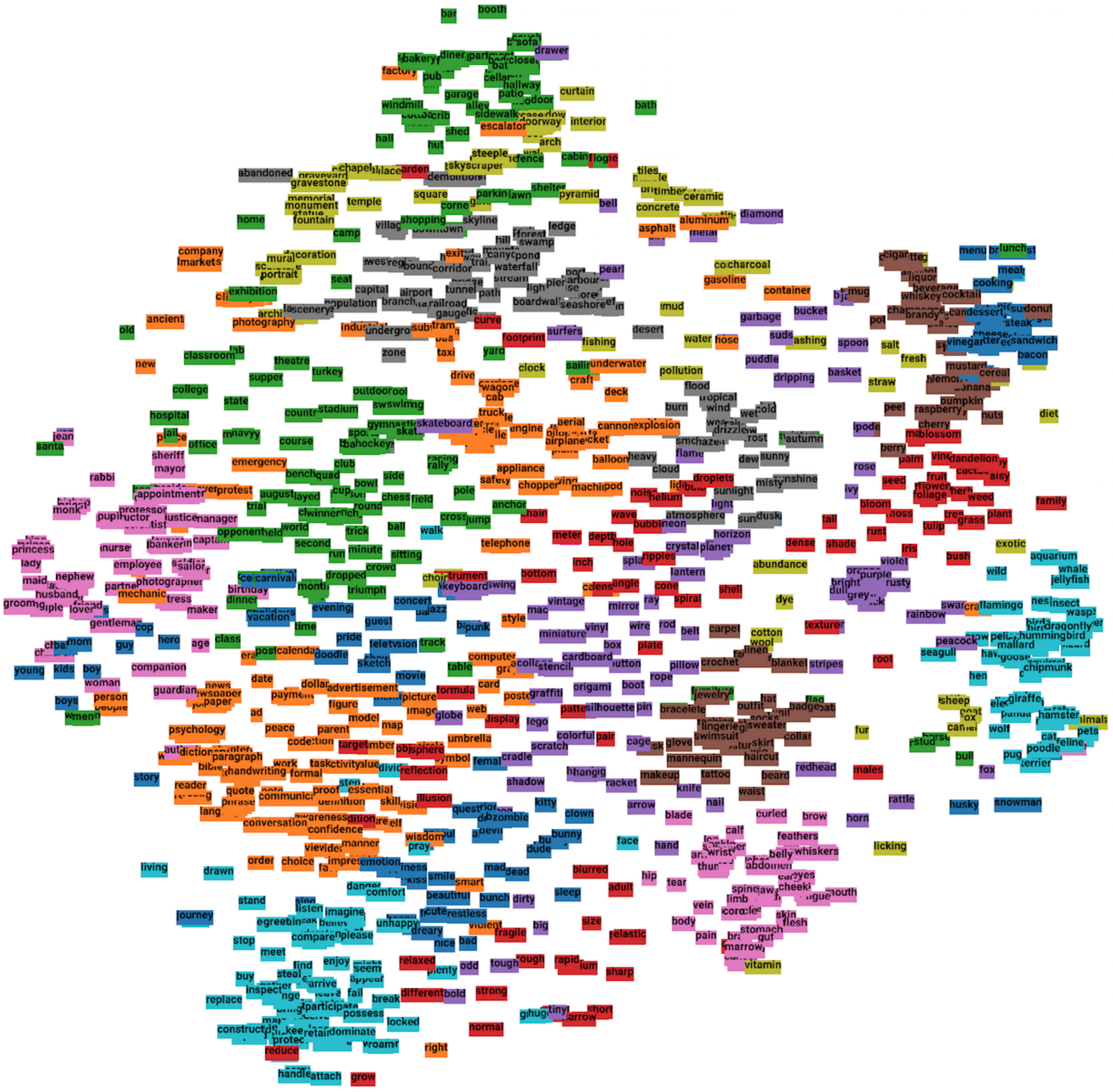


Figure 6.6: T-SNE plot of $E_L$ with 20 cluster labels obtained by K-means clustering.

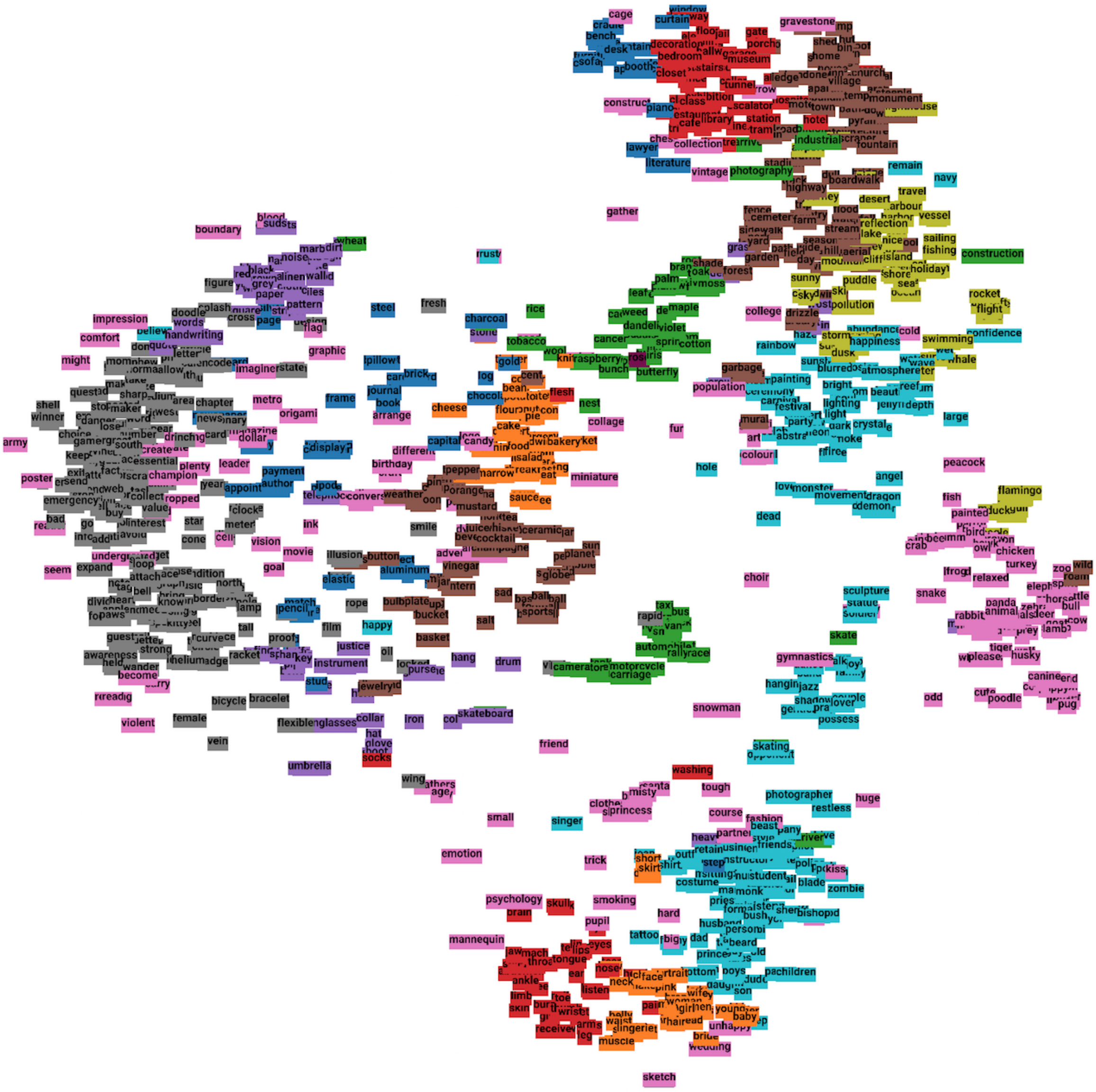


Figure 6.7: T-SNE plot of $E_V$ with 20 cluster labels obtained by K-means clustering.

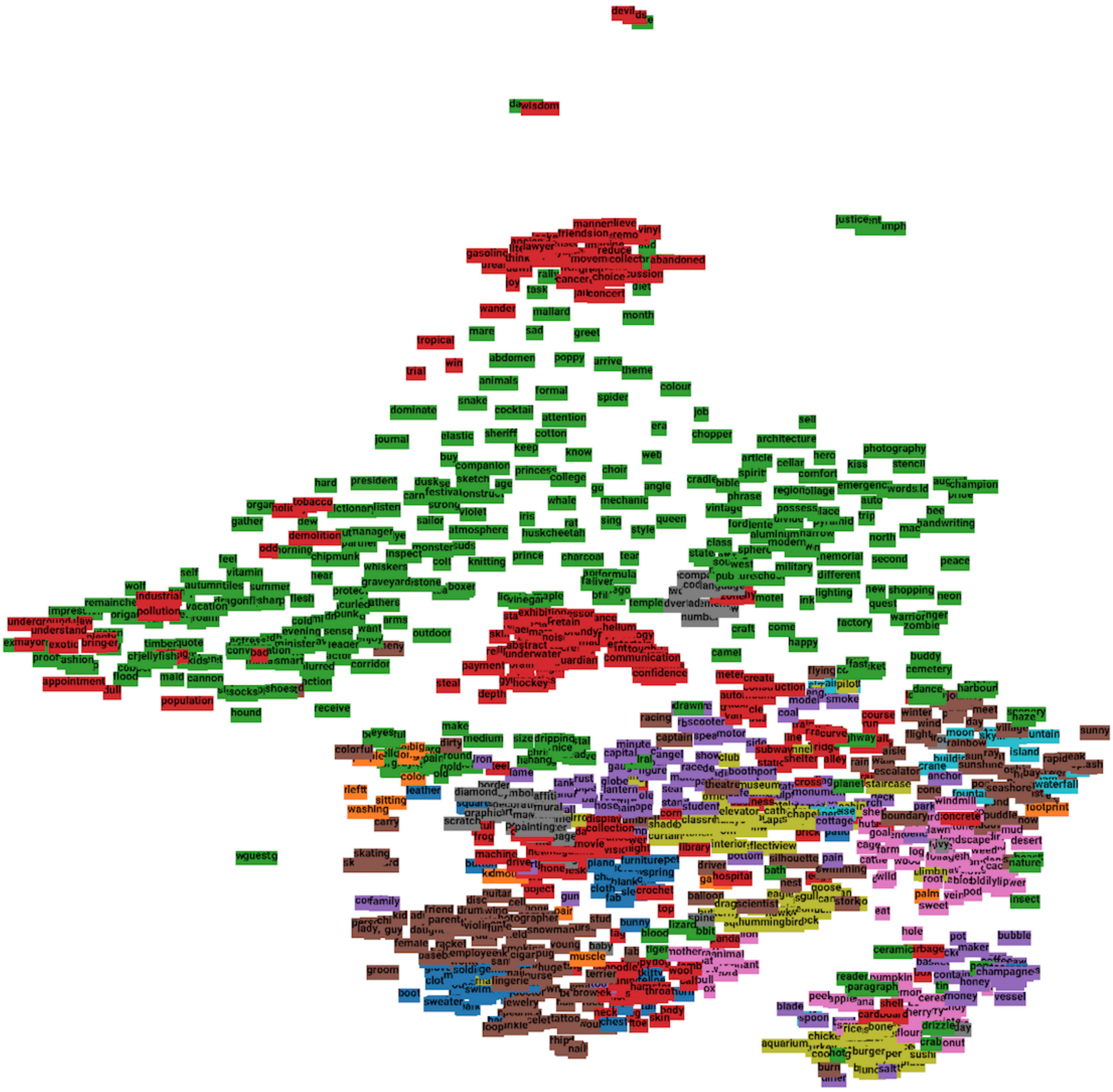


Figure 6.8: T-SNE plot of $E_S$ with 40 cluster labels obtained by K-means clustering. TSNE perplexity = 10.

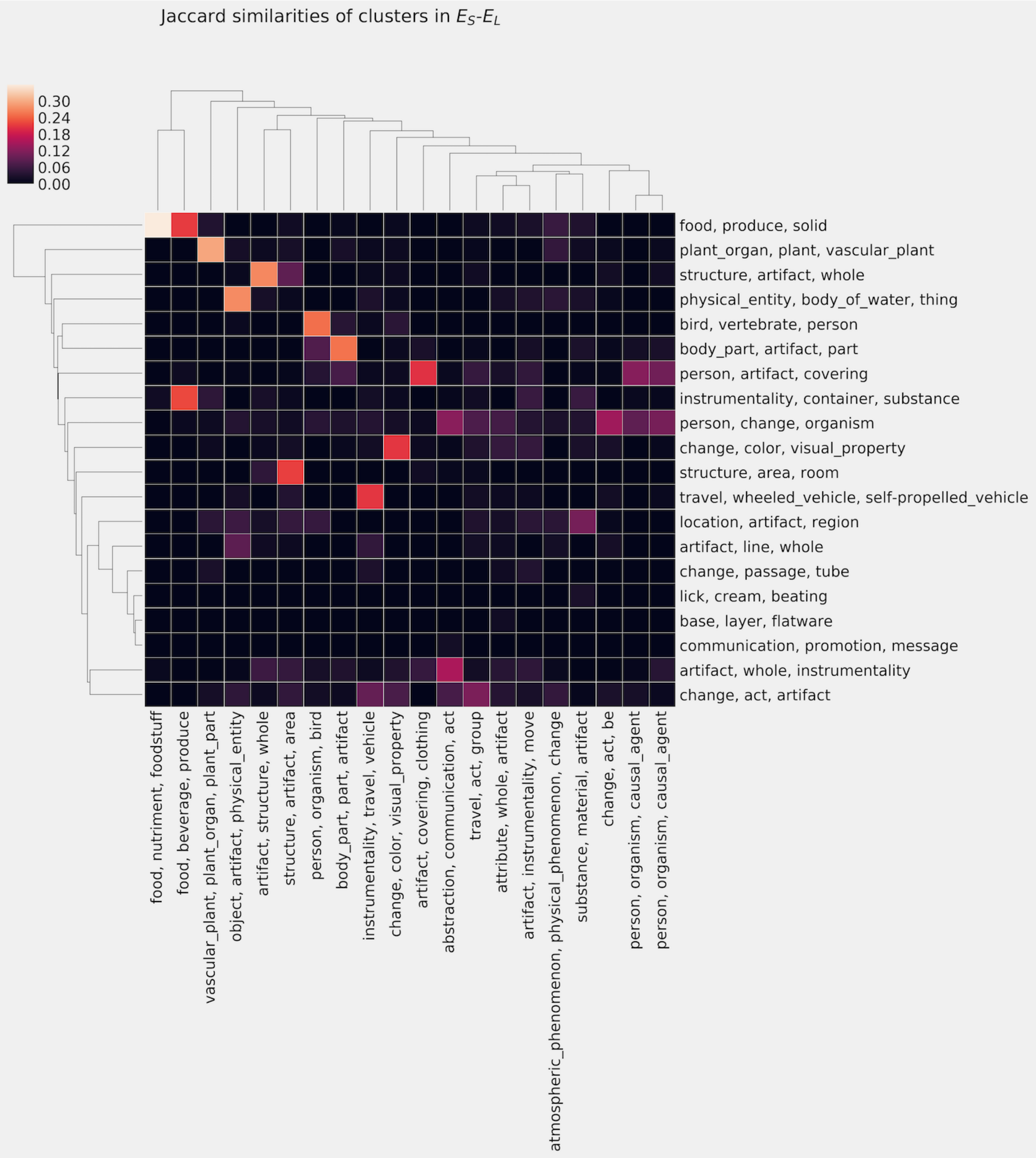


Figure 6.9: Cluster map of Jaccard coefficients between K-means clusters of $E_S$ (axis $y$) and $E_L$ (axis $x$).

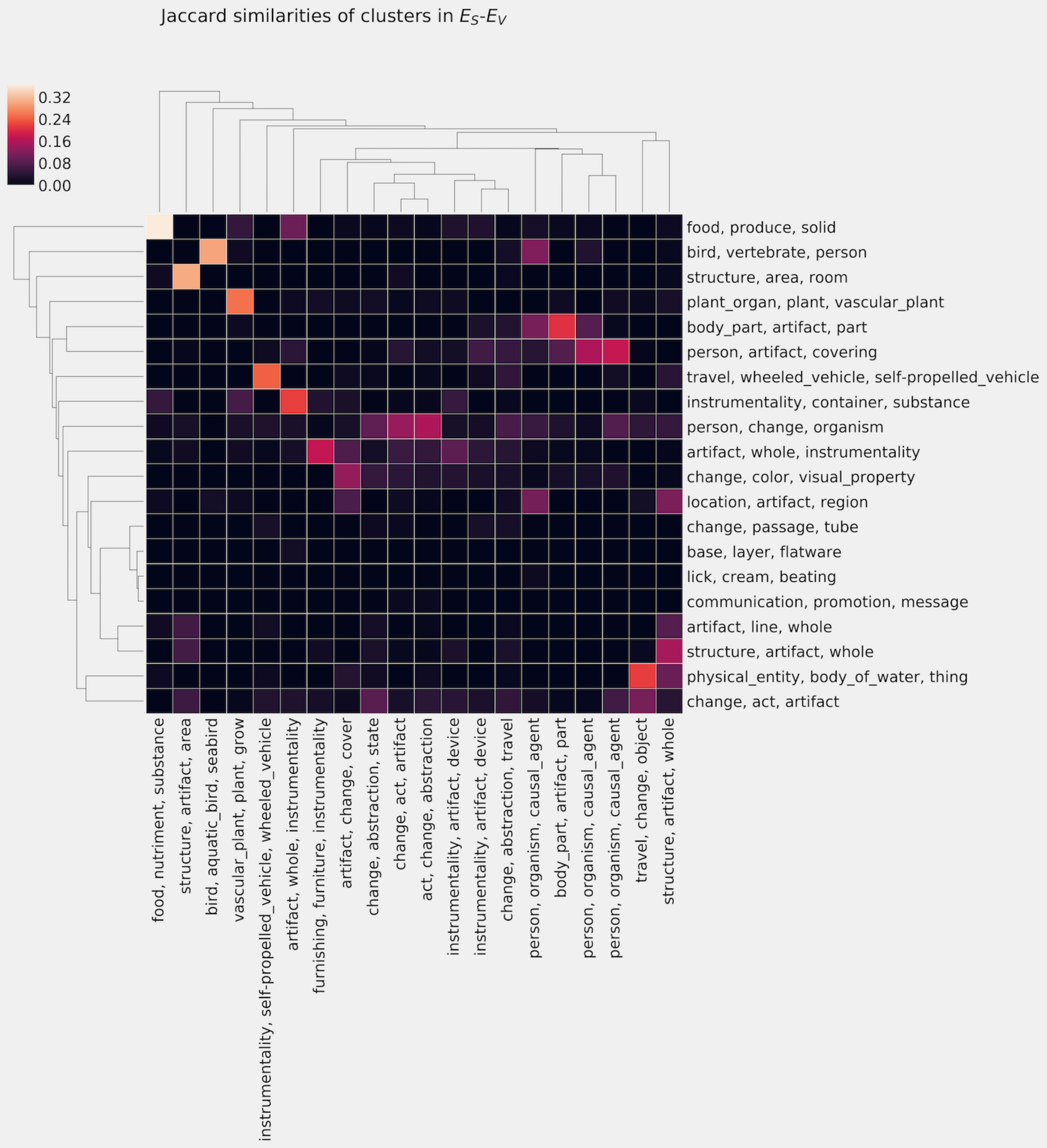


Figure 6.10: Cluster map of Jaccard coefficients between K-means clusters of $E_S$ (axis $y$) and $E_V$ (axis $x$).

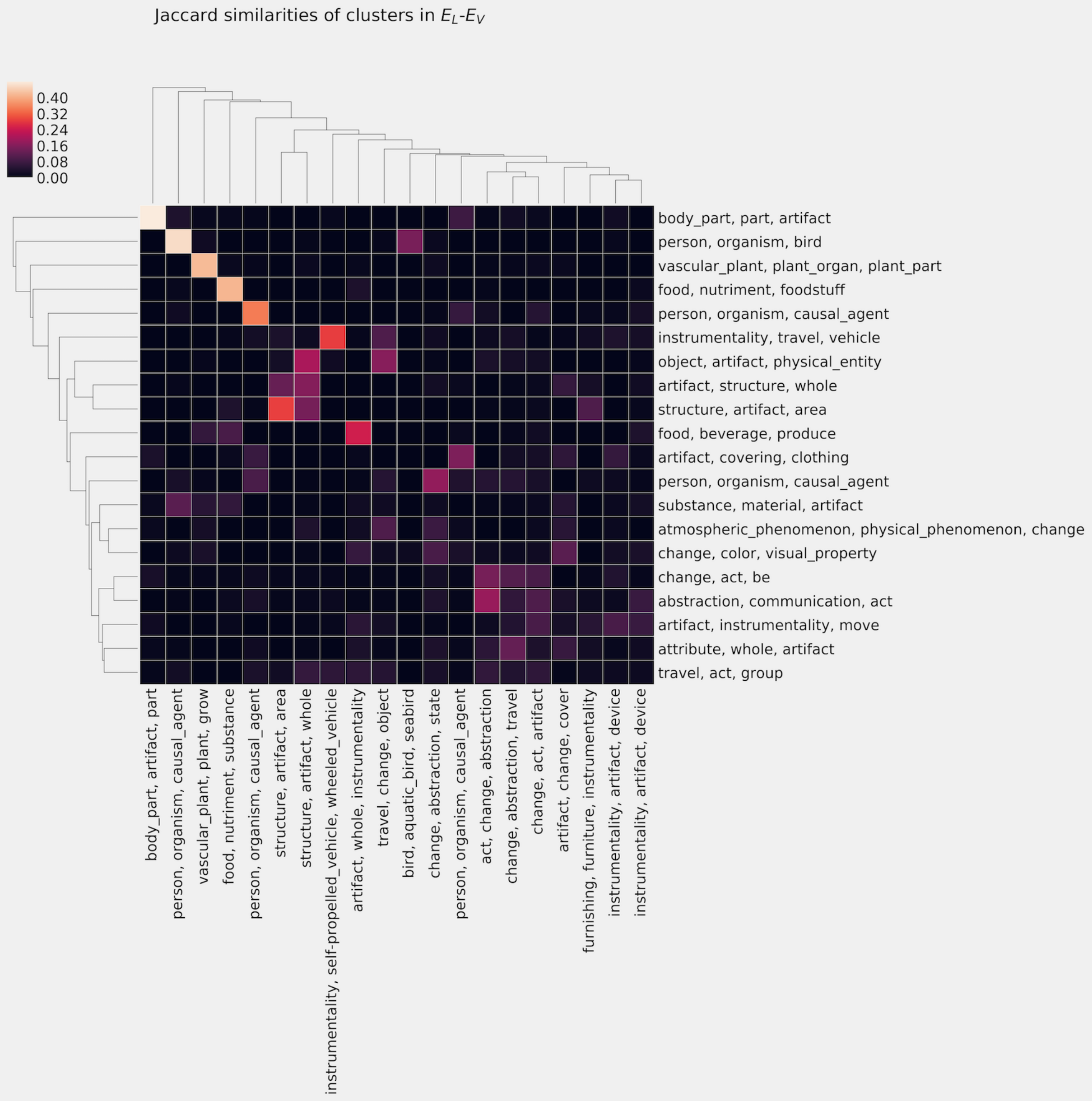


Figure 6.11: Cluster map of Jaccard coefficients between K-means clusters of $E_L$ (axis $y$) and $E_V$ (axis $x$).

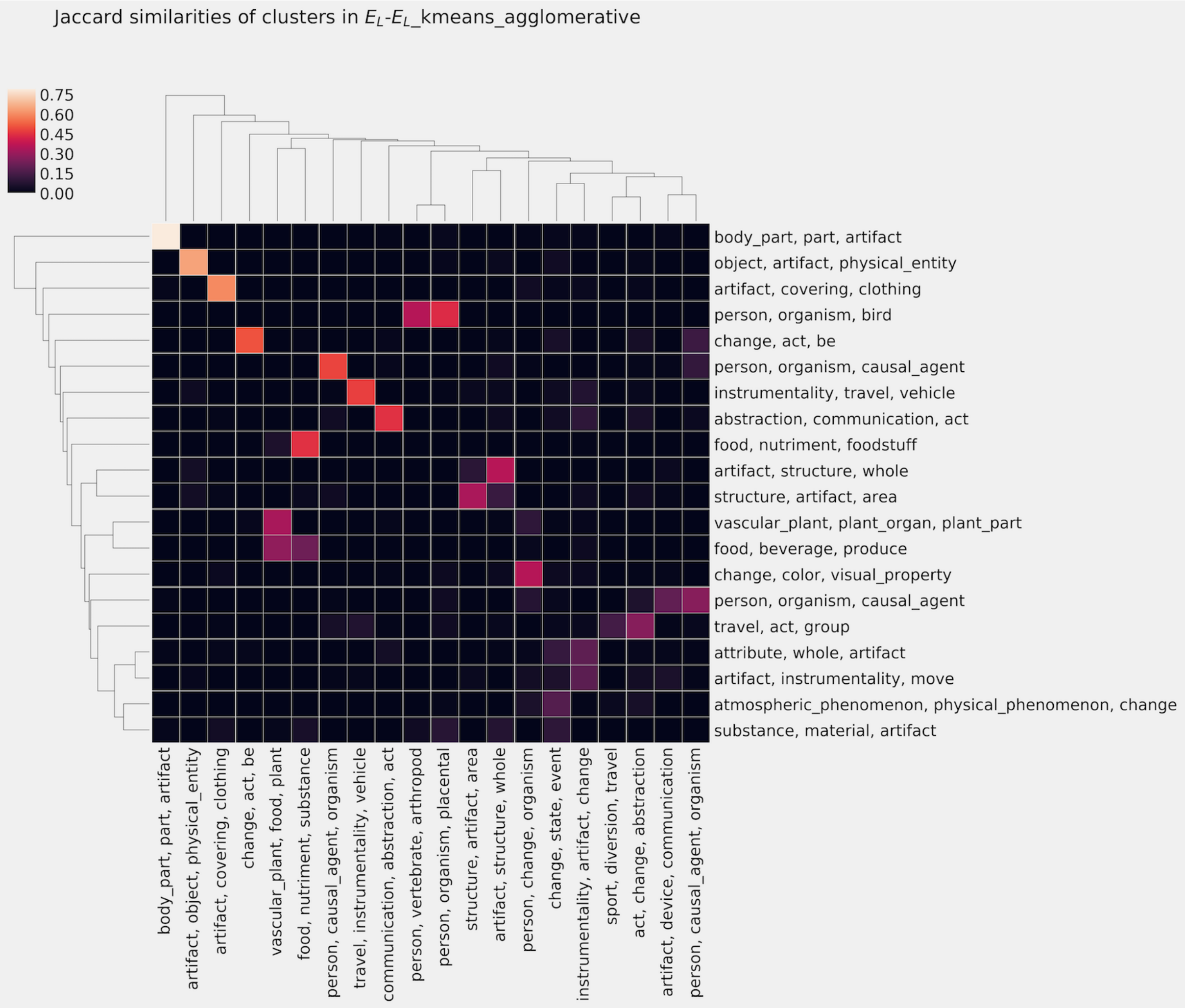


Figure 6.12: Cluster map of Jaccard coefficients between K-means (axis $y$) and Agglomerative (axis $x$) clusters of $E_L$.

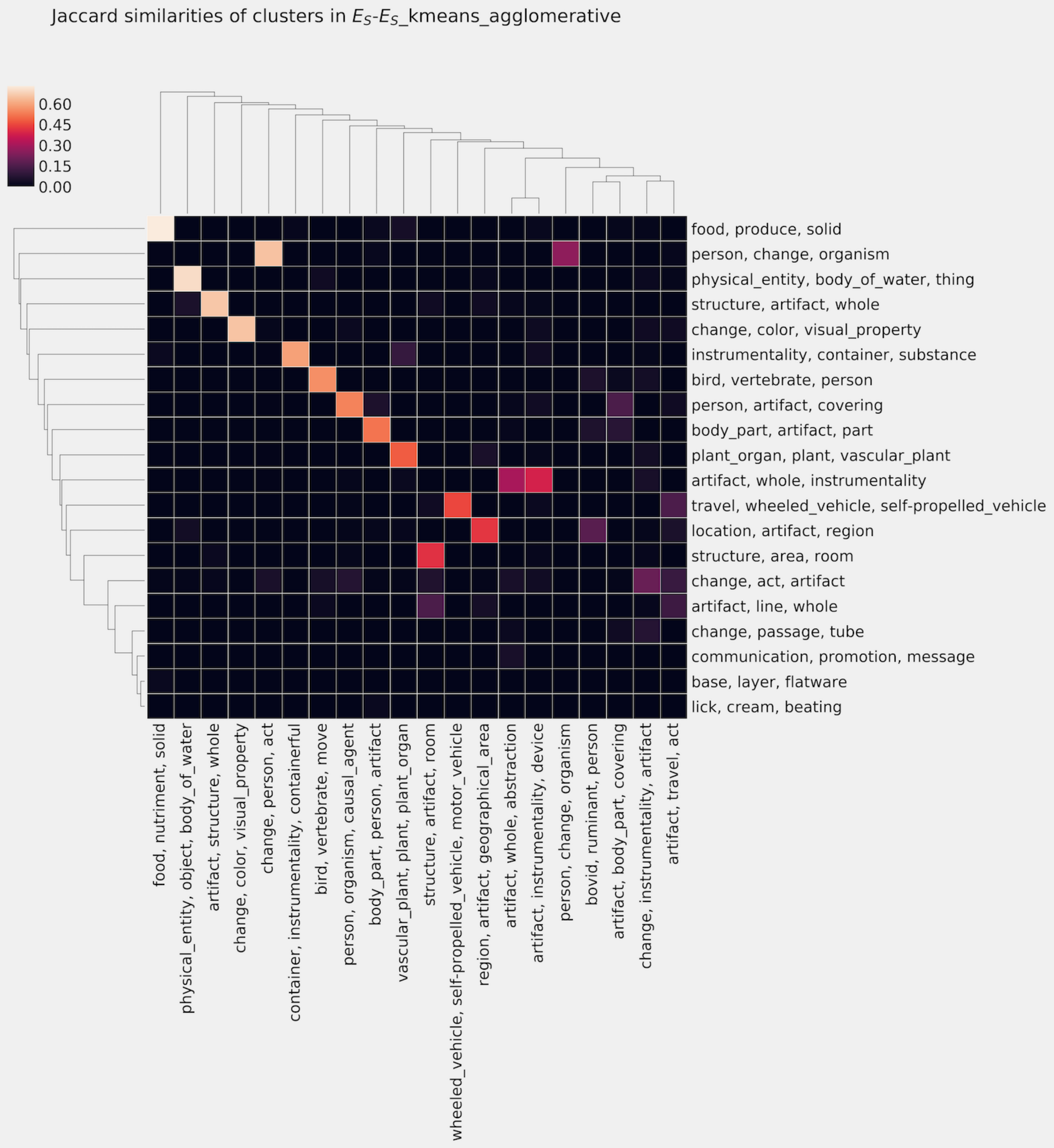


Figure 6.13: Cluster map of Jaccard coefficients between K-means (axis $y$) and Agglomerative (axis $x$) clusters of $E_S$.

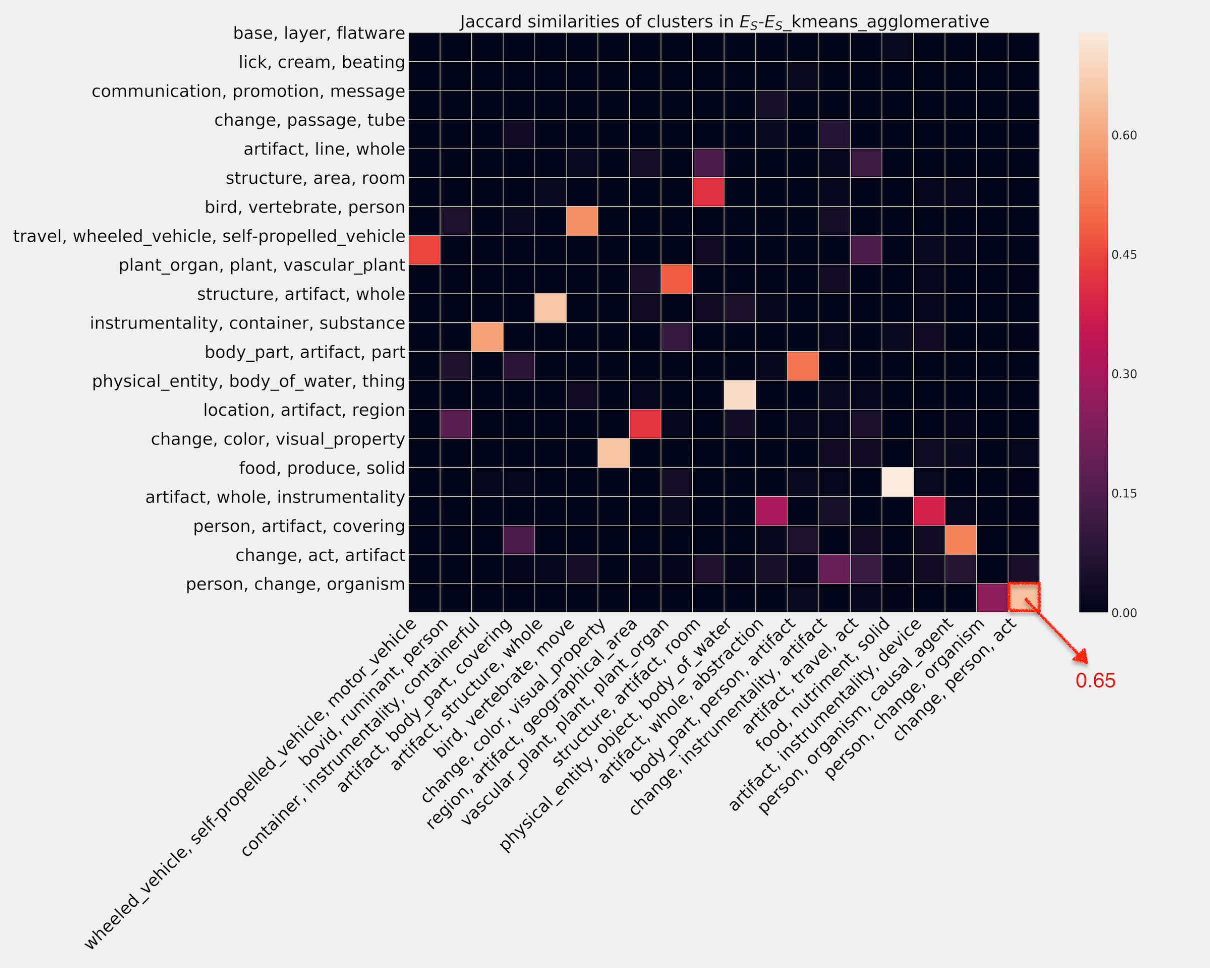


Figure 6.14: Heatmap of Jaccard coefficients between K-means (axis $y$) and Agglomerative (axis $x$) clusters of $E_S$. Clusters are ordered by size.

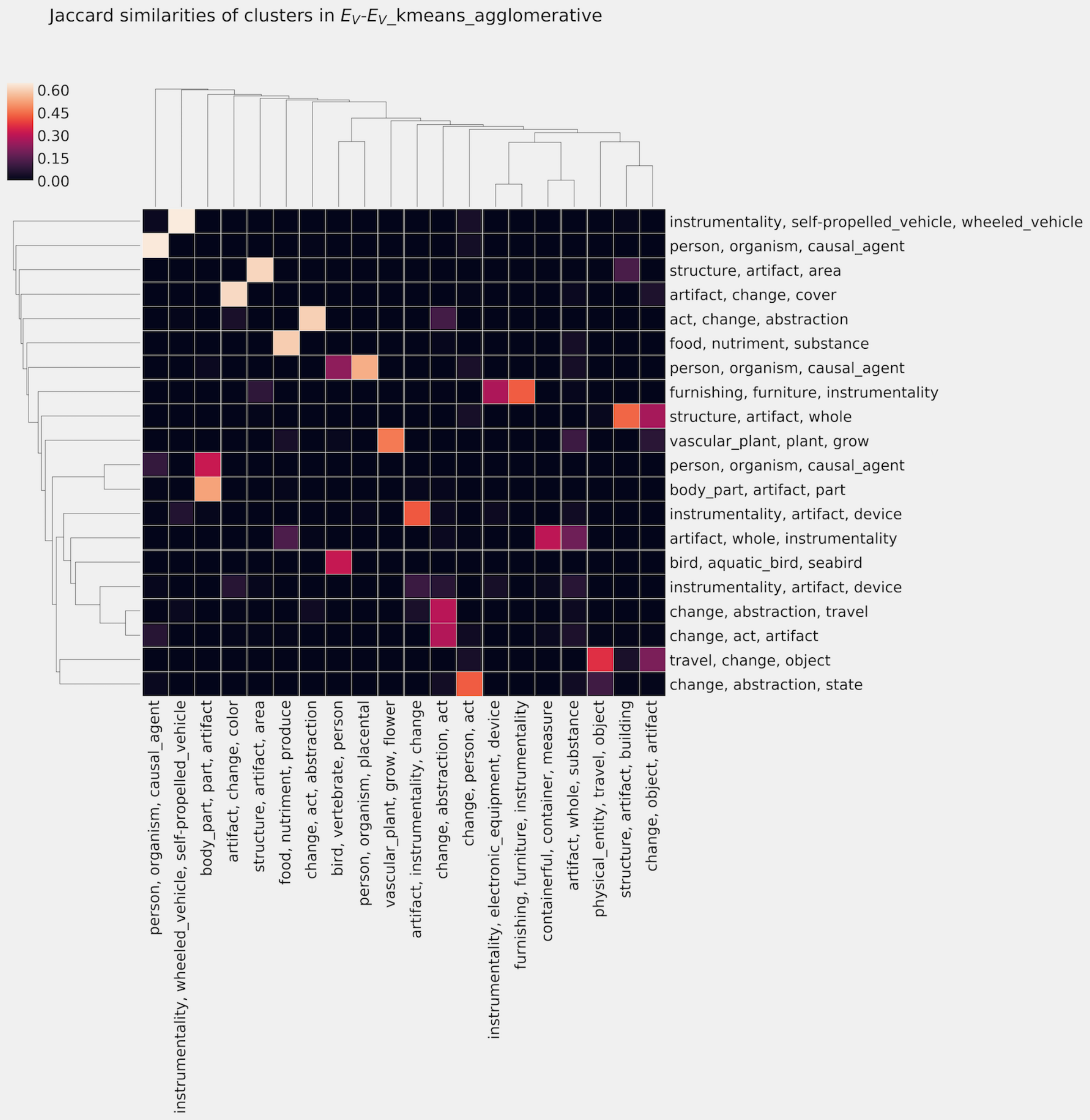
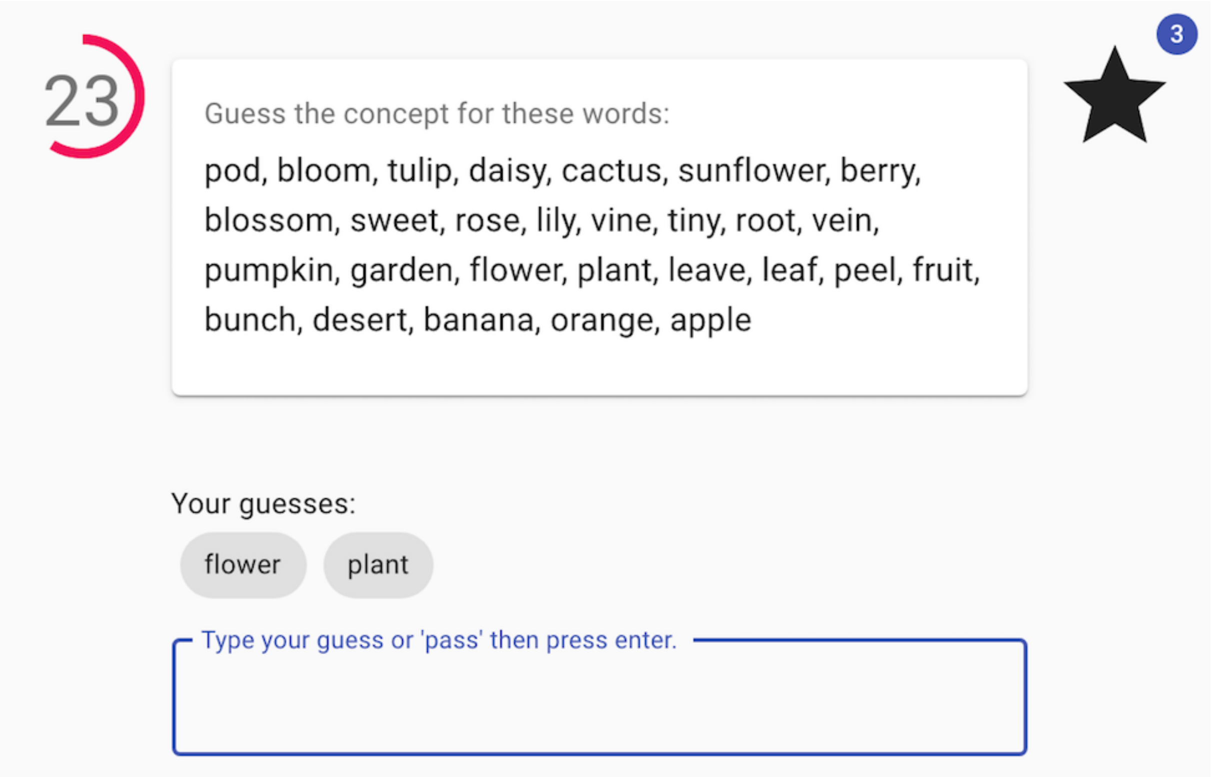


Figure 6.15: Cluster map of Jaccard coefficients between K-means (axis $y$) and Agglomerative (axis $x$) clusters of $E_V$.



Guess the concept for these words:

pod, bloom, tulip, daisy, cactus, sunflower, berry, blossom, sweet, rose, lily, vine, tiny, root, vein, pumpkin, garden, flower, plant, leave, leaf, peel, fruit, bunch, desert, banana, orange, apple

Your guesses:

flower plant

Type your guess or 'pass' then press enter.

Figure 6.16: Screen-shot of Concept Game, a two player, collaborative gamified data collection app, for acquiring cluster label annotations.

We developed a two player, collaborative gamified data collection app, called *Concept Game*[2], similar to ESP Game [Von Ahn and Dabbish, 2004], but with word lists (clusters) instead of images (Figure 6.16). The pair of players have to guess the concept for a list of words, which are the elements of all the clusters from this section. They get a score if their guesses have one word/expression in common. This way we aim to collect more human cluster label annotation for different modalities in the future.

The back-end involves a Sqlite Database on an AWS server[3], where we collect data. The dataset includes two tables:

- **Game**: It stores each game rounds, which is each time the users see a new word list they are guessing a concept for. We log the following attributes:
  - $game\_id$ = TextField()
  - $start\_time$ = DateTimeField()
  - $cluster\_id$ = TextField()
  - $user1$ = TextField(): firsts user's id
  - $user2$ = TextField(): second user's id
  - $guess$ = TextField(): the guessed word – NONE if they ran out of time
- **Answer**: This table stores the log for each word the users typed in with time stamps. This way, later, the time needed for agreeing on a cluster label can be

[2] http://concept-guessing-game.com/

[3] https://aws.amazon.com/

used to infer the difficulty / ambiguity of a cluster word list. It logs the following attributes:

- *game* = ForeignKeyField(Game, backref='answers'): reference to a *game_id* in Game.
- *cluster_id* = TextField()
- *user* = TextField(): id of the user who typed in a word as an answer
- *word* = TextField()
- *e_time* = TimeField(): elapsed time since the beginning of the game

The project is still under development in order to make it more accessible. Currently, people can only play if there are enough players active on the platform. So far only test data has been collected. In the future an auto replay functionality would greatly improve the usability of the game.

The code is publicly available on Github[4]. The web technology development was helped by Krisztián Gergely[5].

[4] `https://github.com/anitavero/concept_game`

[5] `http://krisoft.hu/`

### 6.2.3 Supervised Visualisation

In this Section we use the same T-SNE algorithm as in Section 6.2.2. However, for the labelled projections we apply a WordNet based automatic labelling technique on the words beforehand. This is fundamentally different from the previous Section, where the labelling came from the clustering method in an unsupervised fashion. In that case, WordNet was used only for analysing the cluster outputs, whereas here we label the data first. This way we can inspect our embedding spaces based on pre-defined concepts. The previous method is more generic, this approach contributes to the interpretation of embeddings.

#### Automatic Class Label Annotation

Figures 6.20 – 6.24 show coloured plots where the colours correspond to 13 class labels. We used the same coarse categories as in [Gupta et al., 2019]. They labelled their data manually, which we were not able to do due to the size of our data. Therefore, we developed a technique to automatically label our words using the WordNet hierarchy. Let $C$ be the set of class labels, $C = \{$*transport, food, building, animal, appliance, action, clothes, utensil, body, colour, electronics, number, human*$\}$. All words in the embeddings' common subset vocabulary $V_{common}$ were labelled with a class in the following way: First, we queried the synset list $S(c)$ for each class $c \in C$. Then we obtained the synset closure of each word $w$ up to the third level in the hypernym hierarchy: $S_3^{cl}(w)$. The class with the maximum number of synset overlap with each word synset closure is assigned as the word's class label: $class(w) = \max_{c \in C}[S(c) \cap S_3^{cl}(w)]$. We only show words where this maximum exists.

#### Results

Figure 6.17 depicts a 2D projection of a 3D T-SNE plot of a 100 000 sample from the SGNS Wikipedia 2020 model. After looking at the word labels, clear clusters became apparent, such as words in different languages, topics (e.g., math, mental health, numbers). The thin curves usually contain numbers with the same number of digits and in order. Figures 6.18 and 6.19 show two examples for the clusters.

Figure 6.20 shows a 2D T-SNE plot of our Wikipedia 2020 model trained on the whole corpus. Despite the simple heuristic we used to generate class labels, clearly separable clusters emerged for many of them. We can see *colours* indicated by orange, *numbers* by blue, *clothes* by red, *food* related words by light green, *buildings* by brown, *animals* by purple etc. Some of the confused labels visibly come from the failure of our labelling technique, but looking at it, many mislabelled words cluster around other words in the same topic / category.

In Figure 6.21 – 6.24 we show similar projections for $E_L, E_V, E_S$ and a random embedding $E_R$, where we restricted the vocabulary to the intersection of the three modalities, then kept the ones with an existing WordNet label, resulting 252 words. All $E_L, E_V, E_S$ clearly show much more distinct clusters with much better defined class labels than the random embedding. This may seem obvious, however, it is worth noting, since in very high dimensions even random vector spaces can show some structure. In our projection in Figure 6.24, both data points as well as labels are uniformly distributed.

Looking at the projections in Figure 6.21 – 6.23 the three modalities have different cluster shapes: $E_V$ having the most and $E_S$ having the least coherent and separable clusters. This is consistent with the results on clusterization metrics in Figure 6.1. In general, classes *transport, food, building, animal, clothes, colour, number, action* look to be better captured by this labelling and projection technique than *appliance, utensil, body,* *electronics, human.* This is probably due to the coarse labelling method, and could be alleviated by collecting human annotation. [Gupta et al., 2019] reported that their visual-context model showed more distinct clusters than their linguistic one using GloVe. In our T-SNE projections we did not find such patterns, although our method is fundamentally different from theirs, as they use early-fusion, GloVe, they do not exploit the Visual Genome graph structure, and they apply manual labelling. Overall, it is remarkable how much structure can already be revealed without the need for acquiring additional human effort.

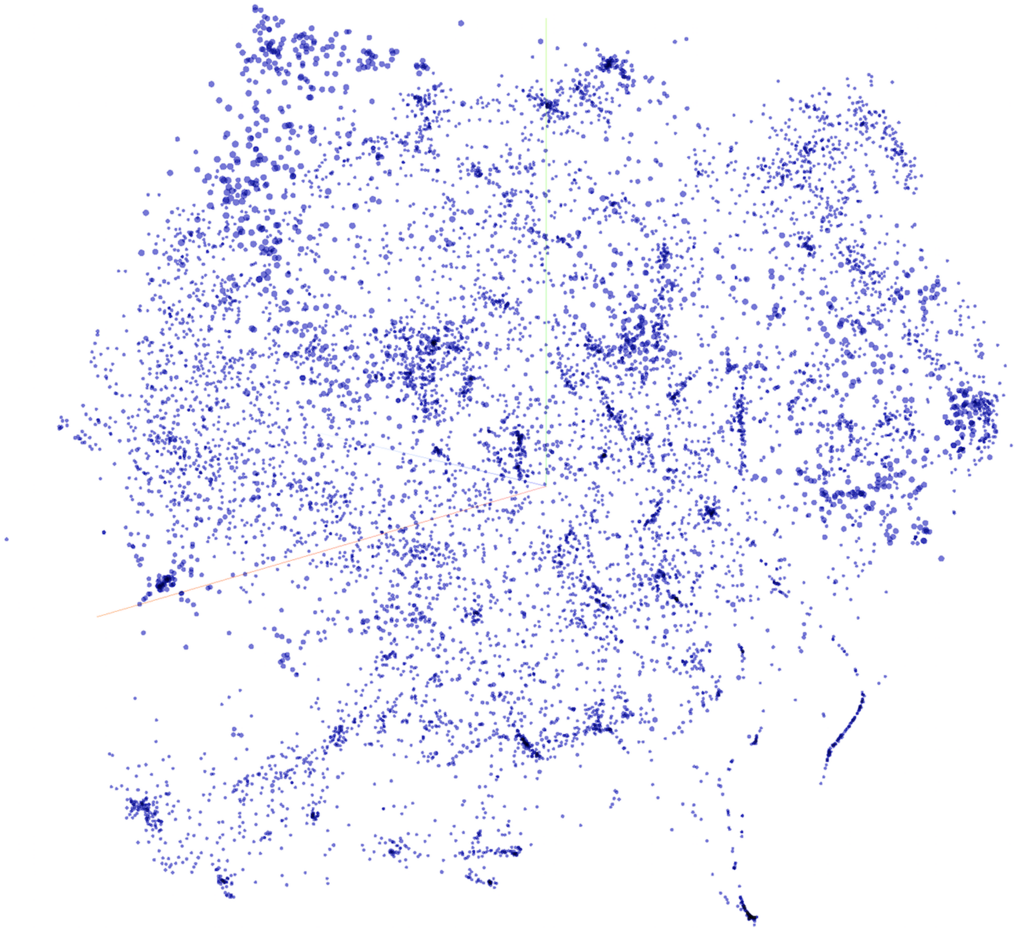
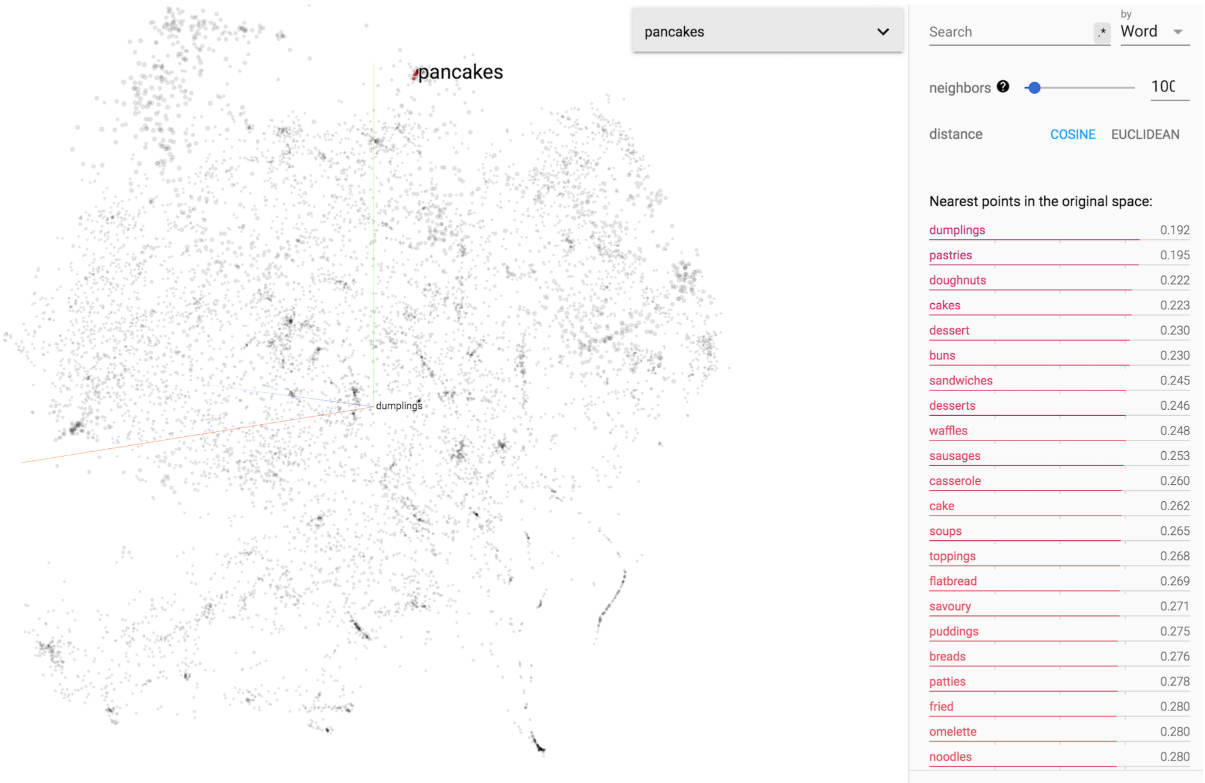


Figure 6.17: T-SNE plot of a trained SGNS model on a 2020 dump of Wikipedia.

## 6.3 Information Gain from Multi-modal Data

So far we compared our embedding spaces based on their cluster structure. In this section we move on to pillar 3 in our analysis. This second type of *transparency* analysis involved experiments for measuring similarity between distributions, based on an information-theoretical approach introduced in Section 2.7.5. We aim to measure the information gain $E_S$ and $E_V$ each contribute when combined with $E_L$. By treating the embedding spaces as samples from multivariate distributions we formulate the question in the following way: *Are two semantic spaces from different modalities independent from each other?*

We employ empirical Mutual Information prediction methods, described in Section 3.2.4.



pancakes

by Word

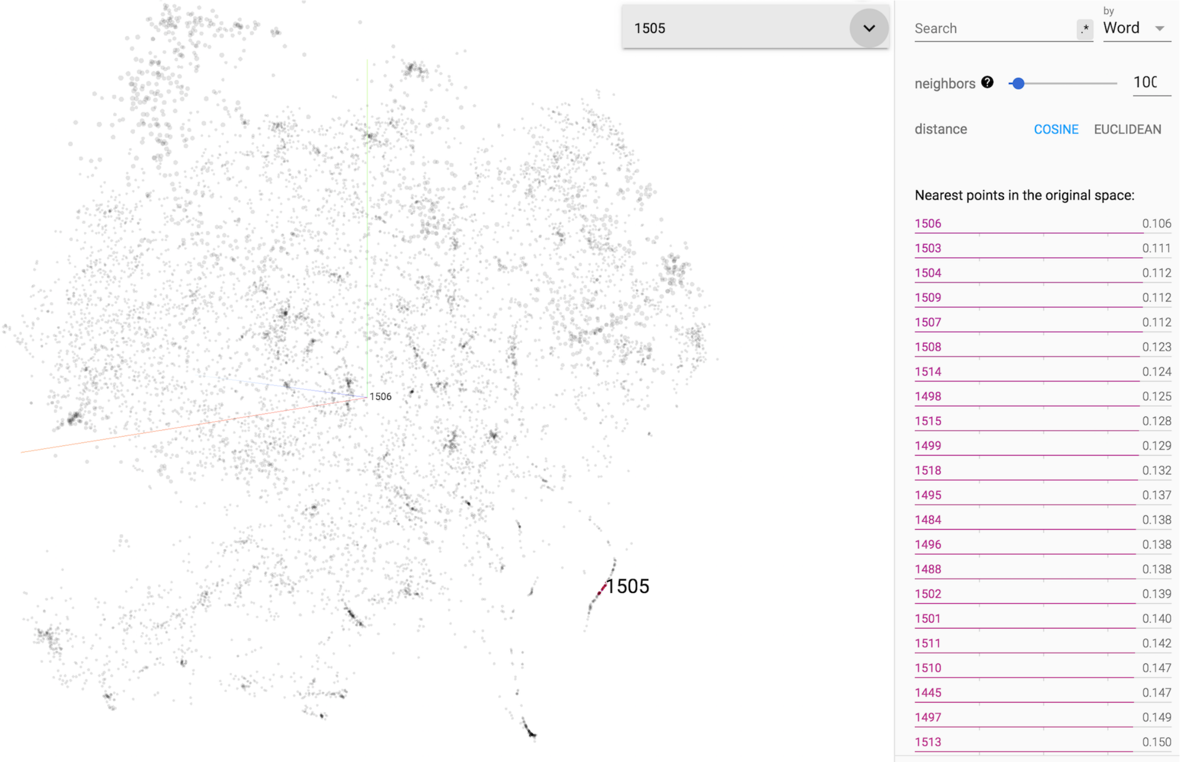
neighbors 100

distance COSINE EUCLIDEAN

Nearest points in the original space:

| | |
|---|---|
| dumplings | 0.192 |
| pastries | 0.195 |
| doughnuts | 0.222 |
| cakes | 0.223 |
| dessert | 0.230 |
| buns | 0.230 |
| sandwiches | 0.245 |
| desserts | 0.246 |
| waffles | 0.248 |
| sausages | 0.253 |
| casserole | 0.260 |
| cake | 0.262 |
| soups | 0.265 |
| toppings | 0.268 |
| flatbread | 0.269 |
| savoury | 0.271 |
| puddings | 0.275 |
| breads | 0.276 |
| patties | 0.278 |
| fried | 0.280 |
| omelette | 0.280 |
| noodles | 0.280 |

Figure 6.18: Cluster, containing the word "pancakes" on the T-SNE plot of a trained SGNS model on a 2020 dump of Wikipedia.



1505

by Word

neighbors 100

distance COSINE EUCLIDEAN

Nearest points in the original space:

| | |
|---|---|
| 1506 | 0.106 |
| 1503 | 0.111 |
| 1504 | 0.112 |
| 1509 | 0.112 |
| 1507 | 0.112 |
| 1508 | 0.123 |
| 1514 | 0.124 |
| 1498 | 0.125 |
| 1515 | 0.128 |
| 1499 | 0.129 |
| 1518 | 0.132 |
| 1495 | 0.137 |
| 1484 | 0.138 |
| 1496 | 0.138 |
| 1488 | 0.138 |
| 1502 | 0.139 |
| 1501 | 0.140 |
| 1511 | 0.142 |
| 1510 | 0.147 |
| 1445 | 0.147 |
| 1497 | 0.149 |
| 1513 | 0.150 |

Figure 6.19: Cluster, containing the number "1505" on the T-SNE plot of a trained SGNS model on a 2020 dump of Wikipedia.

Section 6.3.1 describes details of the analysis, results are presented in Section 6.3.2.[6]

[6] We would like to thank Zoltán Szabó for his counsel on the theoretical background for these studies.

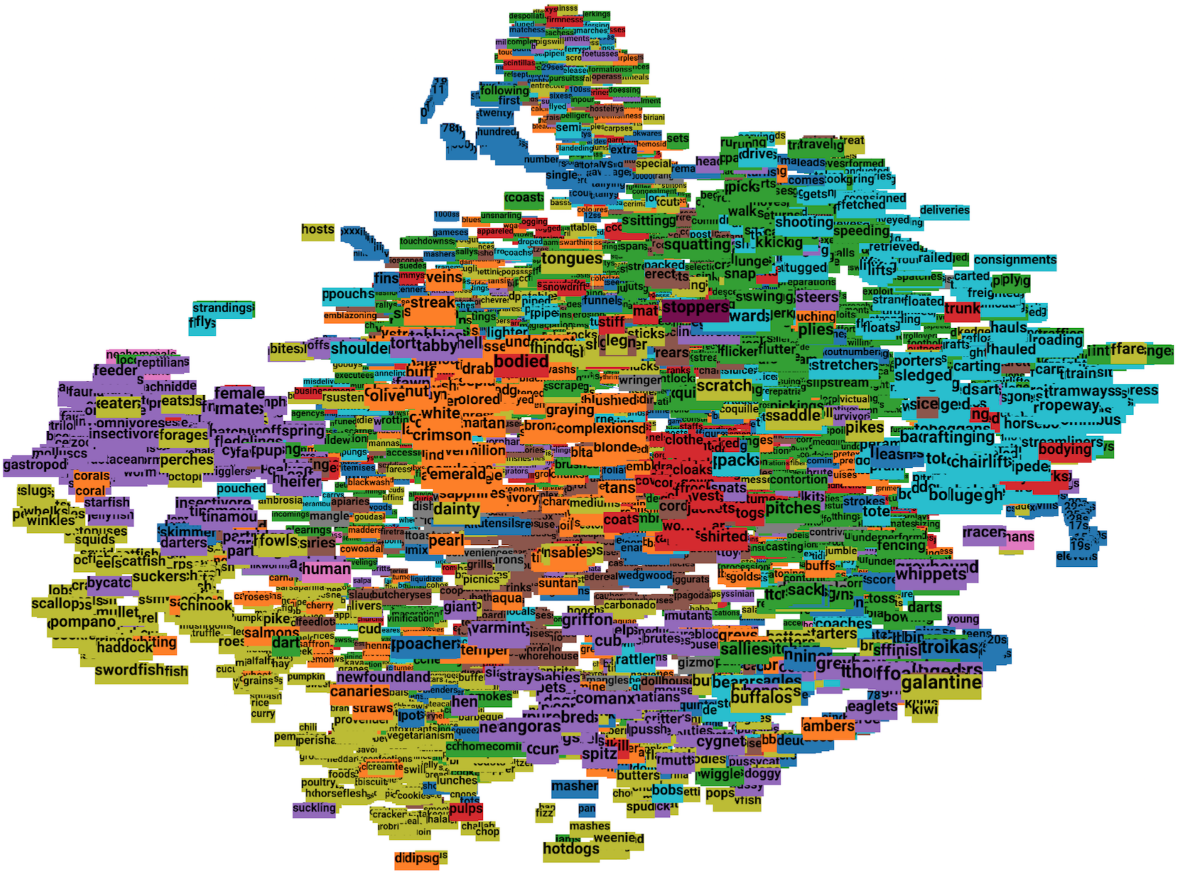


Figure 6.20: T-SNE plot of a trained SGNS model on a 2020 dump of Wikipedia. The colours correspond to 13 classes automatically generated using the WordNet hierarchy: *transport, food, building, animal, appliance, action, clothes, utensil, body, colour, electronics, number, human*

### 6.3.1 Hyper Parameters and Dimensionality Reduction

Since $I_{KNN}$ is not robust in very high dimensions we explore the hyper parameters of $I_{HSIC}$. We used the Gaussian Radial Basis Function (RBF) Kernel [Vert et al., 2004] with parameter settings $\sigma = 1$ and using median heuristic [Garreau et al., 2017].

Furthermore, in order to test the robustness of the results we ran the method after projecting our spaces onto lower dimensional spaces using Principal Component Analysis (PCA) [Wold et al., 1987]. We tested the embeddings with dimensions $d = \{10, 100, max\}$, where $max$ is the full dimension of each space. For further robustness, we ran the $I_{HSIC}$ algorithm for $d = \{3, 11, 12, 13, 50\}$ (Appendix E).

### 6.3.2 Results

The main benefit of this experiment is that we may be able to understand how data of different modalities contribute to the performance of multi-modal embeddings if they contribute at all. In case they do, is it just an artefact of introducing more data or is it due to meaningful information which changes the structure of the vector space in a useful way?

In Figure 6.25 and 6.26 axis $y$ shows $I(E_L, E_V)$ (red) and $I(E_L, E_S)$ (blue), where $I$ is the estimated Shannon mutual information using either a k-Nearest Neighbor based, linear algorithm ($I_{KNN}$) or the HSIC kernel method ($I_{HSIC}$).

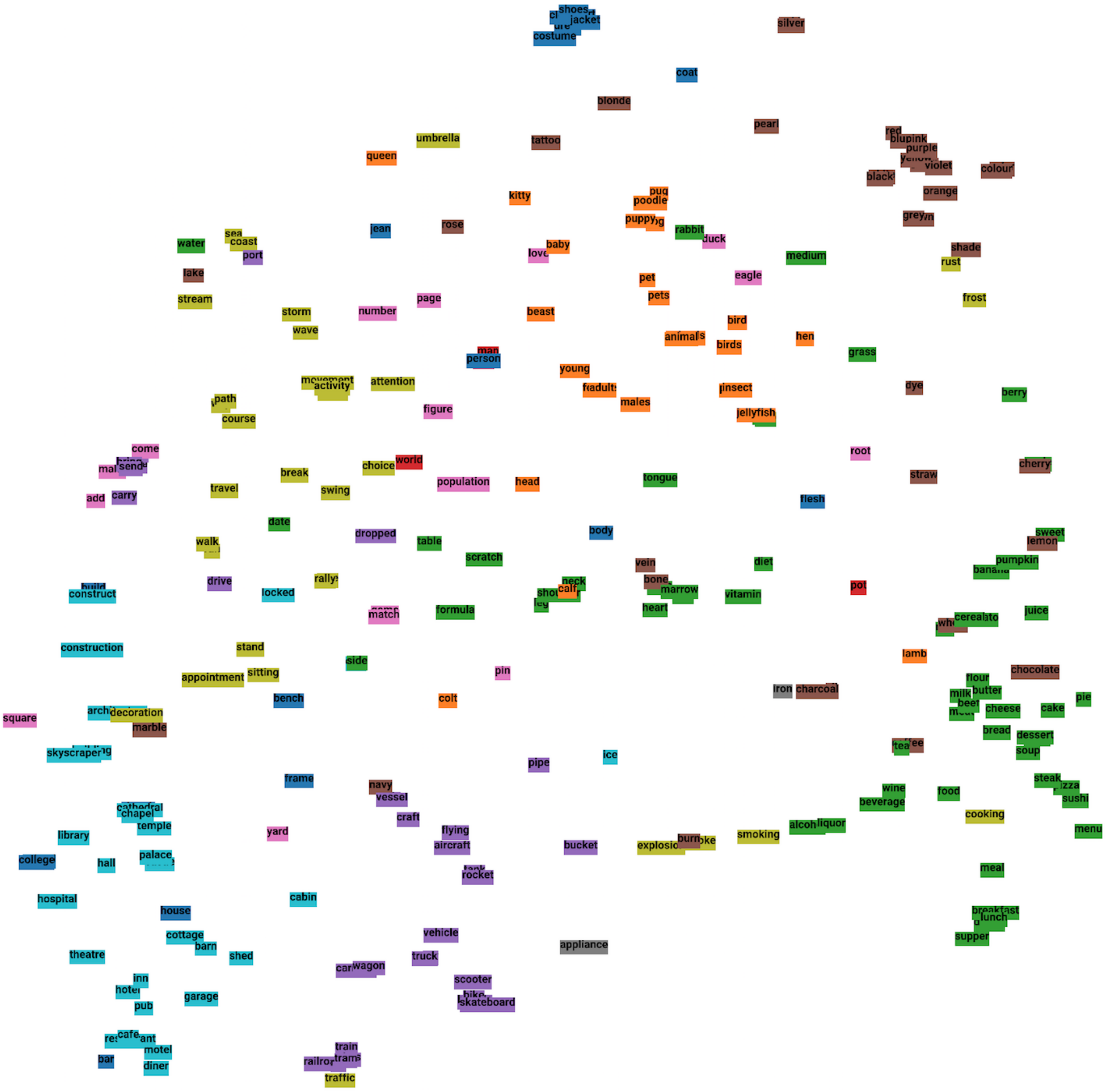


Figure 6.21: T-SNE plot of $E_L$ with its vocabulary restricted to the common subset of $E_L, E_V, E_S$ and the ones with an existing automatic WordNet class label, resulting 252 words. The colours correspond to 13 classes automatically generated using the WordNet hierarchy: *transport, food, building, animal, appliance, action, clothes, utensil, body, colour, electronics, number, human*

In Figure 6.25 axis $x$ represents the size of the training corpus $e_1, \ldots, e_N$ (in terms of the number of tokens) for $E_L$. Apart from $I_{HSIC}$ with $\sigma = 1$ the models agree on $I(E_L, E_V)$ being greater than $I(E_L, E_S)$, which suggests that the Visual Genome Scene Graph based structured embedding $E_S$ is "more independent" from the linguistic model $E_L$, than the image based $E_V$. This is surprising after observing the two models behaving similarly in Chapter 4. Moreover, the results are interesting, since, while the creation of this type of training data was highly visually directed, yet it is a text based model. Nevertheless, it is "farther" from the linguistic model in distribution than the visual one. $I(E_L, E_V)$ appears to be lower for lower volumes of text data. This may be because with more data they contain more related information. Although, in the case of $I_{HSIC}$ with maximal dimensions, using the median heuristic for $\sigma$ this pattern cannot be seen. In $I(E_L, E_S)$ no such tendency can be observed.

Figure 6.26 reports the effect of word frequency (in the $E_L$ training corpus) on the estimated $I$. Similarly to [Sahlgren and Lenci, 2016] we split the vocabulary into three

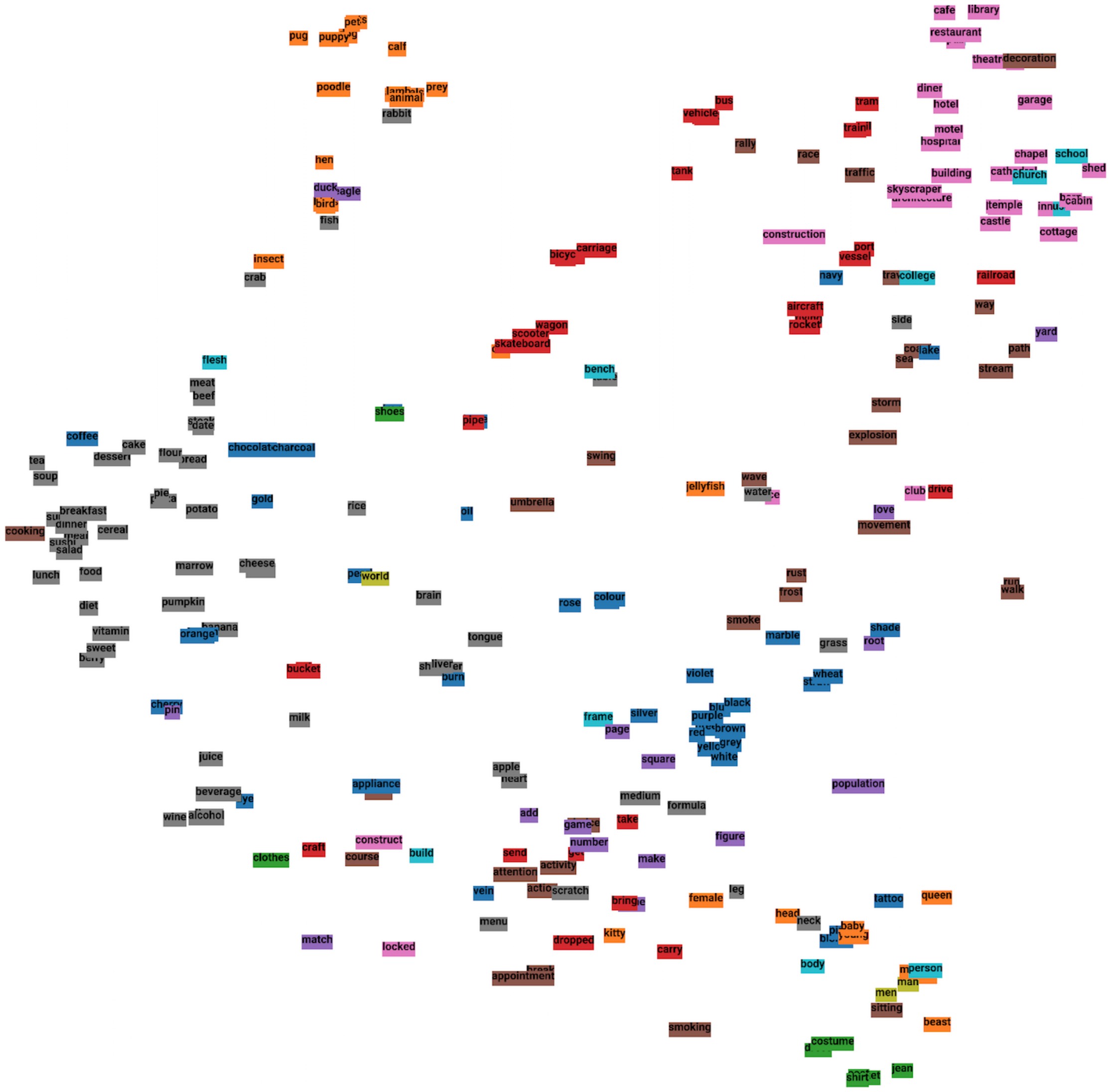


Figure 6.22: T-SNE plot of $E_V$ with its vocabulary restricted to the common subset of $E_L, E_V, E_S$ and the ones with an existing automatic WordNet class label, resulting 252 words. The colours correspond to 13 classes automatically generated using the WordNet hierarchy: *transport, food, building, animal, appliance, action, clothes, utensil, body, colour, electronics, number, human*

equally large parts; HIGH, MEDIUM and LOW range. This way we generate samples for $E_L, E_V$ and $E_S$ for the different frequency ranges in the text corpus. Again, higher mutual information between the linguistic and the visual embeddings can be observed. The negative $I_{KNN}$ in Figure 6.26a is due to the oscillating nature of the approximation, and shows that the k-Nearest Neighbor method is not robust enough in this high dimension.

In terms of the effect of word frequency, the only pattern that emerges is the relative low mutual information between $E_L$ and $E_V$ on low frequency words. However, this may be an artefact of sparse data, since the coverage drops dramatically with filtering pairs which fall in the same frequency category (see in Figure 5.2).

In order to further test the robustness of the results we ran the $I_{HSIC}$ algorithm for further dimensions in the very low range and one medium size: $d = \{3, 11, 12, 13, 50\}$. The results are shown in Appendix E. They support the the overall pattern in the above figures, adding that the results lose their robustness for $d = 3$.

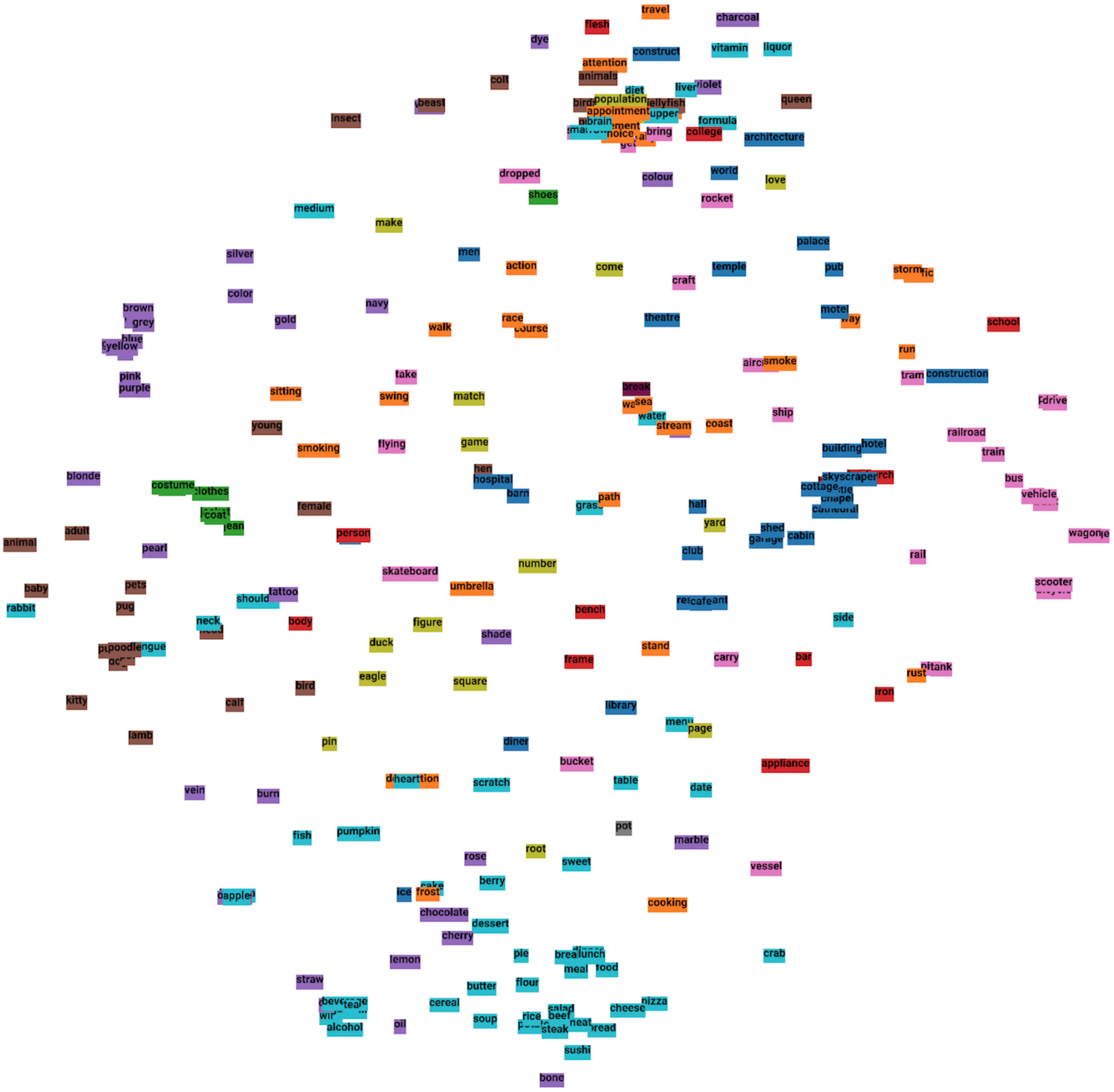


Figure 6.23: T-SNE plot of $E_S$ with its vocabulary restricted to the common subset of $E_L, E_V, E_S$ and the ones with an existing automatic WordNet class label, resulting 252 words. The colours correspond to 13 classes automatically generated using the WordNet hierarchy: *transport, food, building, animal, appliance, action, clothes, utensil, body, colour, electronics, number, human*

## 6.4 Dataset Distribution

Finally, we analyse the text based data source distributions $D_L$ and $D_S$ directly to get another perspective on the type of information they convey. We present words in the respective datasets with the 10 highest probability of co-occurrence with each centroid word from Section 6.2.2[7]. To estimate this probability we calculated Pointwise Mutual Information (PMI), Positive PMI (PPMI) (Equation 2.1), a modified PMI (PMI$^3$), $\chi^2$ [Manning and Schutze, 1999, Section 5.3.3.] and Fisher's exact test [Pedersen, 1996]. PMI$^3$ has an exponent of 3 for the numerator and no logarithm. We used the NLTK package implementations of all the above metrics[8].

[7] Duplicated words for appearing as left and right context as well are removed. Therefore the number of words are $\leq 10$.

[8] `https://www.nltk.org/api/nltk.html#module-nltk.collocations`

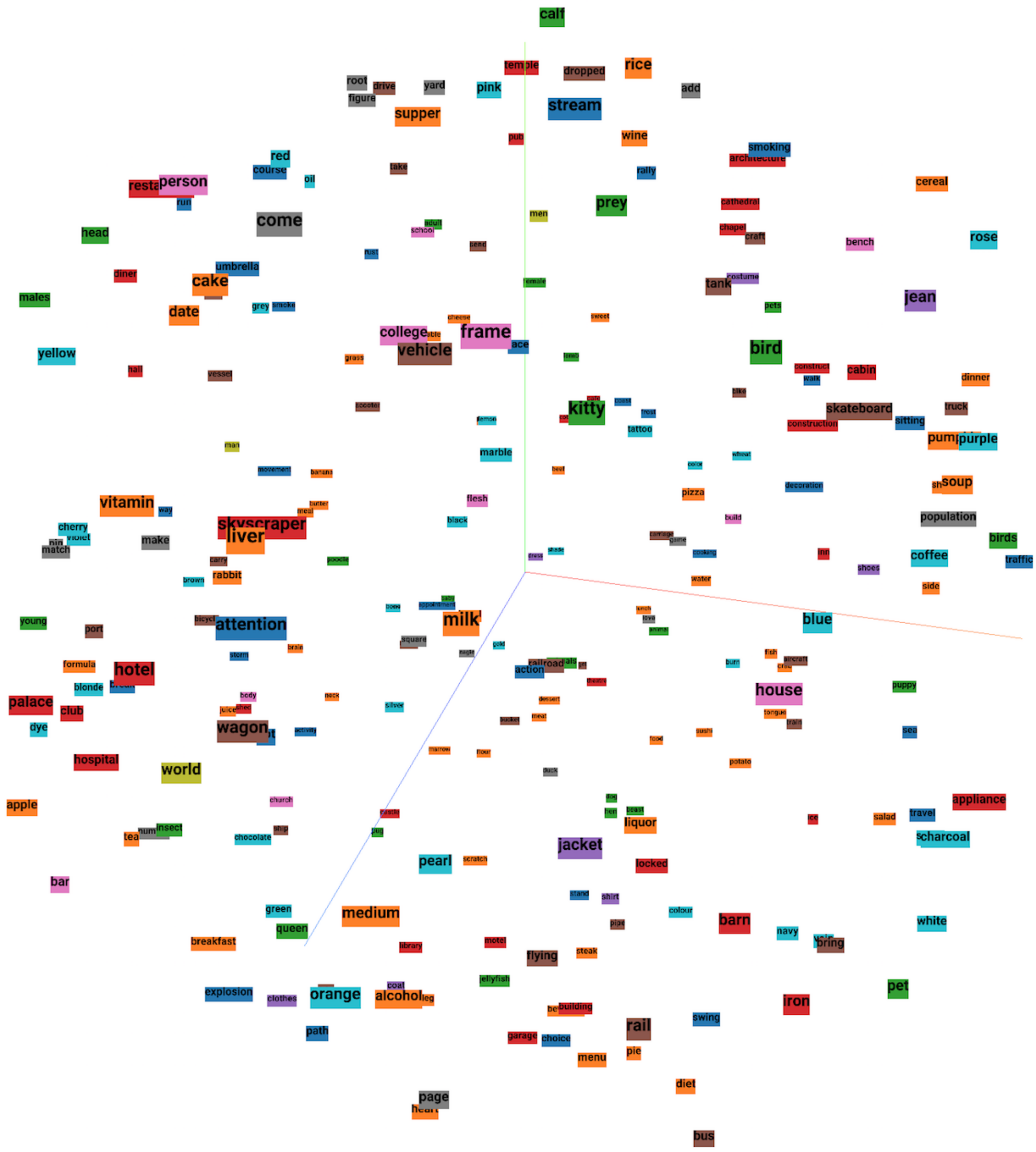


Figure 6.24: T-SNE plot of a random embedding $E_R \in \mathbb{R}^{252x300}$. The colours correspond to 13 classes automatically generated using the WordNet hierarchy: *transport, food, building, animal, appliance, action, clothes, utensil, body, colour, electronics, number, human.* The colour labels are evenly distributed on the projection.

Since PMI, PPMI and Fisher's test suffered from over-representing low frequency bigrams, we only present results for $\chi^2$ and PMI$^3$, which outputted fairly similar results. Table 6.5 presents examples for words closest to cluster centroids with the 10 highest $\chi^2$ score. Results for the full set of centroid words using $\chi^2$ and PMI$^3$ can be found in Appendix F.

| **Centroid** | **Wikipedia** | **Visual Genome** |
|---|---|---|
| plate | tectonics, nazca, restrictor, farallon, subducts, license, cribriform, tectonic, subducting, eurasian | plate, lying_on_top_of, on, has, on_top_of, in |
| rust | epique, cronartium, oleum, cohle, obritzberg, blister, belt, puccinia, windexed, colored | rust, stains_down, around_side_of, rusted_onto, on_fire, with_a_lot |

| | | |
|---|---|---|
| hummingbird | amazilia, selasphorus, mellisuga, calypte, cynanthus, berylline, scintillant, orthorhyncus, eupherusa, chinned | hummingbird, eat_nectar_from, in_flight_below, flapping_its, flapping, windspan |
| fun | poked, poking, pokes, poke, loving, lot, lovin, yidishn, fun, weäsell | are_having, are_having_great, fun, facing_away, planning, having |
| hand | right, sleight, grenades, left, hand, cranked, grenade, claps, gloved, upper | hand, holding, held_in, on, in_mans, man |
| bird | passerine, migratory, caged, sanctuary, watchers, watching, topley, species, prey, furnariidae | bird, perched_on, flying_in, flying_over, beak, flying_ahead_of |

Table 6.5: Example for context words of cluster centroids with the 10 highest $\chi^2$ score. See all cluster centroids in Appendix F.

The samples reveal that while Wikipedia includes more encyclopaedic synonyms as most likely bigrams, Visual Genome conveys more functional, specific type of contexts including more actions and attributes. For example “tectonics” in Wikipedia vs. “lying_on_top_of” in Visual Genome as the most likely co-occurrence for “plate”.

Our observations are in line with the word distributions in VG published in [Krishna et al., 2016]. The most common concepts (Figure 6.27), objects (Figure 6.28), attributes (Figure 6.29) and relationships (Figure 6.30) all paint a picture of how visually oriented VG annotations are. The published statistics also support our observation that VG mostly includes specific descriptions of smaller scenes.

These support our previous findings that Visual Genome can contribute with complementary information to a text based meaning representation by having denser annotations of visual scenes.

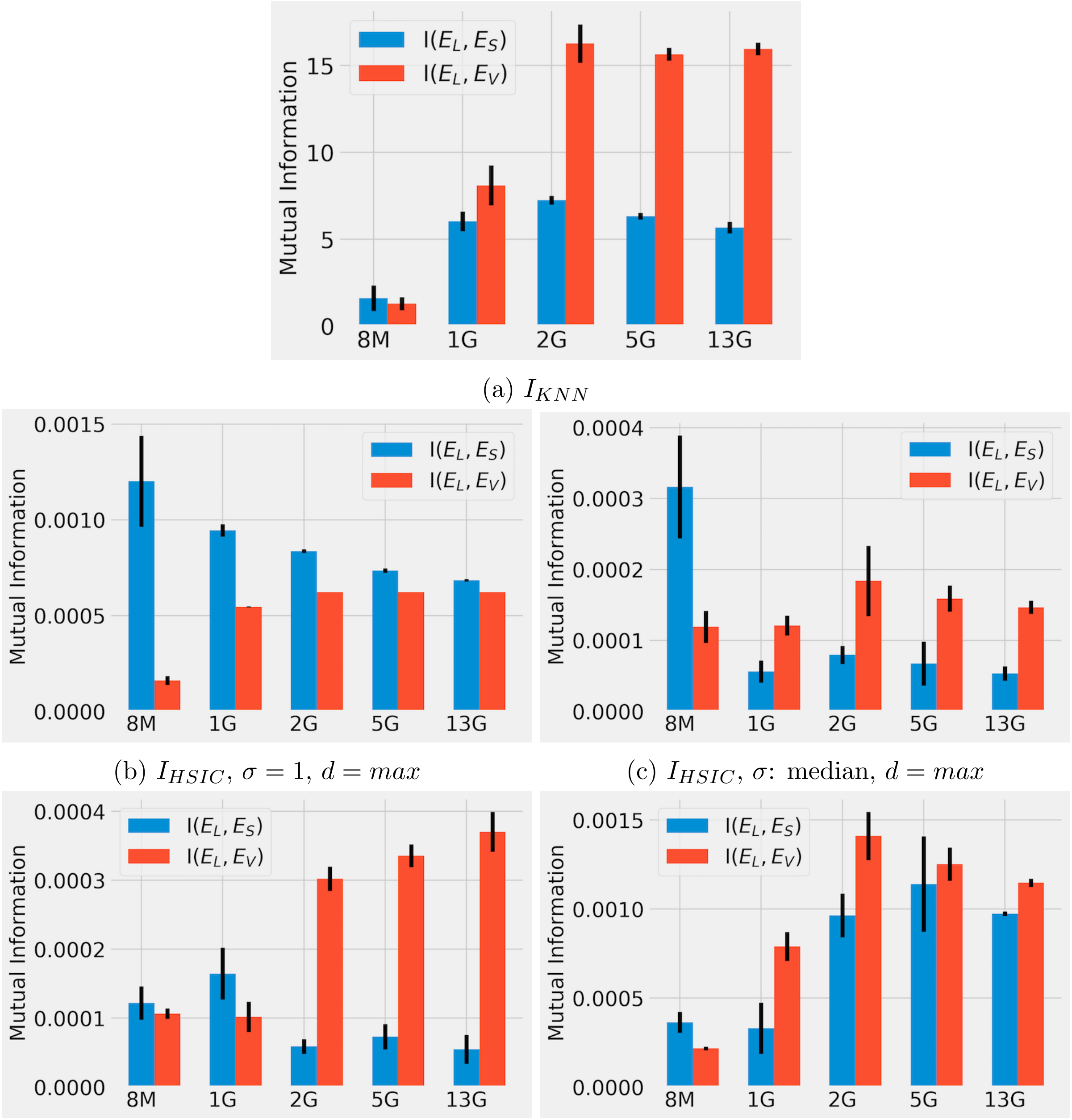
## 6.5 Conclusion

In this chapter we presented proof-of-concept studies of interpretable *Transparency analysis*, forming the second and third pillars of our analysis (Section 3.3).

***Qualitative / Quantitative Structural Analysis*** Firstly, our aim was to interpret our models by zooming into the distributional properties of linguistic, visual, structured and multi-modal embeddings. We ran K-means and Agglomerative clusterings on each embedding and used standard clustering metrics for evaluation when class labels are not given. The results indicate that while the image based model may have better defined clusters, the Visual Genome Scene Graph structured model can outperform the other ones in terms of consistency when the number of clusters are chosen well. We visualised the clustered embeddings and inspected the individual clusters from the the best K-means clustering. We introduced a WordNet based cluster label annotation technique. Furthermore, we compared the clustering to Agglomerative Clustering results.

The supervised T-SNE visualisations provide further insight into the structure of our semantic spaces, which are in line with the above findings. We introduced a simple method to automatically annotate our data with topic labels saving huge amount of human effort. Remarkably, the results already give further insight into our data, despite the simple heuristic of label generation. We believe the method could be easily improved to gain better coverage on the vocabulary and higher accuracy of labels.

***Independence Analysis*** Secondly, we created an implementation of our information theory based framework to measure the information gain visual and structured embed-



(a) $I_{KNN}$

(b) $I_{HSIC}$, $\sigma = 1$, $d = max$

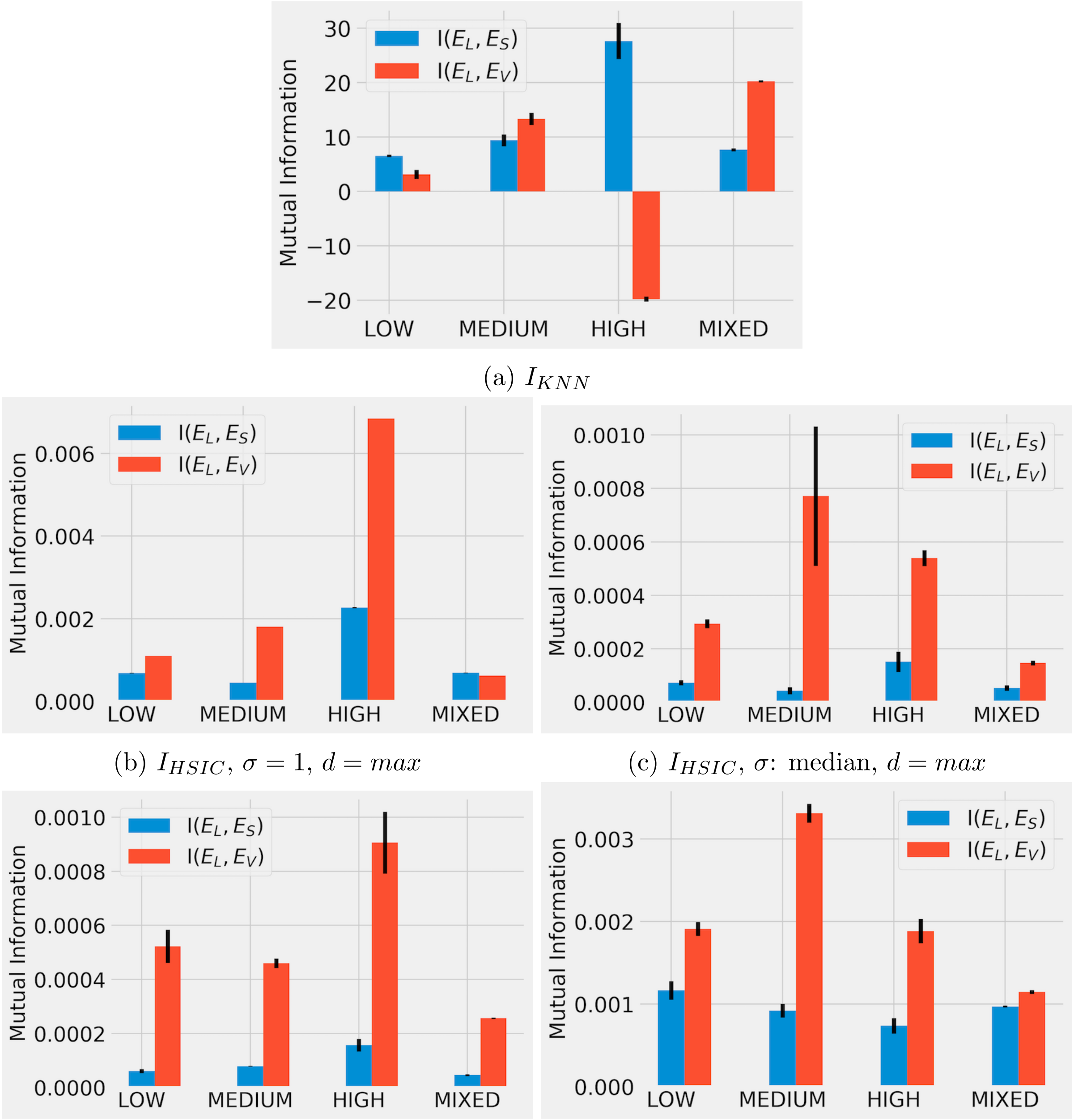
(c) $I_{HSIC}$, $\sigma$: median, $d = max$

(d) $I_{HSIC}$, $\sigma$: median, $d = 100$

(e) $I_{HSIC}$, $\sigma$: median, $d = 10$

Figure 6.25: Estimated Mutual Informations: $I(E_L, E_V)$ (red) and $I(E_L, E_S)$ (blue) for different corpus sizes.

dings may provide by combining them with text based linguistic models. We found that the Visual Genome SceneGraph based structured model is more independent from the Wikipedia based SGNS model than the visual embeddings, trained on images. This may reveal something about why this structural data on its own, as well as combined with linguistic information, can achieve such high accuracies, despite having orders of magnitude less training data than either of the other modalities (as we saw in Chapter 5). Analysing the effect of VG and image data size on this metric would be an important future direction, as we saw that the mutual information of image and text based embeddings increase with corpus size. However, in the context of the structured model's comparable performance, we think that the estimated mutual information is a promising metric for deciding over the usefulness of a new data source.



(a) $I_{KNN}$

(b) $I_{HSIC}$, $\sigma = 1$, $d = max$

(c) $I_{HSIC}$, $\sigma$: median, $d = max$

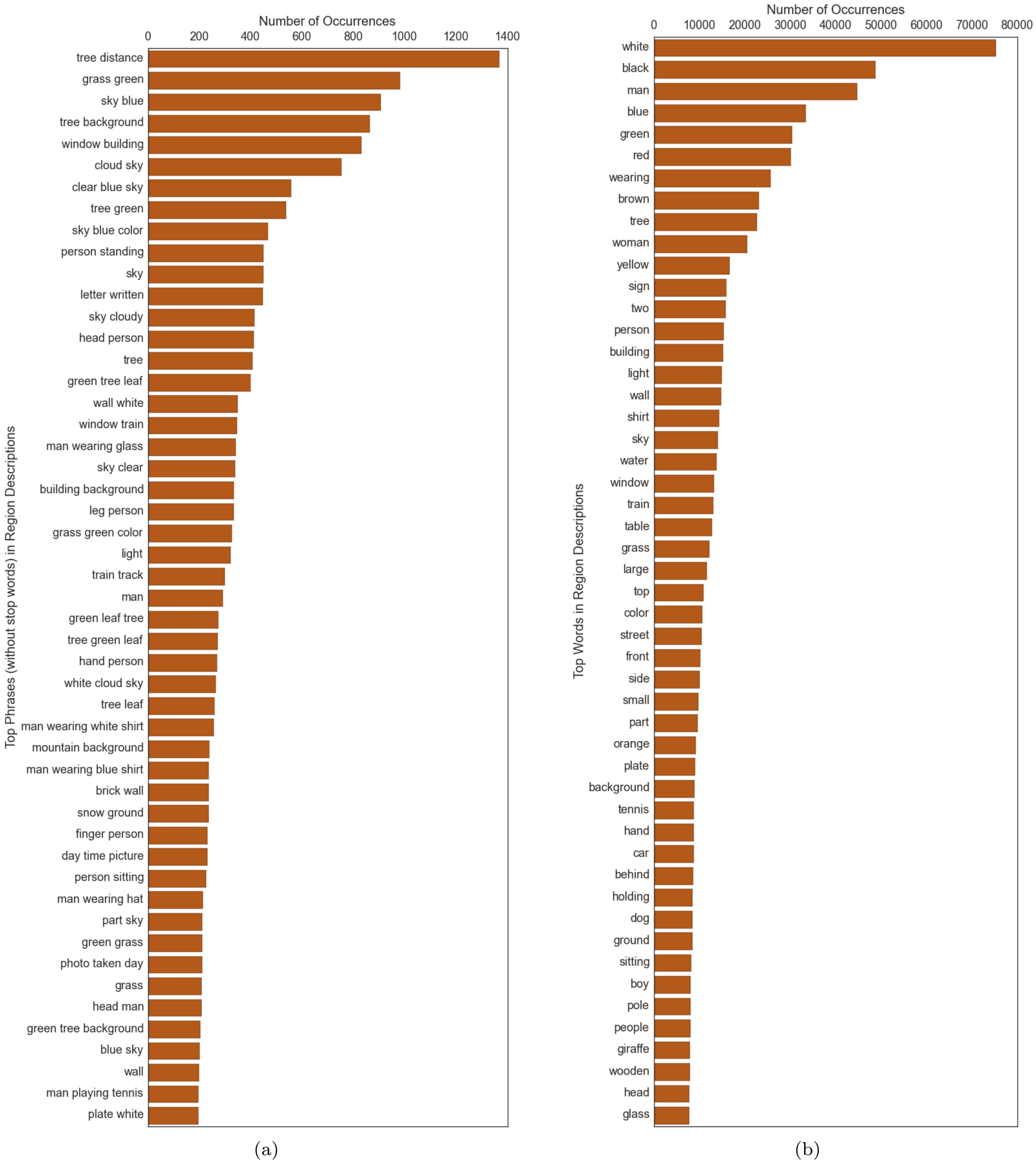
(d) $I_{HSIC}$, $\sigma$: median, $d = 100$

(e) $I_{HSIC}$, $\sigma$: median, $d = 10$

Figure 6.26: Estimated Mutual Informations: $I(E_L, E_V)$ (red) and $I(E_L, E_S)$ (blue) for different word frequency ranges.

**Summary of Transparency Analysis** Let us examine the two hypotheses we made in Section 6.1. All three embedding types show different cluster structures, however, the image based embedding is closer to the linguistic one than our visually structured, textual embedding: both in terms of cluster structure as well as being more mutually dependent. Considering this result in relation to the performance numbers in the previous chapters, we conclude that the image based embedding requires orders of magnitude more data and training time, while not necessarily providing additional useful information to a text based representation in the context of word semantic similarity. Therefore, we weakly reject Hypothesis I. On the other hand, based on the three pillars of our analyses: 1. reaching comparable performance despite being based on a small model trained on small data, 2. the quantitative and qualitative analysis of its cluster structure and 3. independence analysis, we conclude that our structured embedding provides complementary information to our linguistic representation while being highly efficient. Hence, we accept Hypothesis II.

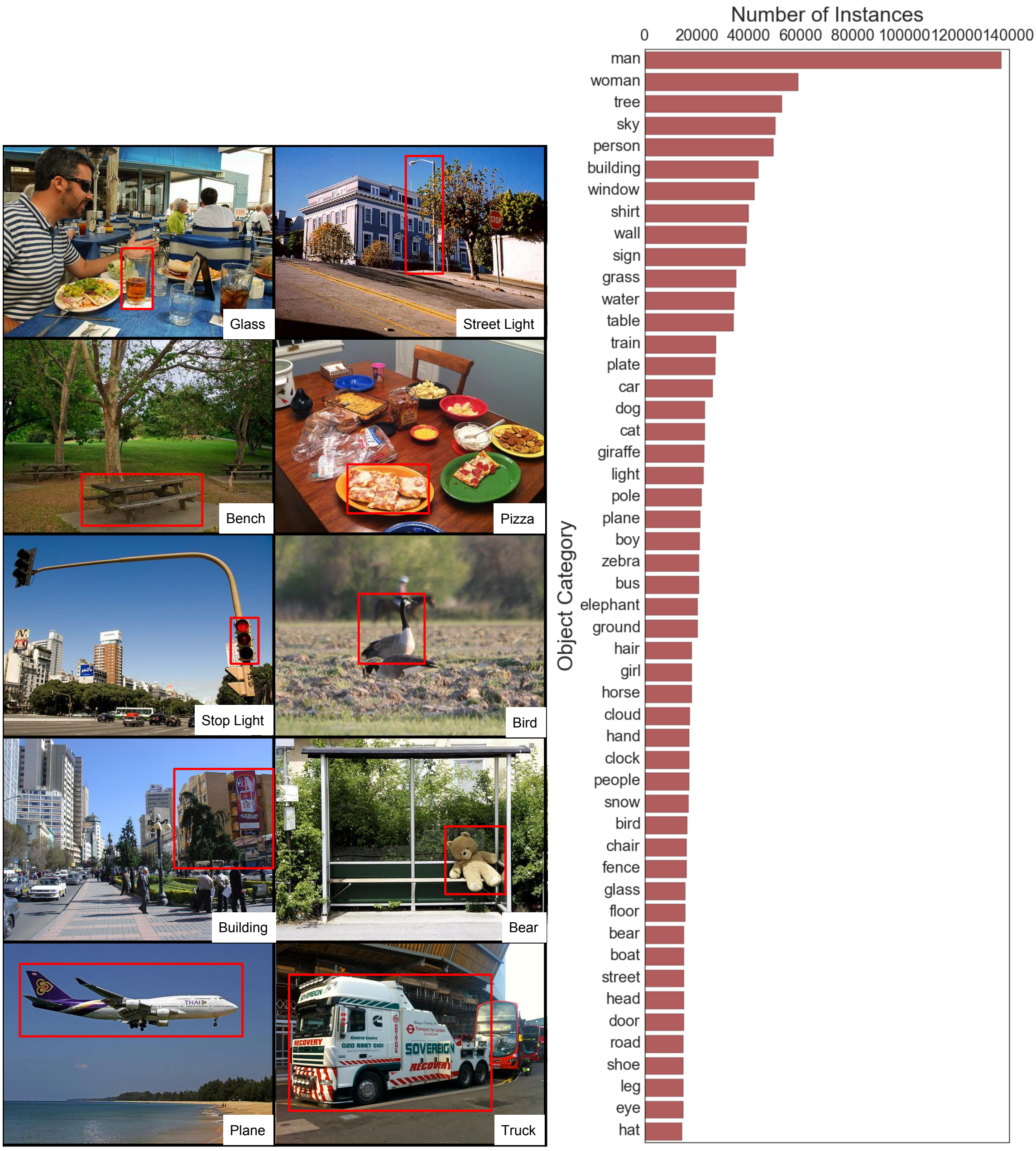
Investigating transformers, Bayesian MI estimators and other evaluations could be potential extensions of these studies. Applying automatically generated scenes graphs [Xu et al., 2020] would mitigate the main limitation of this approach, which is the manual labour required for creating VG. This would serve as a highly effective tool with important applications for low resource languages.



(a)

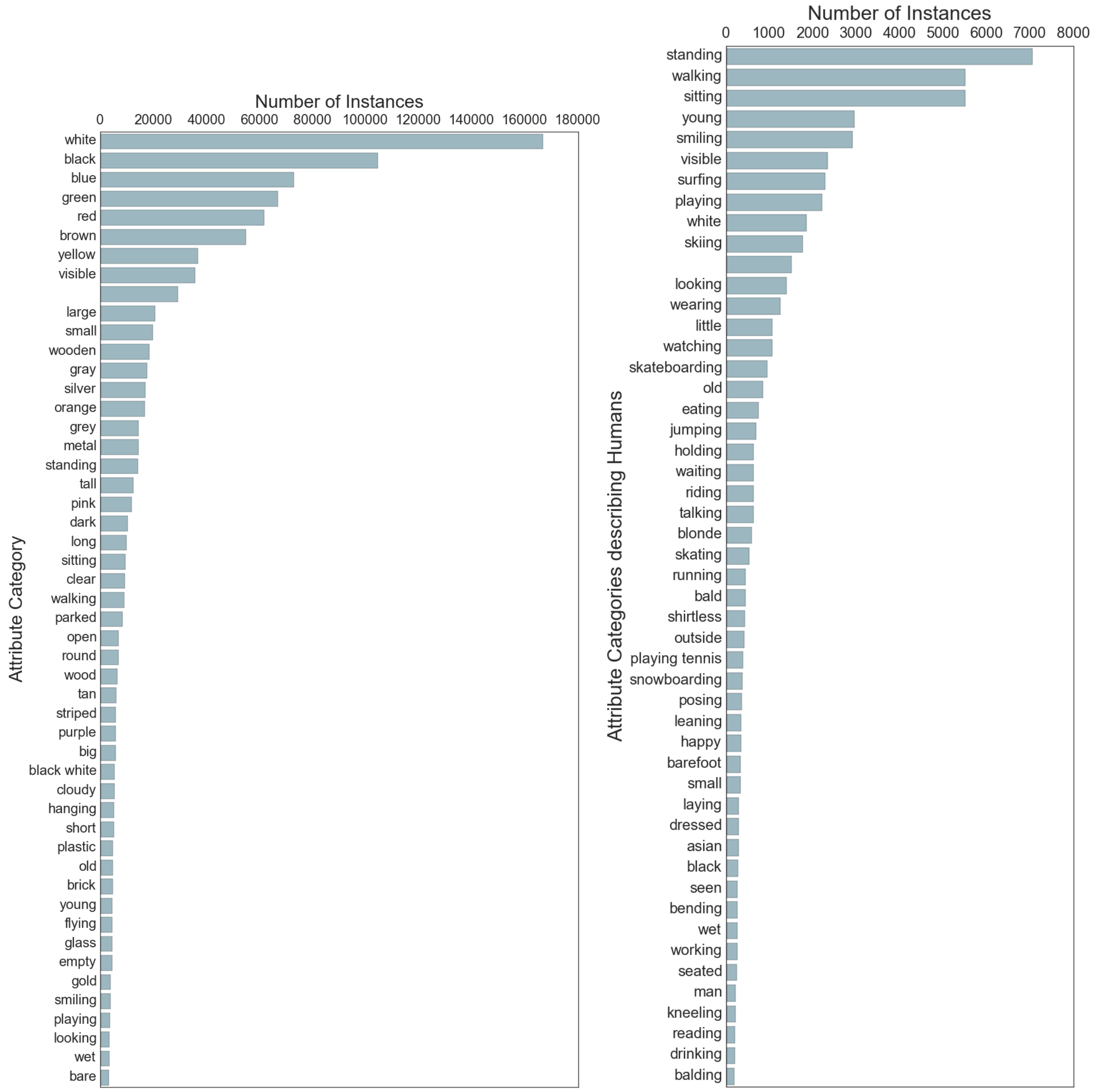
(b)

Figure 6.27: (a) A plot of the most common visual concepts or phrases that occur in region descriptions. The most common phrases refer to universal visual concepts like “blue sky,” “green grass,” etc. (b) A plot of the most frequently used words in region descriptions. Colours occur the most frequently, followed by common objects like “man” and “dog” and universal visual concepts like “sky.” [Krishna et al., 2016]



(a) (b)

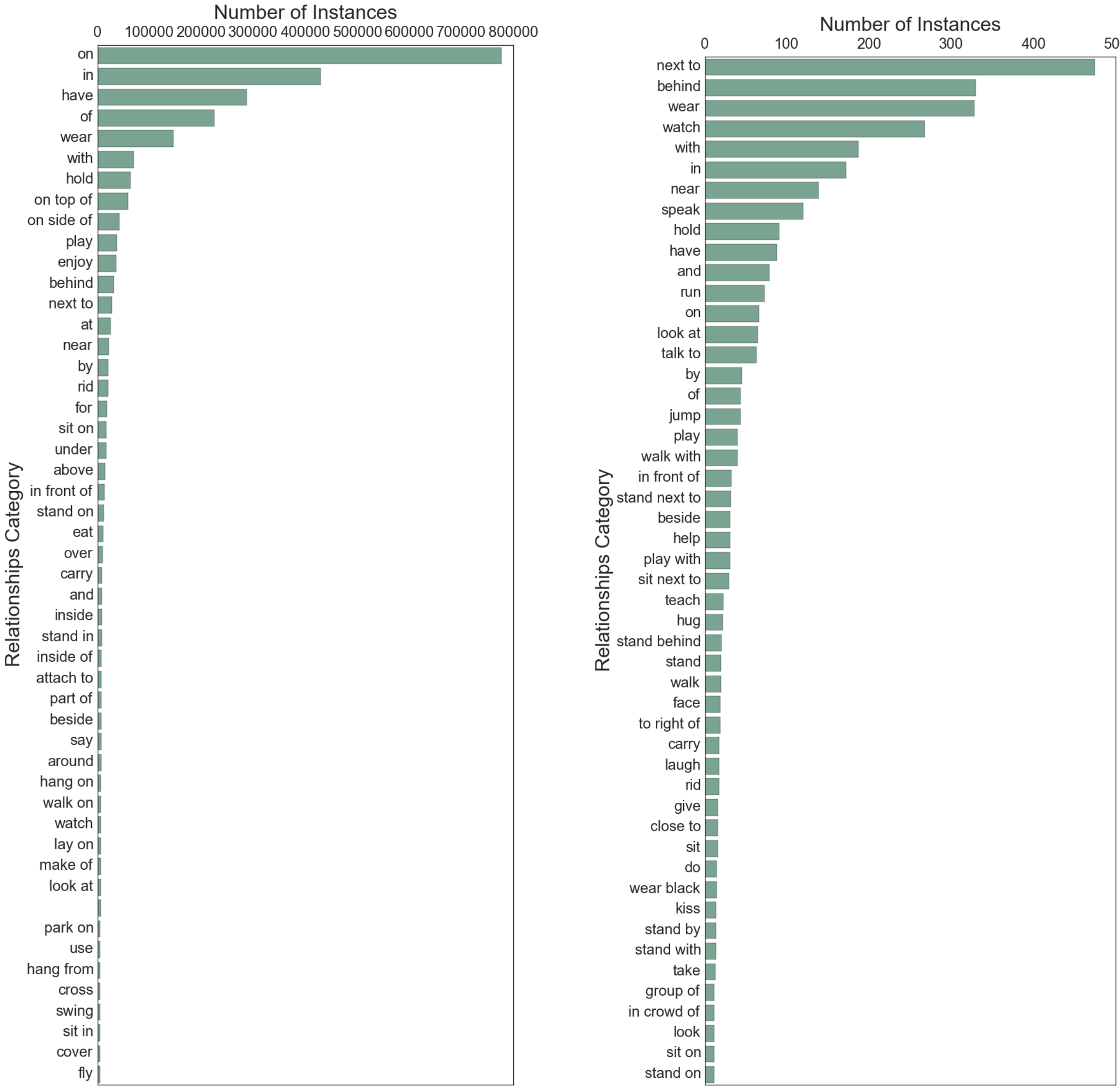
Figure 6.28: (a) Examples of objects in VG. Each object is localized in its image with a tightly drawn bounding box. (b) Plot of the most frequently occurring objects in images. People are the most frequently occurring objects in the dataset, followed by common objects and visual elements like “building”, “shirt”, and “sky”. [Krishna et al., 2016]



(a)

(b)

Figure 6.29: (a) Distribution showing the most common attributes in VG. Colours ("white", "red") and materials ("wooden", "metal") are the most common. (b) Distribution showing the number of attributes describing people. State-of-motion verbs ("standing", "walking") are the most common, while certain sports ("skiing", "surfing") are also highly represented due to an image source bias in the image set. [Krishna et al., 2016]



(a)

(b)

Figure 6.30: (a) A sample of the most frequent relationships in VG. In general, the most common relationships are spatial ("on top of", "on side of", etc.). (b) A sample of the most frequent relationships involving humans in the dataset. The relationships involving people tend to be more action oriented ("walk", "speak", "run", etc.). [Krishna et al., 2016]

# Chapter 7

# Summary and Conclusions

This thesis has been pursuing a better understanding of the impact of visual information on semantic models in non-visual tasks. Since the literature is narrower and more inconclusive on these tasks, here we aimed for constructing a broader evaluation and analysis. We introduced a general embedding formalism and a three pillar framework for transparent analysis of multi-modal semantic embedding models. We proposed and implemented a new type of embedding in between linguistic and visual modalities, based on small data. We analysed its contribution to linguistic representations within our analytical framework. Furthermore, we presented and showcased a framework for treating modalities as partial observers of meaning based on information-theory.

## 7.1 Main Findings

The main findings are the following:

- The source of images affect the performance of multi-modal mid-fused semantic representations.

- The number of images in ordered sources has an impact on performance, but it stabilizes at around 10-20 images.

- Visual information can be complementary for smaller linguistic corpora, but this effect does not necessarily scale with corpus size.

- Images convey complementary statistical information about the co-occurrence of objects in visual scenes, but there is no direct indication of how low level visual features contribute.

- Cluster analysis can provide a useful framework for analysing emergent concept structures. Combined with independence analysis they can serve as a useful framework for transparent embedding analysis.

- VG Scene Graph based, visually structured, textual models achieve comparable or better performance in an economic way, by using orders of magnitude less resources than visual models. When combined, it enriches our linguistic model with more divergent information than the image based one. Its clusters represent more concrete concepts, in-between visual and linguistic domains.

## 7.2 Conclusion and Future Work

Instead of comparing all the latest models at the time, we developed a general analysis framework and presented proof-of-concept studies, which can be applied to various models in the future. To present our methodology, we employed the smallest possible models which allow us to incorporate visual embeddings, thus studying multi-modality. Therefore, in this work we applied the shallow skip-gram network, as visual embeddings fit into them more easily then into count based models, while being the simplest neural models. Furthermore, we used mid-fusion technique, which made it straightforward to study individual modalities. Incorporating this methodology to the evaluation of various recent models would be the next step.

In parallel, the analysis methodology can also be further developed. One direction is to test the level of visual information that impacts abstract semantic representations. One potential test is to gradually reduce the resolution of images we use for visual embeddings and see how the performance changes, in what rate it starts to decline in particular. This way we would see how much visual detail can be omitted while keeping the same gain for conceptually abstract tasks.

Another exciting direction would be to extend the notion of modality and compare semantic representations trained across different data sources in general, such as corpora of different authors, from different times or different styles and social circles. Further extension of the notion of semantic representation could be measuring semantic change in time, such as the polarisation of political discourse. This has the potential to have positive social impact if we are capable of detecting the time and "place" of the source of miscommunication.

Applying automatically generated scenes graphs would mitigate the main limitation of the presented Visual Genome based approach, which is the manual labour required for creating it. This would serve as a highly effective tool with important applications for low resource languages.

For measuring information gain experimenting with Bayesian Mutual Information estimation methods and other evaluation and training datasets would also be a viable future route.

Understanding the information our various data sources convey and the biases our different models have on them is an essential work in Artificial Intelligence. Data driven AI applications surround us, thus we believe there is a surging need for such meta analyses in order to advance this technology in a more conscious way.

# Bibliography

[Agrawal et al., 2016] Agrawal, A., Batra, D., and Parikh, D. (2016). Analyzing the behavior of visual question answering models. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1955–1960.

[Anderson et al., 2017] Anderson, A. J., Kiela, D., Clark, S., and Poesio, M. (2017). Visually grounded and textual semantic models differentially decode brain activity associated with concrete and abstract nouns. *Transactions of the Association for Computational Linguistics*, 5:17–30.

[Anderson et al., 2016] Anderson, A. J., Zinszer, B. D., and Raizada, R. D. (2016). Representational similarity encoding for fmri: Pattern-based synthesis to predict brain activity using stimulus-model-similarities. *NeuroImage*, 128:44–53.

[Antol et al., 2015] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. (2015). Vqa: Visual question answering. *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

[Artetxe et al., 2018] Artetxe, M., Labaka, G., and Agirre, E. (2018). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. *ACL*.

[Arthur and Vassilvitskii, 2006] Arthur, D. and Vassilvitskii, S. (2006). k-means++: The advantages of careful seeding. *Stanford*.

[Arthur et al., 2016] Arthur, P., Neubig, G., and Nakamura, S. (2016). Incorporating discrete translation lexicons into neural machine translation. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1557–1567.

[Bahdanau et al., 2015] Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural Machine Translation By Jointly Learning To Align and Translate. *Iclr 2015*, 26(1):1–15.

[Barocas et al., 2019] Barocas, S., Hardt, M., and Narayanan, A. (2019). *Fairness and Machine Learning*. `http://www.fairmlbook.org`.

[Baroni and Lenci, 2008] Baroni, M. and Lenci, A. (2008). Concepts and properties in word spaces. *Italian Journal of Linguistics*, 20(1):55–88.

[Batchkarov et al., 2016] Batchkarov, M., Kober, T., Reffin, J., Weeds, J., and Weir, D. (2016). A critique of word similarity as a method for evaluating distributional semantic models. *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 7–12.

[Bender et al., 2021] Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? 🦜. *Proceedings of FAccT 2021*.

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

[Bergsma and Goebel, 2011] Bergsma, S. and Goebel, R. (2011). Using visual information to predict lexical preference. *Proceedings of RANLP*, pages 399–405.

[Boleda, 2020] Boleda, G. (2020). Distributional semantics and linguistic theory. *Annual Review of Linguistics*, 6:213–234.

[Bowker and Star, 2000] Bowker, G. C. and Star, S. L. (2000). *Sorting things out: Classification and its consequences.*

[Bowman et al., 2015] Bowman, S., Angeli, G., Potts, C., and Manning, C. D. (2015). A large annotated corpus for learning natural language inference. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

[Bruni et al., 2014] Bruni, E., Tran, N.-K., and Baroni, M. (2014). Multimodal distributional semantics. *J. Artif. Intell. Res.(JAIR)*, 49(2014):1–47.

[Bucci, 1985] Bucci, W. (1985). Dual coding: A cognitive model for psychoanalytic research. *Journal of the American Psychoanalytic Association*, 33(3):571–607.

[Bulat et al., 2017] Bulat, L., Clark, S., and Shutova, E. (2017). Speaking, seeing, understanding: Correlating semantic models with conceptual representation in the brain. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1081–1091.

[Cho et al., 2014] Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734.

[Chomsky et al., 2000] Chomsky, N. et al. (2000). *New horizons in the study of language and mind.*

[Clark, 2015] Clark, S. (2015). Vector space models of lexical meaning. *Handbook of Contemporary Semantics*, 10:9781118882139.

[Conneau et al., 2018] Conneau, A., Kruszewski, G., Lample, G., Barrault, L., and Baroni, M. (2018). What you can cram into a single &!#* vector: Probing sentence embeddings for linguistic properties. *ACL 2018-56th Annual Meeting of the Association for Computational Linguistics*, 1:2126–2136.

[Cover and Thomas, 2012] Cover, T. and Thomas, J. (2012). *Elements of Information Theory.*

[Davis et al., 2019] Davis, C., Bulat, L., Verő, A. L., and Shutova, E. (2019). Deconstructing multimodality: visual properties and visual context in human semantic processing. *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (* SEM 2019)*, pages 118–124.

[Deerwester et al., 1990] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.

[Deng et al., 2009] Deng, J., Dong, W., Socher, R., Li, L., Li, K., and Li, F. (2009). Imagenet: A large-scale hierarchical image database. *Proceedings of CVPR*, pages 248–255.

[Devlin et al., 2019] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL-HLT (1).*

[Dinu et al., 2015] Dinu, G., Lazaridou, A., and Baroni, M. (2015). Improving zero-shot learning by mitigating the hubness problem. *International Conference on Learning Representations, Workshop Track.*

[Dubossarsky et al., 2019] Dubossarsky, H., Hengchen, S., Tahmasebi, N., and Schlechtweg, D. (2019). Time-out: Temporal referencing for robust modeling of lexical semantic change. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 457–470.

[Erk, 2016] Erk, K. (2016). What do you know about an alligator when you know the company it keeps? *Semantics and Pragmatics*, 9:17–1.

[Ernst and Banks, 2002] Ernst, M. O. and Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870):429.

[Faruqui et al., 2016] Faruqui, M., Tsvetkov, Y., Rastogi, P., and Dyer, C. (2016). Problems with evaluation of word embeddings using word similarity tasks. *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 30–35.

[Fergus et al., 2005] Fergus, R., Li, F., Perona, P., and Zisserman, A. (2005). Learning object categories from Google's image search. *Proceedings of ICCV*, pages 1816–1823.

[Firth, 1957] Firth, J. R. (1957). A synopsis of linguistic theory. *Studies in Linguistic Analysis, Oxford: Philological Society, (1–32), reprinted in F.R. Palmer (ed.), Selected Papers of J.R. Firth 1952-1959, London: Longman (1968).*

[Fouhey and Zitnick, 2014] Fouhey, D. F. and Zitnick, C. L. (2014). Predicting object dynamics in scenes. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2019–2026.

[Gabrilovich et al., 2007] Gabrilovich, E., Markovitch, S., et al. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. *IJcAI*, 7:1606–1611.

[Garreau et al., 2017] Garreau, D., Jitkrittum, W., and Kanagawa, M. (2017). Large sample analysis of the median heuristic. *arXiv preprint arXiv:1707.07269.*

[Gasparri and Marconi, 2021] Gasparri, L. and Marconi, D. (2021). Word Meaning. *The Stanford Encyclopedia of Philosophy.*

[Gerz et al., 2016] Gerz, D., Vulić, I., Hill, F., Reichart, R., and Korhonen, A. (2016). SimVerb-3500: A Large-Scale Evaluation Set of Verb Similarity. *EMNLP.*

[Ghorbani et al., 2019] Ghorbani, A., Wexler, J., Zou, J. Y., and Kim, B. (2019). Towards automatic concept-based explanations. *Advances in Neural Information Processing Systems*, 32:9277–9286.

[González et al., 2006] González, J., Barros-Loscertales, A., Pulvermüller, F., Meseguer, V., Sanjuán, A., Belloch, V., and Ávila, C. (2006). Reading cinnamon activates olfactory brain regions. *Neuroimage*, 32(2):906–912.

[Gretton et al., 2005] Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005). Measuring statistical dependence with hilbert-schmidt norms. *International conference on algorithmic learning theory*, pages 63–77.

[Grice, 1975] Grice, H. P. (1975). Logic and conversation. pages 41–58.

[Gupta et al., 2019] Gupta, T., Schwing, A., and Hoiem, D. (2019). Vico: Word embeddings from visual co-occurrences. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7425–7434.

[Handjaras et al., 2016] Handjaras, G., Ricciardi, E., Leo, A., Lenci, A., Cecchetti, L., Cosottini, M., Marotta, G., and Pietrini, P. (2016). How concepts are encoded in the human brain: a modality independent, category-based cortical organization of semantic knowledge. *Neuroimage*, 135:232–242.

[Harnad, 1990] Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346.

[He et al., 2016] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

[Herbelot, 2020] Herbelot, A. (2020). Re-solve it: simulating the acquisition of core semantic competences from small data. *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 344–354.

[Hill et al., 2015] Hill, F., Reichart, R., and Korhonen, A. (2015). SimLex-999: Evaluating Semantic Models With (Genuine) Similarity Estimation. *Association for Computational Linguistics.*

[Hinton et al., 2012] Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580.*

[Hooker, 2021] Hooker, S. (2021). Moving beyond "algorithmic bias is a data problem". *Patterns*, 2(4):100241.

[Jitkrittum et al., 2017] Jitkrittum, W., Szabó, Z., and Gretton, A. (2017). An adaptive test of independence with analytic kernel embeddings. *International Conference on Machine Learning*, pages 1742–1751.

[Johnson et al., 2017] Johnson, J., Hariharan, B., Van Der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., and Girshick, R. (2017). Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910.

[Kabbach et al., 2019] Kabbach, A., Gulordava, K., and Herbelot, A. (2019). Towards incremental learning of word embeddings using context informativeness. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 162–168.

[Kaur et al., 2020] Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., and Wortman Vaughan, J. (2020). Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

[Kay et al., 2015] Kay, M., Matuszek, C., and Munson, S. A. (2015). Unequal representation and gender stereotypes in image search results for occupations. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3819–3828.

[Kelly jr, 1956] Kelly jr, J. (1956). A new interpretation of information rate. *the bell system technical journal.*

[Kendall et al., 2017] Kendall, A., Badrinarayanan, V., and Cipolla, R. (2017). Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *British Machine Vision Conference 2017, BMVC 2017.*

[Kiela and Bottou, 2014] Kiela, D. and Bottou, L. (2014). Learning image embeddings using convolutional neural networks for improved multi-modal semantics. *Proceedings of EMNLP*, pages 36–45.

[Kiela and Clark, 2014] Kiela, D. and Clark, S. (2014). A Systematic Study of Semantic Vector Space Model Parameters. *Proceedings of EACL 2014, Workshop on Continuous Vector Space Models and their Compositionality (CVSC).*

[Kiela et al., 2014] Kiela, D., Hill, F., Korhonen, A., and Clark, S. (2014). Improving multi-modal representations using image dispersion: Why less is sometimes more. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2:835–841.

[Kiela et al., 2016] Kiela, D., Verő, A. L., and Clark, S. (2016). Comparing Data Sources and Architectures for Deep Visual Representation Learning in Semantics. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-16).*

[Kilgarriff and Yallop, 2000] Kilgarriff, A. and Yallop, C. (2000). What's in a thesaurus? *LREC*, pages 1371–1379.

[Kiros et al., 2014] Kiros, R., Salakhutdinov, R., and Zemel, R. (2014). Multimodal neural language models. *International Conference on Machine Learning*, pages 595–603.

[Kiros et al., 2015] Kiros, R., Zhu, Y., Salakhutdinov, R., Zemel, R. S., Torralba, A., Urtasun, R., and Fidler, S. (2015). Skip-Thought Vectors. *ArxiV*, 58(786):1–11.

[Kottur et al., 2015] Kottur, S., Vedantam, R., Moura, J. M. F., and Parikh, D. (2015). Visual Word2Vec (vis-w2v): Learning Visually Grounded Word Embeddings Using Abstract Scenes. *arXiv preprint.*

[Kripke, 1972] Kripke, S. A. (1972). Naming and necessity. pages 253–355.

[Krishna et al., 2016] Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., Bernstein, M., and Fei-Fei, L. (2016). Visual genome: Connecting language and vision using crowdsourced dense image annotations.

[Krizhevsky et al., 2012] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Proceedings of NIPS*, pages 1106–1114.

[Kuhnle, 2020] Kuhnle, A. (2020). Evaluating visually grounded language capabilities using microworlds. Technical report, University of Cambridge, Computer Laboratory.

[Kuhnle and Copestake, 2017] Kuhnle, A. and Copestake, A. (2017). Shapeworld-a new test methodology for multimodal language understanding. *arXiv preprint arXiv:1704.04517*.

[Kuzmenko and Herbelot, 2019] Kuzmenko, E. and Herbelot, A. (2019). Distributional semantics in the real world: building word vector representations from a truth-theoretic model. *Proceedings of the 13th International Conference on Computational Semantics-Short Papers*, pages 16–23.

[Lazaridou et al., 2015] Lazaridou, A., Baroni, M., et al. (2015). Combining language and vision with a multimodal skip-gram model. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 153–163.

[Lazaridou et al., 2016] Lazaridou, A., Pham, N. T., and Baroni, M. (2016). Towards Multi-Agent Communication-Based Language Learning.

[LeCun et al., 1989] LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551.

[Lecun et al., 1998] Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

[Lenci, 2008] Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Italian journal of linguistics*, 20(1):1–31.

[Lenci, 2018] Lenci, A. (2018). Distributional models of word meaning. *Annual review of Linguistics*, 4:151–171.

[Levy and Goldberg, 2014a] Levy, O. and Goldberg, Y. (2014a). Dependency-based word embeddings. *ACL (2)*, pages 302–308.

[Levy and Goldberg, 2014b] Levy, O. and Goldberg, Y. (2014b). Neural word embedding as implicit matrix factorization. *Advances in neural information processing systems*, pages 2177–2185.

[Levy et al., 2015] Levy, O., Goldberg, Y., and Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

[Lin et al., 2013] Lin, M., Chen, Q., and Yan, S. (2013). Network in network. *CoRR*, abs/1312.4400.

[Lin et al., 2014] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. *European conference on computer vision*, pages 740–755.

[Lin and Parikh, 2015] Lin, X. and Parikh, D. (2015). Don't just listen, use your imagination: Leveraging visual common sense for non-visual tasks. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 07-12-June:2984–2993.

[Lu et al., 2019] Lu, J., Batra, D., Parikh, D., and Lee, S. (2019). Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in Neural Information Processing Systems*, 32.

[Lucey et al., 2017] Lucey, J. A., Otter, D., and Horne, D. S. (2017). A 100-year review: Progress on the chemistry of milk and its components. *Journal of Dairy Science*, 100(12):9916–9932.

[Maaten and Hinton, 2008] Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

[MacKay, 2003] MacKay, D. J. (2003). *Information theory, inference and learning algorithms.*

[MacQueen et al., 1967] MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 1(14):281–297.

[Majumdar et al., 2020] Majumdar, A., Shrivastava, A., Lee, S., Anderson, P., Parikh, D., and Batra, D. (2020). Improving vision-and-language navigation with image-text pairs from the web. *European Conference on Computer Vision*, pages 259–274.

[Manning and Schutze, 1999] Manning, C. and Schutze, H. (1999). *Foundations of statistical natural language processing.*

[Marconi, 1997] Marconi, D. (1997). *Lexical competence.*

[Margolis and Laurence, 2021] Margolis, E. and Laurence, S. (2021). Concepts. *The Stanford Encyclopedia of Philosophy.*

[Mervis and Rosch, 1981] Mervis, C. B. and Rosch, E. (1981). Categorization of natural objects. *Annual review of psychology*, 32(1):89–115.

[Mikolov et al., 2013a] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings.*

[Mikolov et al., 2018] Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., and Joulin, A. (2018). Advances in pre-training distributed word representations. *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018).*

[Mikolov et al., 2013b] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, pages 3111–3119.

[Miller, 1995] Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

[Minnema and Herbelot, 2019] Minnema, G. and Herbelot, A. (2019). From brain space to distributional space: the perilous journeys of fmri decoding. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 155–161.

[Mitchell and Lapata, 2010] Mitchell, J. and Lapata, M. (2010). Composition in distributional models of semantics. *Cognitive science*, 34(8):1388–429.

[Mitchell et al., 2008] Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., and Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *science*, 320(5880):1191–1195.

[Nair and Hinton, 2010] Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. *Proceedings of ICML*, pages 807–814.

[Navigli, 2009] Navigli, R. (2009). Word sense disambiguation: A survey. *ACM computing surveys (CSUR)*, 41(2):1–69.

[Nelson et al., 2004] Nelson, D. L., McEvoy, C. L., and Schreiber, T. A. (2004). The university of south florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, 36(3):402–407.

[Pedersen, 1996] Pedersen, T. (1996). Fishing for exactness. *arXiv preprint cmp-lg/9608010*.

[Pennington et al., 2014] Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

[Pereira et al., 2018] Pereira, F., Lou, B., Pritchett, B., Ritter, S., Gershman, S. J., Kanwisher, N., Botvinick, M., and Fedorenko, E. (2018). Toward a universal decoder of linguistic meaning from brain activation. *Nature communications*, 9(1):1–13.

[Peters et al., 2018] Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237.

[Ponce et al., 2006] Ponce, J., Berg, T. L., Everingham, M., Forsyth, D. A., Hebert, M., Lazebnik, S., Marszalek, M., Schmid, C., Russell, B. C., Torralba, A., et al. (2006). Dataset issues in object recognition. pages 29–48.

[Putnam, 1970] Putnam, H. (1970). Is semantics possible? *Metaphilosophy*, 1(3):187–201.

[Radford et al., 2018] Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper.pdf*.

[Radovanović et al., 2010] Radovanović, M., Nanopoulos, A., and Ivanović, M. (2010). On the existence of obstinate results in vector space models. *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 186–193.

[Recanati, 2004] Recanati, F. (2004). *Literal meaning.*

[Rocktäschel et al., 2016] Rocktäschel, T., Grefenstette, E., Hermann, K. M., Kočiský, T., and Blunsom, P. (2016). Reasoning about Entailment with Neural Attention. *ICLR.*

[Roy, 2005] Roy, D. (2005). Grounding words in perception and action: Computational insights.

[Sahlgren and Lenci, 2016] Sahlgren, M. and Lenci, A. (2016). The effects of data size and frequency range on distributional semantic models. *EMNLP 2016.*

[Scarselli et al., 2008] Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., and Monfardini, G. (2008). The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80.

[Schuler, 2005] Schuler, K. K. (2005). Verbnet: A broad-coverage, comprehensive verb lexicon.

[Schütze et al., 2008] Schütze, H., Manning, C. D., and Raghavan, P. (2008). Introduction to information retrieval. *Proceedings of the international communication of association for computing machinery conference*, 4.

[Searle, 1985] Searle, J. R. (1985). *Expression and meaning: Studies in the theory of speech acts.*

[Shannon, 2001] Shannon, C. E. (2001). A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review*, 5(1):3–55.

[Sharma et al., 2015] Sharma, S., Kiros, R., and Salakhutdinov, R. (2015). Action Recognition using Visual Attention. *arXiv preprint*, pages 1–11.

[Silberer and Lapata, 2014] Silberer, C. and Lapata, M. (2014). Learning Grounded Meaning Representations with Autoencoders. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, June 23-25:721–732.

[Simonyan and Zisserman, 2014] Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations (ICLR), 2015.*

[Sivic and Zisserman, 2003] Sivic, J. and Zisserman, A. (2003). Video Google: a text retrieval approach to object matching in videos. *IEEE International Conference on Computer Vision*, (Iccv):1470–1477.

[Socher et al., 2014] Socher, R., Karpathy, A., Le, Q. V., Manning, C. D., and Ng, A. Y. (2014). Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218.

[Spärck Jones, 1967] Spärck Jones, K. (1967). A small semantic classification experiment using cooccurrence data. *Report ML*, 196.

[Srivastava and Salakhutdinov, 2012] Srivastava, N. and Salakhutdinov, R. R. (2012). Multimodal learning with deep boltzmann machines. *Advances in neural information processing systems*, pages 2222–2230.

[Su et al., 2019] Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., and Dai, J. (2019). Vl-bert: Pre-training of generic visual-linguistic representations. *International Conference on Learning Representations*.

[Sudre et al., 2012] Sudre, G., Pomerleau, D., Palatucci, M., Wehbe, L., Fyshe, A., Salmelin, R., and Mitchell, T. (2012). Tracking neural coding of perceptual and semantic features of concrete nouns. *NeuroImage*, 62(1):451–463.

[Szabó, 2014] Szabó, Z. (2014). Information theoretical estimators toolbox. *The Journal of Machine Learning Research*, 15(1):283–287.

[Szegedy et al., 2015] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9.

[Tettamanti et al., 2005] Tettamanti, M., Buccino, G., Saccuman, M. C., Gallese, V., Danna, M., Scifo, P., Fazio, F., Rizzolatti, G., Cappa, S. F., and Perani, D. (2005). Listening to action-related sentences activates fronto-parietal motor circuits. *Journal of cognitive neuroscience*, 17(2):273–281.

[Torralba and Efros, 2011] Torralba, A. and Efros, A. A. (2011). Unbiased look at dataset bias. *CVPR 2011*, pages 1521–1528.

[Tsai et al., 2019] Tsai, Y.-H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L.-P., and Salakhutdinov, R. (2019). Multimodal transformer for unaligned multimodal language sequences. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

[Turney, 2010] Turney, P. D. (2010). From Frequency to Meaning : Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

[Vendrov et al., 2015] Vendrov, I., Kiros, R., Fidler, S., and Urtasun, R. (2015). Order-Embeddings of Images and Language. *arXiv preprint*, (2005):1–13.

[Vert et al., 2004] Vert, J.-P., Tsuda, K., and Schölkopf, B. (2004). A primer on kernel methods. *Kernel methods in computational biology*, 47:35–70.

[Voita and Titov, 2020] Voita, E. and Titov, I. (2020). Information-theoretic probing with minimum description length. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196.

[von Ahn and Dabbish, 2004] von Ahn, L. and Dabbish, L. (2004). Labeling images with a computer game. *CHI*, pages 319–326.

[Von Ahn and Dabbish, 2004] Von Ahn, L. and Dabbish, L. (2004). Labeling images with a computer game. *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 319–326.

[Wang et al., 2018a] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2018a). Glue: A multi-task benchmark and analysis platform for natural language understanding. *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355.

[Wang et al., 2018b] Wang, J., Madhyastha, P. S., and Specia, L. (2018b). Object counts! bringing explicit detections back into image captioning. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2180–2193.

[Wang et al., 2005] Wang, Q., Kulkarni, S. R., and Verdú, S. (2005). Divergence estimation of continuous distributions based on data-dependent partitions. *IEEE Transactions on Information Theory*, 51(9):3064–3074.

[Wang et al., 2009] Wang, Q., Kulkarni, S. R., and Verdú, S. (2009). Divergence estimation for multidimensional densities via $k$-nearest-neighbor distances. *IEEE Transactions on Information Theory*, 55(5):2392–2405.

[Wang and Jiang, 2015] Wang, S. and Jiang, J. (2015). Learning Natural Language Inference with LSTM. *Naacl*.

[Wattenberg et al., 2016] Wattenberg, M., Viégas, F., and Johnson, I. (2016). How to use t-sne effectively. *Distill*.

[Wittgenstein, 1953] Wittgenstein, L. (1953). *Philosophical investigations*.

[Wold et al., 1987] Wold, S., Esbensen, K., and Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52.

[Xu et al., 2016] Xu, H., Murphy, B., and Fyshe, A. (2016). Brainbench: A brain-image test suite for distributional semantic models. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2017–2021.

[Xu et al., 2020] Xu, P., Chang, X., Guo, L., Huang, P.-Y., Chen, X., and Hauptmann, A. G. (2020). A survey of scene graph: Generation and application. *IEEE Trans. Neural Netw. Learn. Syst. 2020*.

[Yang et al., 2019] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

[Yeung, 1991] Yeung, R. W. (1991). A new outlook on shannon's information measures. *IEEE transactions on information theory*, 37(3):466–474.

[Yogatama et al., 2019] Yogatama, D., d'Autume, C. d. M., Connor, J., Kocisky, T., Chrzanowski, M., Kong, L., Lazaridou, A., Ling, W., Yu, L., Dyer, C., et al. (2019). Learning and evaluating general linguistic intelligence. *arXiv preprint arXiv:1901.11373*.

[Zhang and Bowman, 2018] Zhang, K. and Bowman, S. (2018). Language modeling teaches you more than translation does: Lessons learned through auxiliary syntactic task analysis. *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 359–361.

[Zhang et al., 2016] Zhang, P., Goyal, Y., Summers-Stay, D., Batra, D., and Parikh, D. (2016). Yin and yang: Balancing and answering binary visual questions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5014–5022.

[Zhang et al., 2018] Zhang, Q., Wang, W., and Zhu, S.-C. (2018). Examining cnn representations with respect to dataset bias. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

# Appendix A

# Cross-validated Semantic Relatedness and Similarity

| **Embedding** | **Spearman** | **P-value** | **Coverage** |
|---|---|---|---|
| *wikinews* | 0.797 (0.004) | <1‰(<1‰) | 2000 |
| *wikinews_sub* | 0.805 (0.005) | <1‰(<1‰) | 2000 |
| *crawl* | 0.843 (0.001) | <1‰(<1‰) | 2000 |
| *w2v13* | 0.684 (0.007) | <1‰(<1‰) | 2000 |
| *Google AlexNet* | 0.506 (0.009) | <1‰(<1‰) | 2000 |
| *VG SceneGraph* | 0.427 (0.006) | <1‰(<1‰) | 1716 |
| *Google VGG* | 0.516 (0.005) | <1‰(<1‰) | 2000 |
| *VG-internal* | 0.377 (0.008) | <1‰(<1‰) | 1856 |
| *VG-whole* | 0.415 (0.006) | <1‰(<1‰) | 1856 |
| *Google ResNet-152* | 0.469 (0.003) | <1‰(<1‰) | 2000 |
| *wikinews+Google AlexNet* | 0.499 (0.003) | <1‰(<1‰) | 2000 |
| *wikinews+VG SceneGraph* | 0.568 (0.013) | <1‰(<1‰) | 2000 |
| *wikinews+Google VGG* | 0.512 (0.005) | <1‰(<1‰) | 2000 |
| *wikinews+VG-internal* | 0.367 (0.008) | <1‰(<1‰) | 2000 |
| *wikinews+VG-whole* | 0.402 (0.007) | <1‰(<1‰) | 2000 |
| *wikinews+Google ResNet-152* | 0.479 (0.008) | <1‰(<1‰) | 2000 |
| *wikinews_sub+Google AlexNet* | 0.506 (0.011) | <1‰(<1‰) | 2000 |
| *wikinews_sub+VG SceneGraph* | 0.380 (0.010) | <1‰(<1‰) | 2000 |
| *wikinews_sub+Google VGG* | 0.514 (0.012) | <1‰(<1‰) | 2000 |
| *wikinews_sub+VG-internal* | 0.364 (0.009) | <1‰(<1‰) | 2000 |
| *wikinews_sub+VG-whole* | 0.387 (0.013) | <1‰(<1‰) | 2000 |
| *wikinews_sub+Google ResNet-152* | 0.463 (0.004) | <1‰(<1‰) | 2000 |
| *crawl+Google AlexNet* | 0.501 (0.004) | <1‰(<1‰) | 2000 |
| *crawl+VG SceneGraph* | 0.778 (0.006) | <1‰(<1‰) | 2000 |
| *crawl+Google VGG* | 0.516 (0.006) | <1‰(<1‰) | 2000 |
| *crawl+VG-internal* | 0.357 (0.012) | <1‰(<1‰) | 2000 |
| *crawl+VG-whole* | 0.398 (0.008) | <1‰(<1‰) | 2000 |
| *crawl+Google ResNet-152* | 0.514 (0.007) | <1‰(<1‰) | 2000 |
| *w2v13+Google AlexNet* | 0.501 (0.012) | <1‰(<1‰) | 2000 |
| *w2v13+VG SceneGraph* | 0.645 (0.008) | <1‰(<1‰) | 2000 |
| *w2v13+Google VGG* | 0.518 (0.010) | <1‰(<1‰) | 2000 |
| *w2v13+VG-internal* | 0.372 (0.005) | <1‰(<1‰) | 2000 |
| *w2v13+VG-whole* | 0.403 (0.004) | <1‰(<1‰) | 2000 |
| *w2v13+Google ResNet-152* | 0.486 (0.002) | <1‰(<1‰) | 2000 |

Table A.1: Cross-validated Spearman correlations on the MEN dataset. Spearman and P-value columns report <mean (STD)> of three samples after leaving out the third of the evaluation pairs. Multi-modal embeddings are created using the Padding technique. The table sections contain linguistic, visual and multi-modal embeddings in this order.

| **Embedding** | **Spearman** | **P-value** | **Coverage** |
|---|---|---|---|
| *wikinews* | 0.463 (0.009) | <1‰(<1‰) | 666 |
| *wikinews_sub* | 0.412 (0.025) | <1‰(<1‰) | 666 |
| *crawl* | 0.506 (0.019) | <1‰(<1‰) | 666 |
| *w2v13* | 0.316 (0.020) | <1‰(<1‰) | 666 |
| *Google AlexNet* | 0.348 (0.025) | <1‰(<1‰) | 666 |
| *VG SceneGraph* | 0.274 (0.019) | <1‰(<1‰) | 395 |
| *Google VGG* | 0.363 (0.017) | <1‰(<1‰) | 666 |
| *VG-internal* | 0.311 (0.059) | 0.023 (0.027) | 68 |
| *VG-whole* | 0.169 (0.024) | 0.178 (0.068) | 68 |
| *Google ResNet-152* | 0.354 (0.007) | <1‰(<1‰) | 666 |
| *wikinews+Google AlexNet* | 0.332 (0.032) | <1‰(<1‰) | 666 |
| *wikinews+VG SceneGraph* | 0.348 (0.018) | <1‰(<1‰) | 666 |
| *wikinews+Google VGG* | 0.332 (0.014) | <1‰(<1‰) | 666 |
| *wikinews+VG-internal* | 0.300 (0.002) | <1‰(<1‰) | 666 |
| *wikinews+VG-whole* | 0.326 (0.017) | <1‰(<1‰) | 666 |
| *wikinews+Google ResNet-152* | 0.350 (0.028) | <1‰(<1‰) | 666 |
| *wikinews_sub+Google AlexNet* | 0.329 (0.022) | <1‰(<1‰) | 666 |
| *wikinews_sub+VG SceneGraph* | 0.187 (0.027) | <1‰(<1‰) | 666 |
| *wikinews_sub+Google VGG* | 0.353 (0.011) | <1‰(<1‰) | 666 |
| *wikinews_sub+VG-internal* | 0.299 (0.013) | <1‰(<1‰) | 666 |
| *wikinews_sub+VG-whole* | 0.304 (0.015) | <1‰(<1‰) | 666 |
| *wikinews_sub+Google ResNet-152* | 0.348 (0.011) | <1‰(<1‰) | 666 |
| *crawl+Google AlexNet* | 0.349 (0.025) | <1‰(<1‰) | 666 |
| *crawl+VG SceneGraph* | 0.434 (0.017) | <1‰(<1‰) | 666 |
| *crawl+Google VGG* | 0.346 (0.017) | <1‰(<1‰) | 666 |
| *crawl+VG-internal* | 0.310 (0.038) | <1‰(<1‰) | 666 |
| *crawl+VG-whole* | 0.321 (0.007) | <1‰(<1‰) | 666 |
| *crawl+Google ResNet-152* | 0.364 (0.009) | <1‰(<1‰) | 666 |
| *w2v13+Google AlexNet* | 0.345 (0.024) | <1‰(<1‰) | 666 |
| *w2v13+VG SceneGraph* | 0.312 (0.007) | <1‰(<1‰) | 666 |
| *w2v13+Google VGG* | 0.362 (0.017) | <1‰(<1‰) | 666 |
| *w2v13+VG-internal* | 0.209 (0.017) | <1‰(<1‰) | 666 |
| *w2v13+VG-whole* | 0.225 (0.007) | <1‰(<1‰) | 666 |
| *w2v13+Google ResNet-152* | 0.352 (0.020) | <1‰(<1‰) | 666 |

Table A.2: Cross-validated Spearman correlations on the SimLex dataset. Spearman and P-value columns report <mean (STD)> of three samples after leaving out the third of the evaluation pairs. Multi-modal embeddings are created using the Padding technique. The table sections contain linguistic, visual and multi-modal embeddings in this order.

| **Embedding** | **Spearman** | **P-value** | **Coverage** |
|---|---|---|---:|
| *wikinews* | 0.792 (0.002) | <1‰(<1‰) | 2000 |
| *wikinews_sub* | 0.804 (0.001) | <1‰(<1‰) | 2000 |
| *crawl* | 0.845 (0.001) | <1‰(<1‰) | 2000 |
| *w2v13* | 0.684 (0.003) | <1‰(<1‰) | 2000 |
| *Google AlexNet* | 0.509 (0.005) | <1‰(<1‰) | 2000 |
| *VG SceneGraph* | 0.413 (0.004) | <1‰(<1‰) | 1716 |
| *Google VGG* | 0.508 (0.008) | <1‰(<1‰) | 2000 |
| *VG-internal* | 0.374 (0.015) | <1‰(<1‰) | 1856 |
| *VG-whole* | 0.412 (0.002) | <1‰(<1‰) | 1856 |
| *Google ResNet-152* | 0.464 (0.007) | <1‰(<1‰) | 2000 |
| *wikinews+Google AlexNet* | 0.497 (0.004) | <1‰(<1‰) | 2000 |
| *wikinews+VG SceneGraph* | 0.654 (0.006) | <1‰(<1‰) | 1716 |
| *wikinews+Google VGG* | 0.504 (0.011) | <1‰(<1‰) | 2000 |
| *wikinews+VG-internal* | 0.374 (0.003) | <1‰(<1‰) | 1856 |
| *wikinews+VG-whole* | 0.415 (0.006) | <1‰(<1‰) | 1856 |
| *wikinews+Google ResNet-152* | 0.476 (0.004) | <1‰(<1‰) | 2000 |
| *wikinews_sub+Google AlexNet* | 0.501 (0.008) | <1‰(<1‰) | 2000 |
| *wikinews_sub+VG SceneGraph* | 0.452 (0.021) | <1‰(<1‰) | 1716 |
| *wikinews_sub+Google VGG* | 0.503 (0.002) | <1‰(<1‰) | 2000 |
| *wikinews_sub+VG-internal* | 0.370 (0.005) | <1‰(<1‰) | 1856 |
| *wikinews_sub+VG-whole* | 0.415 (0.005) | <1‰(<1‰) | 1856 |
| *wikinews_sub+Google ResNet-152* | 0.475 (0.005) | <1‰(<1‰) | 2000 |
| *crawl+Google AlexNet* | 0.502 (0.009) | <1‰(<1‰) | 2000 |
| *crawl+VG SceneGraph* | 0.813 (0.001) | <1‰(<1‰) | 1716 |
| *crawl+Google VGG* | 0.512 (0.008) | <1‰(<1‰) | 2000 |
| *crawl+VG-internal* | 0.392 (0.005) | <1‰(<1‰) | 1856 |
| *crawl+VG-whole* | 0.427 (0.006) | <1‰(<1‰) | 1856 |
| *crawl+Google ResNet-152* | 0.514 (0.003) | <1‰(<1‰) | 2000 |
| *w2v13+Google AlexNet* | 0.502 (0.004) | <1‰(<1‰) | 2000 |
| *w2v13+VG SceneGraph* | 0.696 (0.003) | <1‰(<1‰) | 1716 |
| *w2v13+Google VGG* | 0.528 (0.005) | <1‰(<1‰) | 2000 |
| *w2v13+VG-internal* | 0.369 (0.011) | <1‰(<1‰) | 1856 |
| *w2v13+VG-whole* | 0.423 (0.010) | <1‰(<1‰) | 1856 |
| *w2v13+Google ResNet-152* | 0.484 (0.010) | <1‰(<1‰) | 2000 |

Table A.3: Cross-validated Spearman correlations on the MEN dataset. Spearman and P-value columns report <mean (STD)> of three samples after leaving out the third of the evaluation pairs. Multi-modal embeddings are created using the Intersection technique. The table sections contain linguistic, visual and multi-modal embeddings in this order.

| Embedding | Spearman | P-value | Coverage |
|---|---|---|---|
| *wikinews* | 0.457 (0.006) | <1‰(<1‰) | 666 |
| *wikinews_sub* | 0.443 (0.015) | <1‰(<1‰) | 666 |
| *crawl* | 0.493 (0.013) | <1‰(<1‰) | 666 |
| *w2v13* | 0.300 (0.010) | <1‰(<1‰) | 666 |
| *Google AlexNet* | 0.348 (0.004) | <1‰(<1‰) | 666 |
| *VG SceneGraph* | 0.249 (0.023) | <1‰(<1‰) | 395 |
| *Google VGG* | 0.344 (0.008) | <1‰(<1‰) | 666 |
| *VG-internal* | 0.289 (0.034) | 0.022 (0.015) | 68 |
| *VG-whole* | 0.118 (0.032) | 0.354 (0.135) | 68 |
| *Google ResNet-152* | 0.351 (0.022) | <1‰(<1‰) | 666 |
| *wikinews+Google AlexNet* | 0.331 (0.021) | <1‰(<1‰) | 666 |
| *wikinews+VG SceneGraph* | 0.362 (0.017) | <1‰(<1‰) | 395 |
| *wikinews+Google VGG* | 0.318 (0.019) | <1‰(<1‰) | 666 |
| *wikinews+VG-internal* | 0.289 (0.043) | 0.024 (0.021) | 68 |
| *wikinews+VG-whole* | 0.269 (0.017) | 0.028 (0.009) | 68 |
| *wikinews+Google ResNet-152* | 0.370 (0.017) | <1‰(<1‰) | 666 |
| *wikinews_sub+Google AlexNet* | 0.356 (0.015) | <1‰(<1‰) | 666 |
| *wikinews_sub+VG SceneGraph* | 0.304 (0.022) | <1‰(<1‰) | 395 |
| *wikinews_sub+Google VGG* | 0.336 (0.021) | <1‰(<1‰) | 666 |
| *wikinews_sub+VG-internal* | 0.270 (0.058) | 0.046 (0.048) | 68 |
| *wikinews_sub+VG-whole* | 0.090 (0.119) | 0.528 (0.350) | 68 |
| *wikinews_sub+Google ResNet-152* | 0.348 (0.005) | <1‰(<1‰) | 666 |
| *crawl+Google AlexNet* | 0.358 (0.014) | <1‰(<1‰) | 666 |
| *crawl+VG SceneGraph* | 0.428 (0.027) | <1‰(<1‰) | 395 |
| *crawl+Google VGG* | 0.332 (0.008) | <1‰(<1‰) | 666 |
| *crawl+VG-internal* | 0.305 (0.024) | 0.013 (0.006) | 68 |
| *crawl+VG-whole* | 0.160 (0.074) | 0.271 (0.247) | 68 |
| *crawl+Google ResNet-152* | 0.370 (0.026) | <1‰(<1‰) | 666 |
| *w2v13+Google AlexNet* | 0.338 (0.002) | <1‰(<1‰) | 666 |
| *w2v13+VG SceneGraph* | 0.278 (0.008) | <1‰(<1‰) | 395 |
| *w2v13+Google VGG* | 0.337 (0.019) | <1‰(<1‰) | 666 |
| *w2v13+VG-internal* | 0.306 (0.049) | 0.017 (0.011) | 68 |
| *w2v13+VG-whole* | 0.233 (0.058) | 0.086 (0.080) | 68 |
| *w2v13+Google ResNet-152* | 0.367 (0.004) | <1‰(<1‰) | 666 |

Table A.4: Cross-validated Spearman correlations on the SimLex dataset. Spearman and P-value columns report <mean (STD)> of three samples after leaving out the third of the evaluation pairs. Multi-modal embeddings are created using the Intersection technique. The table sections contain linguistic, visual and multi-modal embeddings in this order.

| **Embedding** | **Spearman** | **P-value** | **Coverage** |
|---|---|---|---|
| *wikinews* | 0.798 (0.005) | <1‰(<1‰) | 1654 |
| *wikinews_sub* | 0.806 (0.004) | <1‰(<1‰) | 1654 |
| *crawl* | 0.844 (0.003) | <1‰(<1‰) | 1654 |
| *w2v13* | 0.667 (0.003) | <1‰(<1‰) | 1654 |
| *Google AlexNet* | 0.511 (0.006) | <1‰(<1‰) | 1654 |
| *VG SceneGraph* | 0.431 (0.015) | <1‰(<1‰) | 1654 |
| *Google VGG* | 0.524 (0.007) | <1‰(<1‰) | 1654 |
| *VG-internal* | 0.381 (0.008) | <1‰(<1‰) | 1654 |
| *VG-whole* | 0.405 (0.009) | <1‰(<1‰) | 1654 |
| *Google ResNet-152* | 0.472 (0.014) | <1‰(<1‰) | 1654 |
| *wikinews+Google AlexNet* | 0.518 (0.004) | <1‰(<1‰) | 1654 |
| *wikinews+VG SceneGraph* | 0.654 (0.006) | <1‰(<1‰) | 1654 |
| *wikinews+Google VGG* | 0.516 (0.003) | <1‰(<1‰) | 1654 |
| *wikinews+VG-internal* | 0.376 (0.002) | <1‰(<1‰) | 1654 |
| *wikinews+VG-whole* | 0.412 (0.008) | <1‰(<1‰) | 1654 |
| *wikinews+Google ResNet-152* | 0.476 (0.014) | <1‰(<1‰) | 1654 |
| *wikinews_sub+Google AlexNet* | 0.516 (0.007) | <1‰(<1‰) | 1654 |
| *wikinews_sub+VG SceneGraph* | 0.452 (0.008) | <1‰(<1‰) | 1654 |
| *wikinews_sub+Google VGG* | 0.515 (0.004) | <1‰(<1‰) | 1654 |
| *wikinews_sub+VG-internal* | 0.364 (0.002) | <1‰(<1‰) | 1654 |
| *wikinews_sub+VG-whole* | 0.406 (0.017) | <1‰(<1‰) | 1654 |
| *wikinews_sub+Google ResNet-152* | 0.483 (0.012) | <1‰(<1‰) | 1654 |
| *crawl+Google AlexNet* | 0.514 (0.015) | <1‰(<1‰) | 1654 |
| *crawl+VG SceneGraph* | 0.813 (0.001) | <1‰(<1‰) | 1654 |
| *crawl+Google VGG* | 0.524 (0.008) | <1‰(<1‰) | 1654 |
| *crawl+VG-internal* | 0.393 (0.007) | <1‰(<1‰) | 1654 |
| *crawl+VG-whole* | 0.423 (0.013) | <1‰(<1‰) | 1654 |
| *crawl+Google ResNet-152* | 0.512 (0.005) | <1‰(<1‰) | 1654 |
| *w2v13+Google AlexNet* | 0.507 (0.007) | <1‰(<1‰) | 1654 |
| *w2v13+VG SceneGraph* | 0.695 (0.004) | <1‰(<1‰) | 1654 |
| *w2v13+Google VGG* | 0.521 (0.008) | <1‰(<1‰) | 1654 |
| *w2v13+VG-internal* | 0.378 (0.005) | <1‰(<1‰) | 1654 |
| *w2v13+VG-whole* | 0.405 (0.002) | <1‰(<1‰) | 1654 |
| *w2v13+Google ResNet-152* | 0.487 (0.006) | <1‰(<1‰) | 1654 |

Table A.5: Cross-validated Spearman correlations on the common subset of the MEN dataset. Spearman and P-value columns report <mean (STD)> of three samples after leaving out the third of the evaluation pairs. Multi-modal embeddings are created using the Intersection technique. The table sections contain linguistic, visual and multi-modal embeddings in this order.

| **Embedding** | **Spearman** | **P-value** | **Coverage** |
|---|---|---|---|
| *wikinews* | 0.299 (0.064) | 0.029 (0.030) | 68 |
| *wikinews_sub* | 0.233 (0.074) | 0.095 (0.064) | 68 |
| *crawl* | 0.361 (0.055) | 0.005 (0.003) | 68 |
| *w2v13* | 0.101 (0.033) | 0.428 (0.145) | 68 |
| *Google AlexNet* | 0.536 (0.042) | <1‰(<1‰) | 68 |
| *VG SceneGraph* | 0.257 (0.038) | 0.044 (0.032) | 68 |
| *Google VGG* | 0.464 (0.031) | <1‰(<1‰) | 68 |
| *VG-internal* | 0.295 (0.030) | 0.018 (0.014) | 68 |
| *VG-whole* | 0.213 (0.049) | 0.108 (0.087) | 68 |
| *Google ResNet-152* | 0.527 (0.034) | <1‰(<1‰) | 68 |
| *wikinews+Google AlexNet* | 0.584 (0.025) | <1‰(<1‰) | 68 |
| *wikinews+VG SceneGraph* | 0.353 (0.070) | 0.008 (0.006) | 68 |
| *wikinews+Google VGG* | 0.547 (0.024) | <1‰(<1‰) | 68 |
| *wikinews+VG-internal* | 0.326 (0.022) | 0.008 (0.003) | 68 |
| *wikinews+VG-whole* | 0.128 (0.074) | 0.377 (0.305) | 68 |
| *wikinews+Google ResNet-152* | 0.456 (0.023) | <1‰(<1‰) | 68 |
| *wikinews_sub+Google AlexNet* | 0.605 (0.027) | <1‰(<1‰) | 68 |
| *wikinews_sub+VG SceneGraph* | 0.317 (0.059) | 0.020 (0.024) | 68 |
| *wikinews_sub+Google VGG* | 0.538 (0.054) | <1‰(<1‰) | 68 |
| *wikinews_sub+VG-internal* | 0.319 (0.062) | 0.019 (0.022) | 68 |
| *wikinews_sub+VG-whole* | 0.165 (0.106) | 0.313 (0.220) | 68 |
| *wikinews_sub+Google ResNet-152* | 0.540 (0.023) | <1‰(<1‰) | 68 |
| *crawl+Google AlexNet* | 0.564 (0.027) | <1‰(<1‰) | 68 |
| *crawl+VG SceneGraph* | 0.339 (0.072) | 0.014 (0.016) | 68 |
| *crawl+Google VGG* | 0.602 (0.023) | <1‰(<1‰) | 68 |
| *crawl+VG-internal* | 0.335 (0.053) | 0.011 (0.012) | 68 |
| *crawl+VG-whole* | 0.178 (0.055) | 0.189 (0.158) | 68 |
| *crawl+Google ResNet-152* | 0.501 (0.018) | <1‰(<1‰) | 68 |
| *w2v13+Google AlexNet* | 0.495 (0.020) | <1‰(<1‰) | 68 |
| *w2v13+VG SceneGraph* | 0.227 (0.084) | 0.136 (0.164) | 68 |
| *w2v13+Google VGG* | 0.485 (0.044) | <1‰(<1‰) | 68 |
| *w2v13+VG-internal* | 0.333 (0.059) | 0.014 (0.018) | 68 |
| *w2v13+VG-whole* | 0.251 (0.049) | 0.055 (0.043) | 68 |
| *w2v13+Google ResNet-152* | 0.498 (0.028) | <1‰(<1‰) | 68 |

Table A.6: Cross-validated Spearman correlations on the common subset of the SimLex dataset. Spearman and P-value columns report <mean (STD)> of three samples after leaving out the third of the evaluation pairs. Multi-modal embeddings are created using the Intersection technique. The table sections contain linguistic, visual and multi-modal embeddings in this order.

# Appendix B

# WordNet Concreteness

Further WordNet concreteness analysis (Section 4.3.4) on the common subset of the datasets for the behavioural tasks, and for *Intersection* type mid-fusion method.

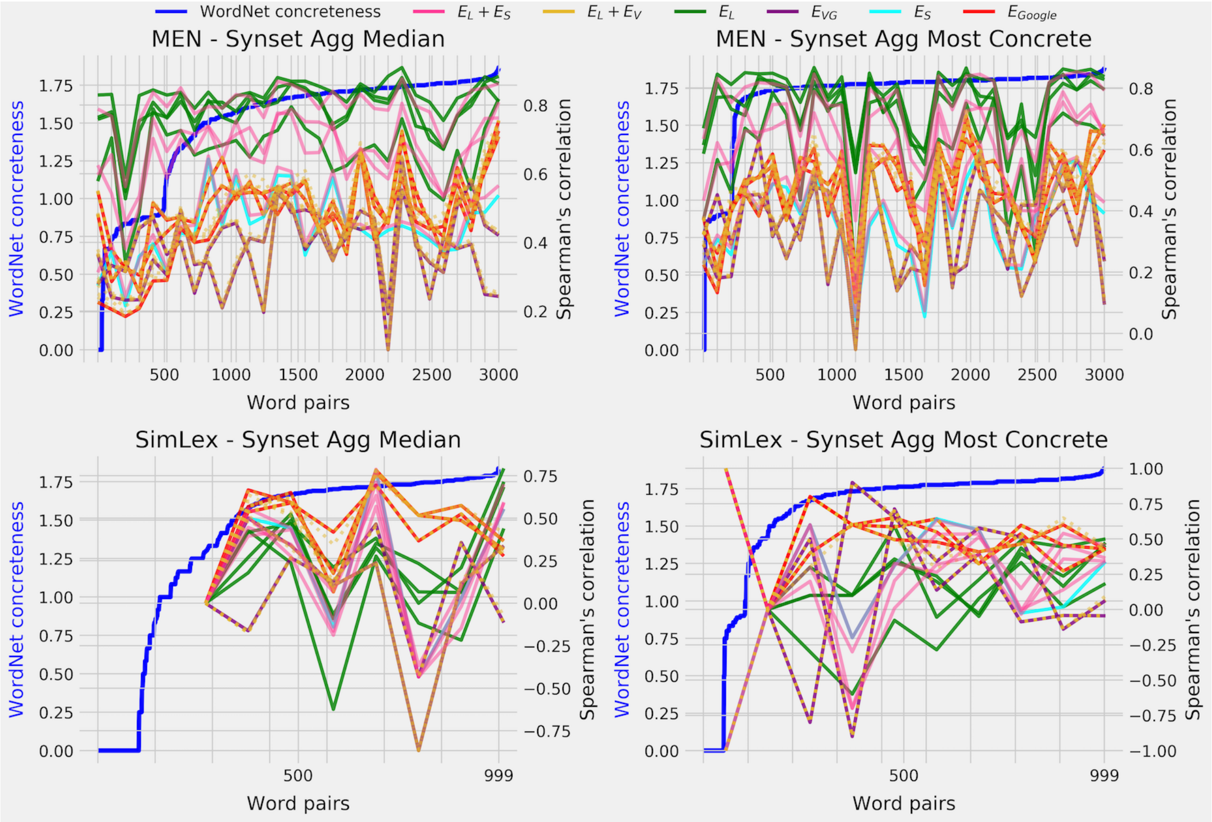


Figure B.1: Scores on the embeddings' common subset of Semantic Similarity dataset splits, ordered by the sum of WordNet concreteness scores of the two words in every word pair. Mid-fusion method: Padding.

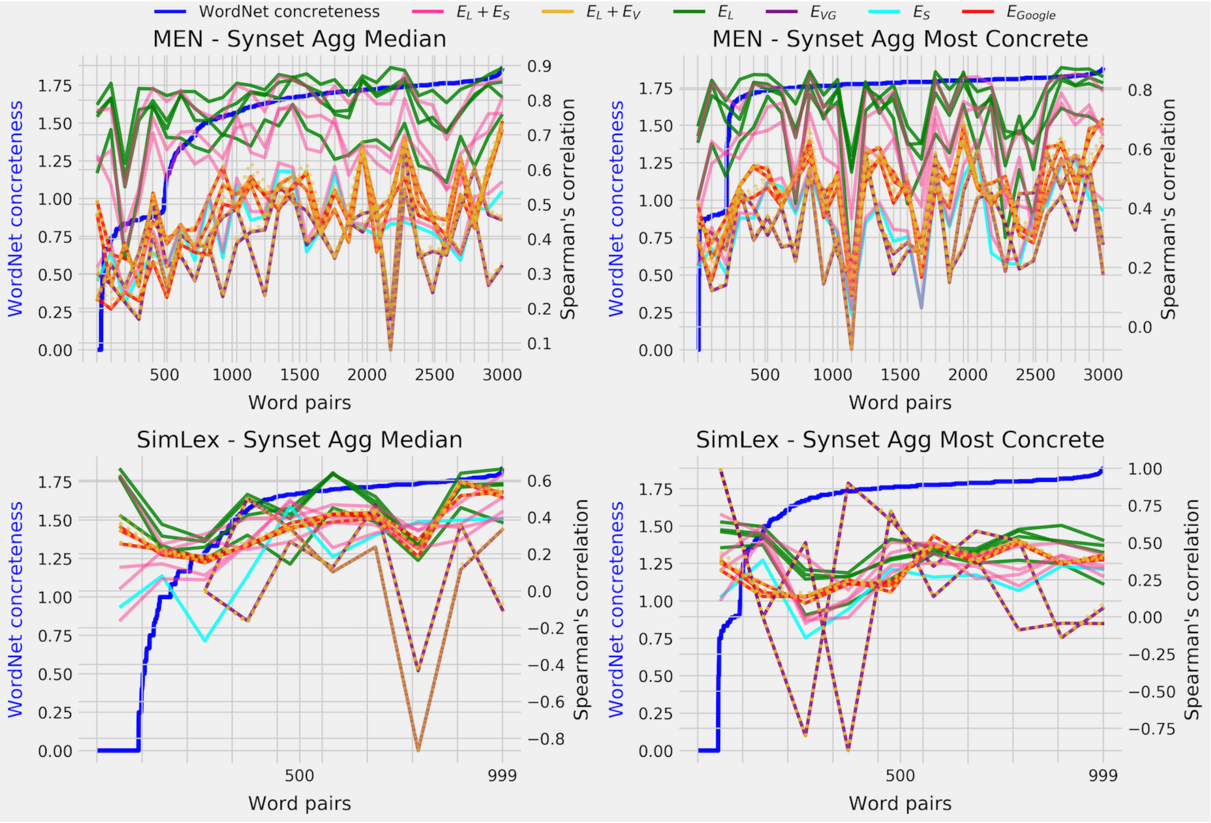


Figure B.2: Scores on the full Semantic Similarity dataset splits, ordered by the sum of WordNet concreteness scores of the two words in every word pair. Mid-fusion method: Intersection.

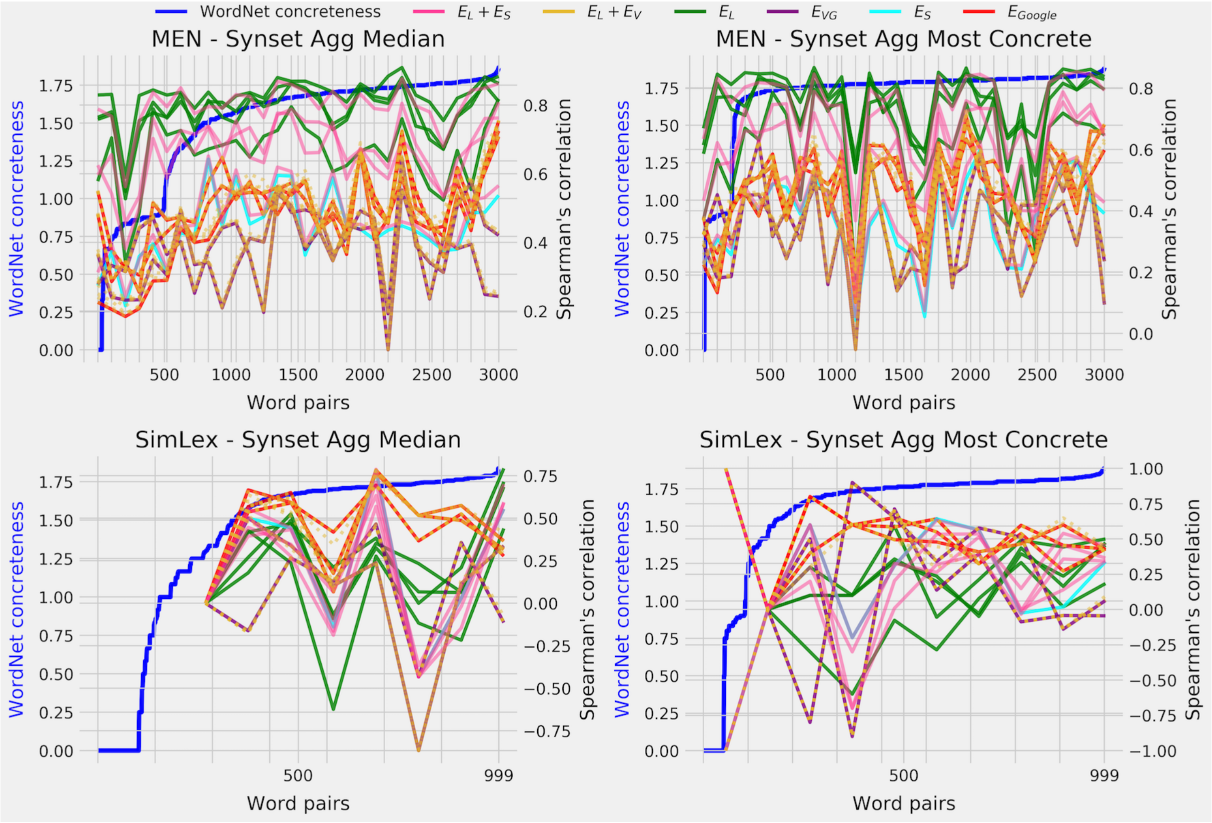


Figure B.3: Scores on the embeddings' common subset of Semantic Similarity dataset splits, ordered by the sum of WordNet concreteness scores of the two words in every word pair. Mid-fusion method: Intersection.

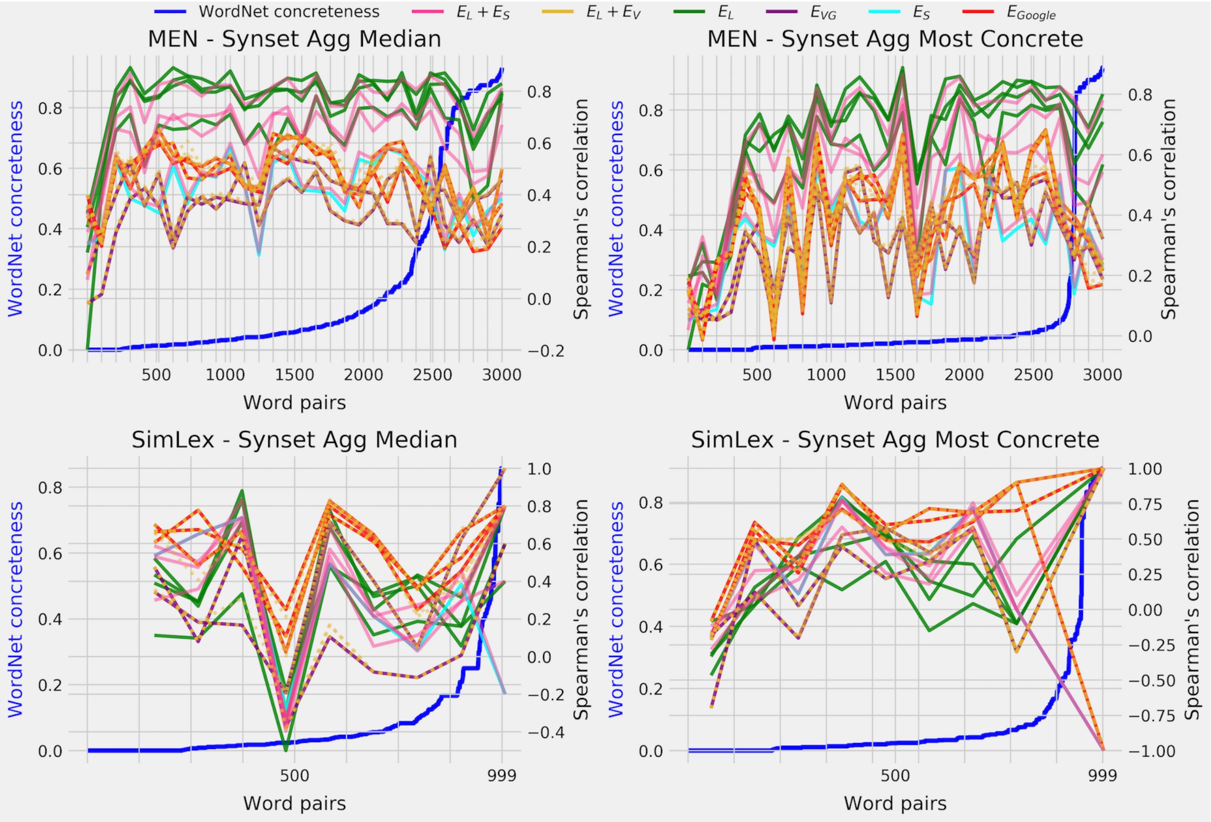


Figure B.4: Scores on the embeddings' common subset of Semantic Similarity dataset splits, ordered by the difference of WordNet concreteness scores of the two words in every word pair. Mid-fusion method: Padding.

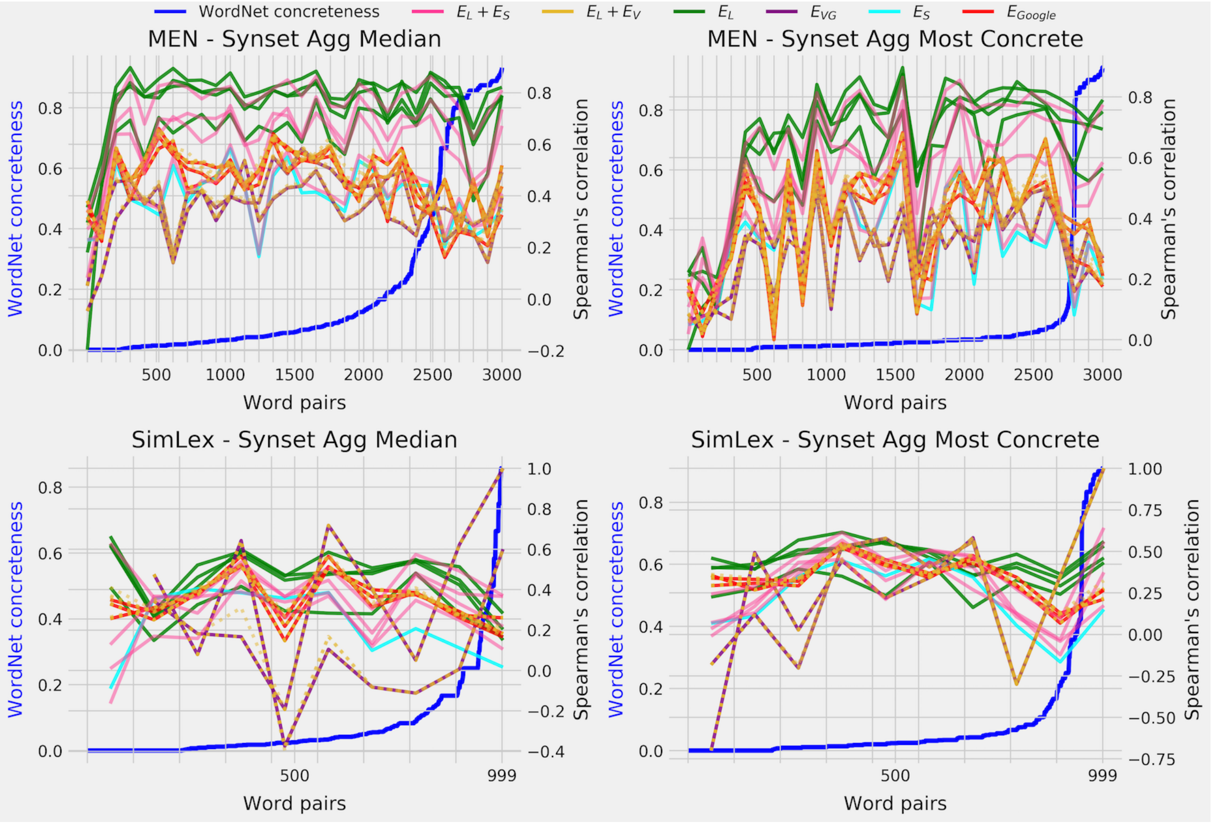


Figure B.5: Scores on the full Semantic Similarity dataset splits, ordered by the difference of WordNet concreteness scores of the two words in every word pair. Mid-fusion method: Intersection.

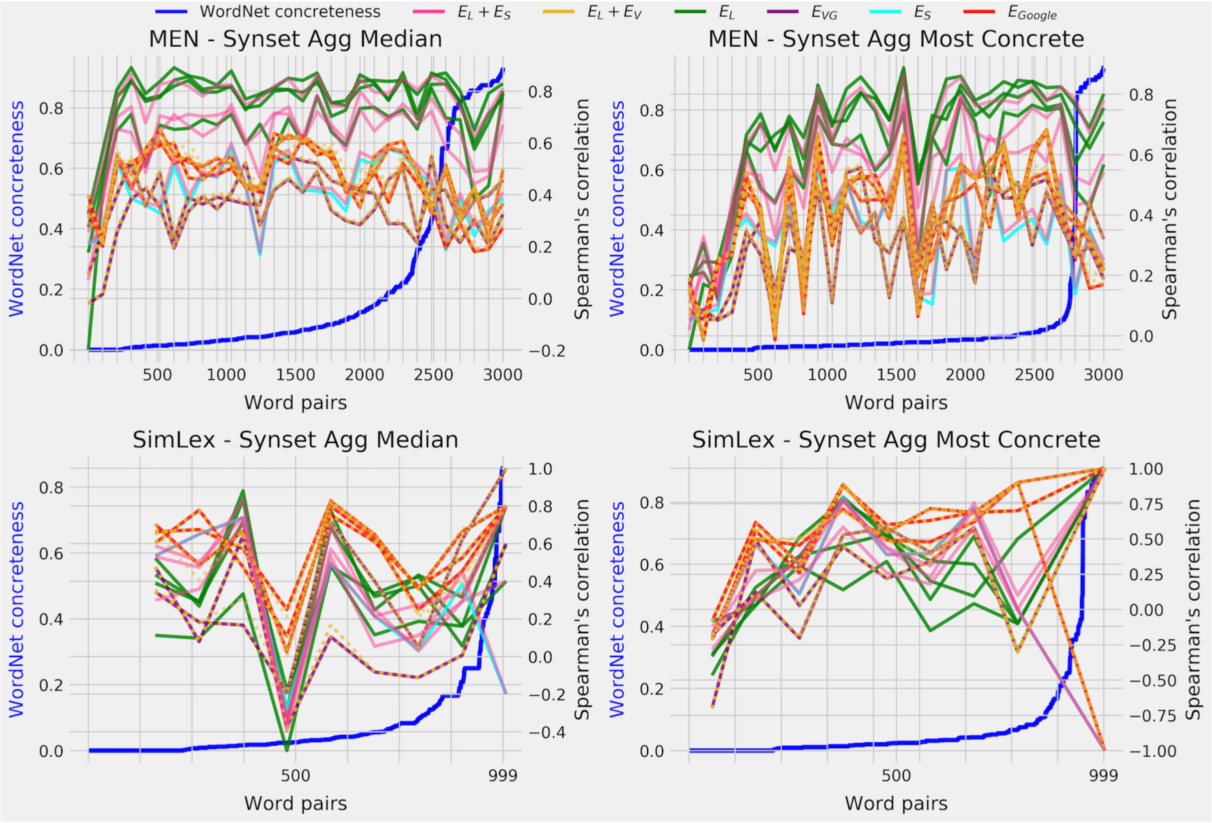


Figure B.6: Scores on the embeddings' common subset of Semantic Similarity dataset splits, ordered by the difference of WordNet concreteness scores of the two words in every word pair. Mid-fusion method: Intersection.

# Appendix C

# EmbEval Toolkit

The code we used to generate the results in this work is openly available[1]. It performs a general evaluation of word embeddings (which we used in Chapters  4, 5 and 6.

The code base loads several embedding models, generates multi-modal embeddings and runs all the evaluations on the semantic similarity and relatedness datasets well as the brain datasets.

The software can also be used to generate the various visualisations and tables of results as well as visualisations of embedding spaces. Details on its usage can be found in the documentation[2].

[1] https://github.com/anitavero/embeval

[2] https://anitavero.github.io/embeval/

# Appendix D

# Cluster Structure

| WordNet label | Own label | Members |
|---|---|---|
| food<br>nutriment<br>foodstuff | food | butter, cheese, bread, chicken, soup, sauce, dessert, beef, salad, meat, cake, steak, tomato, potato, pizza, flour, milk, meal, vinegar, bacon, pie, cooking, sushi, sandwich, breakfast, burger, menu |
| vascular plant<br>plant organ<br>plant part | plants | flower, flowers, tree, blossom, dandelion, foliage, fruit, weed, cactus, lily, bloom, shade, leaf, grass, sunflower, poppy, vine, plant, garden, iris, grow, daisy, oak, bulb, rust, herb, moss, tulip, palm, maple, root, tall, bush, seed, family |
| atmospheric phenomenon<br>physical phenomenon<br>change | weather | rain, snow, fog, weather, mist, drizzle, frost, dew, cold, wet, wind, smoke, sunlight, misty, sunrise, winter, storm, sunset, haze, sunshine, fire, spring, dusk, autumn, heavy, atmosphere, cloud, sunny, burn, flood, desert, sun, hot, ice, tropical |
| food<br>beverage<br>produce | sweets<br>alcohol<br>tobacco<br>"legal drugs" | coffee, lemon, candy, juice, chocolate, sugar, strawberry, honey, tea, beer, bottle, bean, banana, cocktail, whiskey, pumpkin, beverage, pepper, cereal, brandy, sweet, wine, tobacco, mug, cherry, donut, nuts, liquor, berry, rice, mustard, cigar, cigarette, alcohol, raspberry, champagne, pot, apple, peel |
| substance<br>material<br>artifact | material –<br>farm<br>animals | cow, wool, charcoal, sheep, cattle, food, animal, wood, goat, wheat, sand, animals, salt, water, timber, fish, mud, straw, cotton, copper, washing, oil, ox, iron, lamb, fresh, abundance, fur, coal, fishing, exotic, dye, ceramic, camel, pollution, tin, licking, smoking, diet, vitamin |
| artifact<br>covering<br>clothing | clothing /<br>fashion | wig, clothes, dress, shoes, jacket, sweater, skirt, sunglasses, leather, hair, costume, shirt, haircut, cloth, socks, waist, mannequin, collar, jewelry, tattoo, lingerie, beard, blonde, mask, fabric, uniform, necklace, linen, outfit, glove, hat, fashion, blanket, bikini, knitting, swimsuit, crochet, badge, coat, carpet, bracelet, arms, makeup |
| artifact<br>structure<br>whole | classical<br>architecture | tower, building, marble, staircase, fountain, doorway, roof, chapel, steeple, porch, ceiling, mural, glass, wall, brick, statue, stone, arch, monument, dome, window, gravestone, sculpture, aisle, tiles, gate, interior, painted, decoration, concrete, church, graveyard, cathedral, curtain, painting, palace, clock, grave, portrait, choir, architecture, pyramid, memorial, square, castle, skyscraper, museum, cemetery, temple, organ |

| | | |
|---|---|---|
| change<br>color<br>visual property | colour / decor | blue, bright, green, pink, black, yellow, dark, white, purple, red, brown, violet, rainbow, colour, orange, sky, rusty, silhouette, grey, diamond, redhead, light, flame, peacock, mirror, color, tiny, shadow, stripes, dull, rose, neon, colorful, crystal, bell, moon, horizon, arrow, silver, ivy, gold, swan, dragon, lantern, star, pearl, horn, ray, fox, globe, planet, bold, belt |
| body part<br>part<br>artifact | body parts | skin, spine, neck, bone, chest, throat, shoulder, wrist, stomach, ear, jaw, cheek, lips, nose, eyes, eye, limb, toe, belly, skull, abdomen, finger, teeth, elbow, cord, whiskers, knee, thumb, tooth, muscle, ankle, tail, paws, lip, brain, flesh, leg, body, calf, heart, blood, tongue, brow, pain, tear, blade, mouth, liver, gut, arm, marrow, curled, canine, feathers, foot, vein, hip, cancer |
| attribute<br>whole<br>artifact | measures & Misc | flexible, reflection, pattern, sharp, ripples, large, elastic, normal, angle, object, spiral, fragile, dense, different, relaxed, frame, strong, fast, target, small, bottom, wave, long, rough, illusion, cone, narrow, texture, pair, noise, curve, bubble, depth, droplets, display, footprint, condition, wide, sphere, reduce, hole, blurred, lamp, short, shell, rapid, medium, plate, size, lens, instrument, feet, helium, chain, meter, inch, cell, adult, formula, males |
| artifact<br>instrumentality<br>move | objects | bag, cardboard, bucket, wire, hand, nail, pencil, hanging, rope, skateboard, knife, garbage, splash, button, scratch, pipe, ink, dripping, dirty, boot, spoon, drawer, hard, dirt, cage, suds, miniature, box, puddle, graffiti, hang, drum, jar, swing, metal, collage, pin, pillow, tough, rock, surf, cradle, vintage, stencil, origami, keyboard, disc, rod, big, rattle, racket, ipod, vinyl, lego, surfers, odd, basket, tag, van, mac |
| person<br>organism<br>bird | animals | bird, cat, squirrel, owl, rabbit, dog, birds, parrot, zebra, giraffe, stork, duck, goose, pelican, deer, elephant, rat, snake, eagle, pigeon, hamster, wolf, cheetah, hawk, mallard, crab, poodle, chipmunk, frog, flamingo, mouse, tiger, pets, crow, whale, gull, wild, insect, feline, prey, hummingbird, hound, pug, lion, panda, pet, lizard, bee, ant, dragonfly, nest, zoo, jellyfish, hen, seagull, spider, wasp, terrier, aquarium, butterfly |
| structure<br>artifact<br>area | room | kitchen, room, bedroom, bathroom, garage, shop, cafe, motel, cellar, diner, closet, hallway, cottage, hotel, sidewalk, restaurant, barn, house, apartment, door, pub, alley, stairs, sofa, patio, bed, floor, couch, cabin, bakery, store, booth, crib, dinner, desk, furniture, hut, parking, fence, inn, pool, corner, shelter, hall, farm, lawn, street, shed, bar, mill, lab, windmill, sitting, office, hospital, log, classroom, shopping, supper, bath, jail, lunch, theatre, yard |
| person<br>organism<br>causal agent | social roles: family members & professions | father, friend, mother, lover, uncle, wife, daughter, lawyer, woman, brother, teacher, son, child, nurse, nephew, banker, soldier, couple, maid, gentleman, husband, author, bride, doctor, priest, wedding, partner, photographer, worker, actor, lady, captain, employee, sailor, groom, appointment, leader, student, king, secretary, scientist, singer, queen, guardian, professor, president, princess, actress, justice, children, instructor, monk, prince, birthday, maker, sheriff, bishop, manager, mayor, companion, chair, minister, politician, boxer, age, pupil, saint, jean, rabbi |

| | | |
|---|---|---|
| object<br>artifact<br>physical entity | places | shore, corridor, trail, bridge, road, harbour, river, tunnel, area, park, beach, pond, valley, lake, hill, ledge, city, railroad, island, highway, harbor, rail, downtown, seashore, canyon, west, canal, border, coast, north, town, mountain, pier, path, traffic, bay, ocean, cliff, forest, swamp, port, abandoned, skyline, stream, line, south, boundary, waterfall, station, loop, sea, railway, construction, boardwalk, scenery, reef, branch, lighthouse, demolition, landscape, underground, airport, zone, urban, metro, region, capital, gauge, village, population |
| instrumentality<br>travel<br>vehicle | transportation | vehicle, airplane, truck, car, elevator, automobile, aircraft, cab, carriage, bike, jet, chopper, scooter, balloon, bicycle, pilot, deck, train, wagon, gasoline, motorcycle, plane, craft, machine, engine, boat, taxi, cannon, crane, tank, escalator, mechanic, ship, hose, driver, steel, rocket, container, gun, safety, auto, motor, explosion, flying, factory, air, flight, camera, appliance, accident, drive, aluminum, telephone, bus, underwater, lighting, vessel, aerial, phone, emergency, ford, exit, subway, company, police, pod, tram, industrial, asphalt, wing |
| change<br>act<br>be | verbs | bring, get, come, want, go, keep, take, know, find, say, give, make, understand, put, listen, enjoy, feel, leave, think, learn, imagine, gather, believe, fail, arrange, add, lose, create, way, hear, send, meet, collect, carry, avoid, buy, remain, allow, appear, might, enter, arrive, seem, entertain, break, steal, receive, stop, stand, build, locked, compare, retain, sell, handle, danger, eat, wander, face, unhappy, protect, please, pray, become, walk, expand, travel, plenty, greet, inspect, comfort, huge, possess, dominate, attach, roam, participate, speak, step, drawn, construct, replace, divide, great, living |
| person<br>organism<br>causal agent | art /<br>entertainment | smile, fun, happy, love, girl, kid, kids, boy, baby, dad, mom, kiss, dude, friends, funny, man, joy, angel, beautiful, christmas, cute, movie, night, spirit, beast, bunny, mad, sing, puppy, monster, soul, zombie, song, devil, dance, kitty, guy, bunch, happiness, snowman, show, holiday, buddy, music, restless, theme, sketch, nice, boys, dead, clown, young, quest, girls, vacation, celebration, emotion, carnival, dreary, dawn, bad, cop, sleep, journey, concert, pride, hero, evening, story, demon, sad, morning, warrior, jazz, band, guest, film, god, piano, punk, doodle, guitar, tv, television, husky, violin, festival, female |
| travel<br>act<br>group | sport | time, day, year, second, course, run, win, game, home, sports, ball, trip, season, week, country, match, track, dropped, club, parade, trick, world, crowd, august, month, horse, winner, swimming, field, football, left, men, triumph, women, gymnastics, basketball, bench, table, racing, round, jump, outdoor, cup, top, swim, race, side, baseball, sailing, opponent, champion, goal, held, school, trial, played, camp, cross, flag, bowl, summer, rally, squad, head, old, ceremony, military, hockey, exhibition, skating, state, bull, college, purse, army, pole, stadium, ski, chess, navy, minute, class, posted, skate, anchor, colt, seat, stud, turkey, santa, mare |

| | | |
|---|---|---|
| abstraction<br>communication<br>act | writing /<br>Misc | fact, discussion, work, idea, read, sense, quote, manner, words, conversation, information, book, picture, value, image, reader, view, person, advertisement, paper, vision, impression, communication, nature, phrase, page, paragraph, proof, article, interest, job, definition, money, abstract, poster, formal, wisdom, reading, skill, choice, attention, literature, letter, handwriting, art, business, smart, awareness, confidence, word, key, design, new, essential, model, date, computer, action, collection, payment, note, law, graphic, figure, bible, library, protest, task, news, violent, chapter, umbrella, movement, dollar, magazine, symbol, photography, modern, newspaper, web, activity, circle, number, people, peace, market, map, self, card, code, psychology, text, right, parent, dictionary, order, party, language, journal, written, tax, style, era, calendar, cent, ad, ancient |

Table D.1: Members of the 20 clusters in $E_L$. Clusters are ordered by size.

| **WordNet label** | **Own label** | **Members** |
|---|---|---|
| base<br>layer<br>flatware | plate | plate |
| lick<br>cream<br>beating | licking | licking |
| communication<br>promotion<br>message | ad | ad, advertisement |
| change<br>passage<br>tube | pipe | rust, pipe, hose, tank, graffiti, chain |
| artifact<br>line<br>whole | train | railway, railroad, subway, curve, tunnel, run, shelter, train, station, tram, highway, track, rail, way, engine, stop, gate, bridge, smoke |
| structure<br>area<br>room | room | classroom, hallway, hall, closet, bedroom, room, bathroom, garage, office, cafe, museum, doorway, kitchen, shop, restaurant, store, mannequin, stadium, market, ceiling, corner |
| bird<br>vertebrate<br>person | animals | hummingbird, gull, peacock, hawk, pelican, crow, parrot, seagull, wing, swan, pigeon, owl, goose, flamingo, nest, eagle, tail, bird, silhouette, duck, chest, body, ledge, giraffe, zebra |
| travel<br>wheeled vehicle<br>self-propelled vehicle | vehicles | cab, car, taxi, police, vehicle, automobile, drive, racing, scooter, bike, van, street, road, motorcycle, truck, speak, wagon, bus, parade, drawn, asphalt, cop, parking, bicycle, sidewalk, traffic, driver, carriage, meter |
| plant organ<br>plant<br>vascular plant | plants | bloom, foliage, grave, dead, vine, blossom, ivy, pod, cactus, tree, moss, root, leave, limb, forest, bush, plant, lily, branch, weed, leaf, vein, sunshine, log, fence, flower, sunlight, wood, palm, bench, sun |
| structure<br>artifact<br>whole | building<br>parts | chapel, cottage, steeple, castle, dome, story, cathedral, build, skyscraper, arch, lighthouse, apartment, hut, angel, shed, hotel, monument, window, staircase, home, cabin, house, roof, porch, tower, sculpture, patio, bell, deck, brick, church, cross, clock, step, statue |

| | | |
|---|---|---|
| instrumentality<br>container<br>substance | vessel | champagne, tea, beverage, alcohol, honey, milk, pencil, tulip, juice, oil, bakery, ceramic, container, coffee, tin, cup, beer, sunflower, daisy, wine, rose, marble, bowl, sweet, maker, jar, vessel, mug, money, bottle, pumpkin, straw, glass, basket, box, pot, bucket, bunch |
| body part<br>artifact<br>part | pets & body parts | jaw, throat, pupil, cheek, canine, belly, brow, mouth, stomach, tongue, eye, nose, poodle, ear, hamster, lip, fur, tooth, teeth, pet, leg, wool, head, feline, toe, panda, smile, neck, face, beard, puppy, collar, horn, skin, cat, kitty, calf, nail, dog, tag, mother |
| physical entity<br>body of water<br>thing | water | rapid, village, coast, bay, mist, horizon, canal, skyline, valley, sea, cliff, fog, town, waterfall, stream, water, sunset, pier, harbor, boardwalk, break, ocean, lake, fountain, shore, island, river, wave, splash, city, rock, ship, building, sand, hill, crane, mountain, beach, pond, surf, boat, pool |
| location<br>artifact<br>region | farm animal | dandelion, boundary, grass, wild, deer, stork, field, mud, farm, windmill, garden, landscape, desert, cattle, dirt, area, barn, yard, zoo, ox, path, footprint, garbage, puddle, lawn, cow, sheep, concrete, snow, eat, lamb, goat, stone, cone, trail, rain, day, park, animal, cage, horse, bull, elephant |
| change<br>color<br>visual property | colors | bright, beautiful, big, dirty, small, colorful, grey, long, purple, dark, round, men, tiny, pink, eyes, painted, brown, gold, medium, white, hang, iron, silver, old, black, left, tall, red, safety, large, metal, blue, steel, yellow, leather, hanging, make, walk, green, right, color, bath, pair, washing, sitting, carry |
| food<br>produce<br>solid | food | drizzle, nuts, herb, beef, flour, season, cereal, cherry, breakfast, sugar, steak, bacon, burger, butter, rice, meat, meal, sauce, dinner, pie, raspberry, lunch, sushi, bean, mustard, pepper, seed, salt, soup, cheese, tomato, hot, berry, potato, dessert, strawberry, salad, cardboard, food, bone, lemon, burn, frost, chocolate, bread, turkey, sandwich, spoon, pizza, chicken, shell, candy, peel, cooking, bubble, knife, fruit, fish, donut, cake, apple, ice, banana, orange |
| artifact<br>whole<br>instrumentality | furnishing | crochet, calendar, linen, map, painting, work, frog, skull, note, code, stud, lantern, art, telephone, scratch, furniture, information, collection, menu, ipod, page, table, mural, piano, spring, movie, magazine, poster, cell, spine, portrait, appliance, desk, paper, graphic, frame, bed, date, crib, pattern, text, picture, card, globe, butterfly, wall, pillow, fabric, cord, sofa, carpet, guitar, square, cloth, image, tv, book, heart, lamp, star, television, blanket, couch, newspaper, night, decoration, mirror, time, computer, design, keyboard, word, mouse, border, drawer, floor, button, chair, key, display, curtain, reading |
| person<br>artifact<br>covering | people | fun, nurse, lingerie, violin, jewelry, makeup, haircut, cigar, wig, monk, instructor, santa, pug, brother, doctor, dad, terrier, huge, parent, scientist, gentleman, bikini, pearl, badge, bracelet, shirt, swimsuit, sweater, jean, costume, hip, jacket, sleep, daughter, mom, short, skirt, snowman, hat, man, muscle, instrument, necklace, young, basketball, wrist, hair, smoking, glove, outfit, music, coat, rabbit, pets, woman, band, football, father, dude, boot, hand, elbow, tattoo, arm, ankle, soldier, lab, waist, clown, dress, belt, racket, blonde, bunny, uniform, loop, lens, friend, cigarette, held, finger, girl, photographer, purse, person, knee, pin, boy, female, trick, thumb, guy, mask, foot, son, swing, clothes, lady, bride, skate, squirrel, bag, phone, disc, ski, tiger, child, groom, adult, shoulder, student, kid, camera, skateboard, baseball, ball, baby |

| | | |
|---|---|---|
| change<br>act<br>artifact | Misc | pain, downtown, capital, condition, theatre, motel, cemetery, elevator, journey, class, zone, captain, coal, military, navy, school, craft, gauge, texture, exit, storm, language, moon, company, create, club, anchor, country, construction, meet, rainbow, weather, port, alley, hospital, party, take, flight, pilot, dragon, booth, interior, business, race, sky, library, drum, sunny, door, motor, employee, light, model, hen, bulb, goal, gun, wind, cloud, diner, pole, aircraft, course, fox, rod, skating, letter, jump, show, written, flame, symbol, reflection, plane, shadow, object, diamond, airport, ray, circle, line, airplane, swimming, bottom, arrow, flag, crowd, balloon, top, number, aquarium, fire, flying, seat, side, stand, figure, air, handle, game, winter, view, match, blade, bar, machine, family, wire, lion, hole, people, shade, worker, jet, rope, umbrella, couple |
| person<br>change<br>organism | Misc | ant, news, jellyfish, protest, add, imagine, inn, journal, liver, essential, marrow, rattle, arrange, wasp, paragraph, brandy, fact, aerial, devil, unhappy, emotion, chipmunk, god, oak, explosion, prey, proof, vision, activity, chess, movement, danger, gasoline, secretary, jazz, song, send, mayor, tobacco, soul, urban, violent, quote, demon, replace, fragile, manner, misty, receive, ancient, flowers, skill, reef, ripples, rally, living, diet, sketch, awareness, illusion, pollution, abstract, value, wisdom, squad, remain, arrive, saint, trial, impression, avoid, vinyl, minister, maid, concert, believe, jail, learn, please, politician, great, guardian, population, holiday, cancer, psychology, become, college, demolition, payment, brain, army, rabbi, lawyer, literature, prince, task, tropical, bring, lover, bold, inch, interest, companion, exhibition, leader, noise, actor, underwater, supper, communication, helium, sense, happiness, win, sad, gymnastics, entertain, champion, banker, odd, conversation, planet, dawn, dense, camp, law, locked, pray, lose, plenty, abundance, fail, mallard, vacation, chapter, dreary, warrior, origami, might, joy, timber, choice, underground, depth, stencil, formula, friends, allow, retain, participate, understand, paws, mad, pride, stairs, wander, comfort, theme, give, nephew, reduce, funny, bad, idea, droplets, age, ..., surfers |

Table D.2: Members of the 20 clusters in $E_S$. Clusters are ordered by size.

| **WordNet label** | **Own label** | **Members** |
|---|---|---|
| bird<br>aquatic bird<br>seabird | birds | seagull, gull, goose, duck, pelican, swan, mallard, stork, eagle, flamingo |
| furnishing<br>furniture<br>instrumentality | furnishing | furniture, stand, booth, desk, modern, display, bed, chair, container, door, appliance, drawer, sofa, curtain, couch, bench, crib, frame, box, table, tv, window, computer, cradle, television, mac |
| instrumentality<br>artifact<br>device | objects | inspect, protect, collar, find, skateboard, gasoline, heavy, key, belt, steal, instrument, hang, justice, glove, handle, knife, scooter, horn, shoes, pipe, bone, telephone, mouse, bag, hat, spoon, guitar, gun, colt, purse, drum, iron, boot, violin, spine, umbrella, sunglasses |
| instrumentality<br>self-propelled vehicle<br>wheeled vehicle | car related | accident, cord, vehicle, auto, automobile, skate, photography, truck, race, arrive, ford, chopper, cab, rally, seat, industrial, smart, mechanic, racing, car, demolition, triumph, construction, motorcycle, machine, taxi, engine, driver, crane, carriage, van, bus, cannon, motor, tank, hockey, wagon, camera |

| | | |
|---|---|---|
| person<br>organism<br>causal agent | “female topics” | woman, model, brandy, pink, actress, lady, girl, young, wife, tiny, haircut, blonde, women, girls, hot, mother, hair, portrait, body, makeup, cheek, wig, neck, muscle, chest, lingerie, waist, redhead, child, face, bride, belly, bikini, kid, swimsuit, baby, brow, skirt, dress, short |
| instrumentality<br>artifact<br>device | metals & writing | object, aluminum, journal, author, capital, lawyer, step, cardboard, law, silver, elastic, bible, written, book, tin, literature, chocolate, wire, money, cigarette, stud, steel, payment, glass, charcoal, blanket, gold, newspaper, page, cigar, appointment, brick, butter, pencil, mirror, log, phone, ipod, match, pillow, rod, piano, keyboard |
| vascular plant<br>plant<br>grow | plants | weed, bunch, maple, cancer, iris, poppy, dandelion, leave, flower, rose, foliage, grow, plant, cactus, spring, tulip, ivy, palm, lily, leaf, daisy, tree, root, wheat, wool, raspberry, tobacco, flowers, blossom, butterfly, sunflower, cotton, herb, violet, oak, moss, strawberry, nest, dew, berry, rice, branch, coal |
| food<br>nutriment<br>substance | food | sushi, meal, sandwich, pie, breakfast, lunch, food, supper, flour, cereal, sweet, dessert, dinner, subway, diet, cake, date, steak, sauce, bread, copper, nuts, bacon, cooking, beef, meat, bakery, knitting, eat, potato, salad, donut, pizza, burger, coffee, soup, bean, cheese, vitamin, fruit, pumpkin, rock, marrow, market, timber |
| artifact<br>change<br>cover | colours & materials | texture, fabric, cloth, metal, rain, concrete, paper, suds, rough, words, stone, wall, square, dense, leather, quote, wood, frost, mud, noise, text, purple, carpet, blue, tiles, dirt, droplets, red, sand, fog, formula, mist, pattern, handwriting, green, straw, linen, asphalt, stripes, crowd, marble, yellow, black, brown, grey, grass, white |
| body part<br>artifact<br>part | body parts | gut, throat, wrist, burn, ear, thumb, elbow, listen, shoulder, liver, pain, knee, arms, hand, toe, finger, give, tongue, limb, abdomen, jaw, receive, nail, arm, feet, hear, skin, washing, head, ankle, hip, teeth, tear, stomach, brain, foot, lip, mouth, leg, flesh, mask, eyes, nose, skull, eye, socks, lips |
| structure<br>artifact<br>area | room | museum, garage, hall, classroom, kitchen, cellar, interior, office, diner, decoration, exhibition, hotel, ceiling, restaurant, store, bathroom, trial, pub, class, closet, cafe, room, porch, stairs, deck, hospital, living, corridor, aisle, bar, staircase, doorway, hallway, chapel, floor, lab, station, bedroom, gate, elevator, theatre, escalator, tunnel, organ, alley, library, jail, tram |
| artifact<br>whole<br>instrumentality | fruit, drinks & sport | compare, sad, ceramic, tea, rattle, honey, mustard, weather, champagne, pearl, button, wine, sugar, peel, pepper, jewelry, milk, orange, balloon, bulb, lemon, beer, cocktail, salt, beverage, sphere, juice, sports, planet, sun, whiskey, lantern, world, cup, football, pin, diamond, banana, basket, cherry, cent, basketball, globe, ripples, vinegar, pot, bottle, jar, tomato, baseball, plate, bucket, bowl, bubble, mug, ball, moon |
| travel<br>change<br>object | vacation | island, view, reflection, harbor, nice, side, sea, summer, tropical, pollution, port, aircraft, pier, travel, surfers, journey, sunny, coast, flying, morning, ocean, seashore, horizon, mare, holiday, lake, surf, shore, vacation, bay, airport, cliff, sunlight, air, river, storm, ship, fishing, beach, desert, harbour, puddle, flight, sailing, evening, sunrise, skyline, vessel, lighthouse, dawn, sunset, rocket, mountain, whale, underwater, boat, swimming, swim, plane, dusk, jet, cloud, sky, airplane, ski |

| | | |
|---|---|---|
| change<br>abstraction<br>state | festival | theme, wisdom, soul, image, possess, large, confidence, happiness, beautiful, joy, love, ceremony, festival, movement, abundance, dead, depth, celebration, lover, run, demon, blurred, pray, happy, remain, wet, dance, navy, family, carnival, angel, sculpture, ray, dragon, drive, atmosphere, night, shadow, band, god, believe, party, dark, hanging, abstract, show, christmas, monster, devil, jump, lighting, sunshine, warrior, painting, water, aquarium, zombie, concert, haze, crystal, statue, explosion, jazz, jellyfish, wave, bright, rainbow, ice, light, smoke, club, neon, colorful, hole, protest, autumn, rust, reef, flame, fire |
| person<br>organism<br>causal agent | animals | animals, animal, picture, painted, zoo, turkey, curled, goat, companion, pets, canine, pet, prey, relaxed, horse, spirit, tail, dog, chipmunk, squirrel, pigeon, fox, cute, please, sheep, owl, birds, military, giraffe, lion, lamb, bee, insect, hamster, hawk, licking, bird, cat, puppy, feline, terrier, deer, calf, rat, chicken, camel, dragonfly, whiskers, poodle, cow, hound, cattle, lizard, fish, bunny, crow, wolf, tiger, parrot, zebra, cheetah, fur, panda, bull, wasp, ox, hen, frog, crab, snake, boxer, hummingbird, rabbit, elephant, pupil, husky, peacock, spider, pug, ant |
| change<br>abstraction<br>travel | Misc | think, condition, understand, know, meet, sing, symbol, bring, speak, awareness, say, strong, sense, music, song, come, stencil, badge, loop, avoid, long, tag, idea, feel, bell, helium, guest, held, heart, proof, film, tall, information, oil, meter, anchor, female, drawn, flexible, smile, peace, break, note, paragraph, figure, attach, gauge, apple, wander, kitty, paws, silhouette, footprint, hose, locked, vinyl, corner, round, divide, curve, cross, target, wing, lens, necklace, tooth, border, rope, lamp, bracelet, minute, north, time, illusion, cone, swing, racket, angle, circle, chain, clock, bike, bicycle, pole, spiral |
| person<br>organism<br>causal agent | people | monk, manager, student, males, banker, instructor, parent, politician, minister, worker, adult, professor, played, employee, pilot, bottom, husband, style, uncle, business, men, boys, son, captain, dude, teacher, man, mayor, top, beard, dad, boy, retain, cop, fail, uniform, outfit, company, priest, nurse, daughter, maid, opponent, father, scientist, police, children, sailor, friends, beast, restless, sitting, kids, old, bishop, prince, punk, costume, people, tattoo, groom, president, couple, blade, secretary, saint, sheriff, singer, mad, walk, pod, doctor, photographer, guy, skating, person, formal, bush, actor, gentleman, rabbi, queen, sleep, funny, soldier, jacket, sweater, coat, shirt, jean |
| structure<br>artifact<br>whole | landmark | village, mill, cemetery, country, graveyard, boardwalk, bath, memorial, outdoor, wide, ancient, temple, inn, path, town, abandoned, windmill, landscape, canal, downtown, trip, cottage, scenery, architecture, farm, patio, roam, palace, camp, drizzle, factory, monument, road, apartment, street, shelter, nature, tower, grave, wind, fountain, season, way, flood, castle, barn, exotic, city, cabin, shade, school, aerial, arch, ledge, garbage, motel, railroad, railway, hill, house, bridge, highway, dreary, garden, train, dome, trail, day, church, winter, urban, parade, home, waterfall, dull, canyon, traffic, cathedral, building, yard, skyscraper, steeple, pool, rail, wild, stadium, forest, mural, pyramid, track, park, field, hut, pond, roof, shed, fence, sidewalk, stream, valley, snow, swamp, lawn |

| | | |
|---|---|---|
| change<br>act<br>artifact | Misc | learn, seem, course, dropped, reading, gather, create, reader, impression, might, champion, partner, advertisement, friend, hard, dye, comfort, trick, vision, construct, craft, small, goal, violent, poster, movie, conversation, participate, communication, read, population, huge, smoking, discussion, underground, tough, become, build, carry, leader, college, pair, tax, fashion, fast, graphic, misty, miniature, odd, big, imagine, cold, collage, shopping, shop, graffiti, magazine, color, dirty, choir, ink, unhappy, different, vintage, wedding, king, seed, arrange, psychology, kiss, birthday, cell, plenty, bloom, princess, boundary, lego, snowman, crochet, sketch, gymnastics, emotion, santa, art, origami, clown, narrow, mannequin, army, chess, rusty, blood, collection, dripping, cage, colour, clothes, alcohol, liquor, candy, flag, age, metro, dollar, gravestone, feathers, map |
| act<br>change<br>abstraction | Misc | activity, great, put, replace, lose, want, order, buy, allow, august, reduce, south, essential, keep, posted, bold, pride, fun, west, game, job, action, safety, buddy, story, entertain, get, week, maker, collect, skill, language, fact, normal, interest, hero, value, work, bad, self, attention, brother, greet, chapter, danger, appear, nephew, ad, size, medium, year, dominate, enjoy, era, task, mom, emergency, sell, news, go, zone, guardian, send, take, left, second, choice, word, card, web, quest, add, make, phrase, dictionary, sharp, winner, line, scratch, arrow, vein, number, shell, splash, parking, enter, rapid, disc, new, right, win, stop, manner, fresh, calendar, squad, month, vine, exit, fragile, region, article, expand, menu, design, area, state, inch, definition, doodle, code, letter, star |

Table D.3: Members of the 20 clusters in $E_V$. Clusters are ordered by size.

| **WordNet label** | **Own label** | **Members** |
|---|---|---|
| baby<br>organism<br>work | baby | baby |
| device<br>weapon<br>hurt | knife | knife |
| area<br>communication<br>mark | footprint | footprint |
| atmosphere<br>condition<br>obscure | sky | cloud, sky |
| line<br>brandish<br>gesticulate | ocean | wave, ocean |
| artifact<br>animal tissue<br>implementation | teeth | tooth, teeth |
| way<br>road<br>artifact | road | road, street, highway |
| organism<br>animal<br>bad person | animal | fox, hen, game |

| | | |
|---|---|---|
| substance<br>food<br>grass | food | cereal, soup, oil |
| nonvascular organism<br>moss<br>bryophyte | alpine plant | moss, ivy, cliff |
| aircraft<br>craft<br>airplane | airplane | aircraft, airplane, jet, plane |
| instrumentality<br>device<br>artifact | computer | keyboard, mouse, computer, key |
| food<br>beverage<br>substance | drink | beverage, wine, beer, juice |
| body part<br>process<br>part | body parts | ear, head, eye, horn, tail |
| instrumentality<br>artifact<br>substance | pottery | ceramic, tin, pencil, marble, hot |
| bird<br>vertebrate<br>artifact | flying animal | parrot, limb, hummingbird, hawk, owl, dragon, squirrel, branch, butterfly |
| bird<br>aquatic bird<br>seabird | bird | gull, seagull, pelican, swan, peacock, crow, pigeon, goose, flamingo, wing, bird, duck, eagle |
| thing<br>body of water<br>physical entity | water | bay, canal, harbor, water, lake, sea, pier, river, ship, pond, shore, boat, splash, pool |
| change<br>move<br>visual property | body, color | left, long, small, big, muscle, purple, pink, right, washing, green, pair, color, sitting, palm |
| food<br>fruit<br>change | desserts | nuts, sugar, cherry, frost, chocolate, raspberry, flour, dessert, butter, pie, strawberry, candy, lemon, ice, donut, cake |
| group<br>event<br>act | event | party, parade, crowd, booth, race, cafe, stadium, show, family, restaurant, match, people, market, stand, park, airport, student, couple |
| change<br>color<br>visual property | visual property | bright, grey, dark, round, painted, white, gold, silver, black, red, old, brown, blue, tall, yellow, metal, large, hanging |
| object<br>structure<br>artifact | landscape | horizon, skyline, fog, valley, sunset, town, skyscraper, waterfall, moon, lighthouse, stream, city, building, castle, island, fountain, mountain, crane, hill |
| container<br>instrumentality<br>measure | drink, vessel | tea, champagne, alcohol, honey, milk, coffee, cup, container, salt, bowl, mug, maker, spoon, jar, bottle, money, vessel, straw, diner, glass, bucket, basket, pot, bubble |
| artifact<br>whole<br>furnishing | furnishing, pet | linen, sleep, furniture, blanket, bed, spring, crib, pillow, carpet, couch, pattern, sofa, feline, fabric, bunny, cloth, piano, floor, chair, square, cat, leather, chest, patio, kitty, button |
| clothing<br>covering<br>consumer goods | clothing | wig, instructor, jacket, bikini, costume, badge, sweater, shirt, swimsuit, outfit, gentleman, skirt, short, jean, boot, hat, coat, dude, dress, glove, uniform, clothes, soldier, belt, mask, cop, pin, ski |

| | | |
|---|---|---|
| reproductive structure<br>plant organ<br>vascular plant | plants | pod, bloom, tulip, daisy, cactus, sunflower, berry, blossom, sweet, rose, lily, vine, tiny, root, vein, pumpkin, garden, flower, plant, leave, leaf, peel, fruit, bunch, desert, banana, orange, apple |
| artifact<br>part<br>body part | body parts<br>house animals | jaw, throat, canine, belly, pupil, cheek, stomach, hamster, tongue, poodle, mouth, nose, fur, pet, lip, leg, wool, panda, toe, neck, collar, puppy, skin, licking, body, calf, dog, tag, lamb |
| food<br>nutriment<br>meat | food | beef, herb, season, steak, meat, breakfast, bacon, burger, rice, meal, sauce, lunch, mustard, cheese, pepper, dinner, bean, sushi, tomato, seed, potato, salad, food, bone, sandwich, turkey, bread, chicken, pizza, cooking, plate, fish |
| artifact<br>instrumentality<br>substance | office | crochet, calendar, collection, telephone, menu, note, movie, ipod, appliance, magazine, table, frog, cardboard, date, desk, paper, hospital, skull, card, library, box, shell, book, cord, picture, television, steel, tv, drawer, object, newspaper, garbage, night, top, ledge, machine, corner, display, fire |
| abstraction<br>communication<br>change | communication | language, code, information, text, ad, company, graphic, painting, map, written, exit, mural, letter, word, work, art, scratch, poster, symbol, heart, advertisement, star, graffiti, image, page, spine, border, time, arrow, frame, diamond, say, portrait, number, birthday, design, circle, decoration, reading |
| structure<br>artifact<br>area | building | elevator, chapel, hallway, apartment, closet, garage, hall, window, classroom, bedroom, doorway, cathedral, door, bathroom, story, interior, build, museum, cabin, room, arch, mannequin, shop, office, club, staircase, store, hotel, reflection, kitchen, tunnel, mirror, pilot, house, ceiling, aquarium, view, curtain, shade, church |
| artifact<br>travel<br>whole | transportation | zone, railway, construction, curve, create, taxi, run, subway, car, cab, drive, automobile, railroad, business, parking, alley, shelter, tram, vehicle, stop, asphalt, course, way, light, train, police, station, bus, rail, gate, van, sidewalk, home, line, truck, track, concrete, traffic, bridge, cross, meter, brick |
| artifact<br>structure<br>whole | farm &<br>wild animals | deer, dandelion, wild, grass, farm, foliage, windmill, field, mud, bush, forest, weed, landscape, shed, barn, zoo, hut, tree, cattle, area, dirt, fence, rock, log, goal, ox, yard, cow, sheep, goat, lawn, eat, animal, giraffe, stone, cage, wood, zebra, mother, horse, lion, bull, elephant, hole |
| artifact<br>body part<br>instrumentality | body<br>accessories | cigar, haircut, makeup, brow, pug, hip, bracelet, wrist, pearl, tattoo, elbow, stud, smile, ankle, hand, necklace, finger, arm, band, smoking, hair, snowman, beard, waist, thumb, lens, cigarette, loop, woman, burn, cell, knee, purse, racket, face, nail, foot, shoulder, bride, phone, bag, camera, lady, groom, skateboard |
| change<br>travel<br>object | travel | rapid, village, journey, seashore, swamp, theatre, mist, storm, scientist, stork, boundary, sunny, coast, country, boardwalk, sunshine, wet, weather, break, rainbow, dirty, aisle, flight, rain, meet, ray, sand, day, puddle, escalator, lab, trail, beach, path, surf, silhouette, nest, walk, snow, wind, shadow, sunlight, flying, cone, sun, balloon, umbrella |

| | | |
|---|---|---|
| artifact<br>instrumentality<br>device | building<br>vehicle | pain, capital, minute, gauge, coal, cottage, rust, lantern, anchor, angel, speak, steeple, motor, dome, port, iron, pole, globe, rod, pipe, bulb, engine, hose, bell, model, seat, roof, porch, sculpture, monument, flame, handle, tank, lamp, gun, flag, bar, chain, wall, deck, bike, side, bottom, figure, wagon, rope, tower, wire, clock, scooter, step, blade, motorcycle, bench, bicycle, smoke, statue, carriage |
| person<br>organism<br>causal agent | people<br>activities | fun, violin, nurse, brother, lingerie, monk, parent, dad, jewelry, huge, played, santa, doctor, basketball, terrier, instrument, music, captain, take, football, man, father, young, daughter, drum, mom, trick, son, jump, held, pets, men, blonde, friend, employee, colorful, skating, person, guy, boy, swing, girl, safety, photographer, racing, swimming, female, clown, disc, skate, adult, kid, winter, guitar, child, baseball, driver, ball, air, carry, worker |
| person<br>change<br>organism | Misc | rattle, news, song, ant, imagine, send, emotion, arrange, living, jazz, ripples, inn, god, learn, please, violent, fragile, marrow, aerial, misty, inch, unhappy, devil, essential, avoid, squad, tobacco, prey, flowers, banker, urban, protest, replace, saint, psychology, demon, movement, holiday, rabbi, pollution, mayor, illusion, dense, entertain, wisdom, underwater, manner, awareness, politician, pray, give, lawyer, become, participate, supper, trial, vinyl, law, gymnastics, droplets, odd, believe, dawn, brain, secretary, brandy, retain, fail, communication, wasp, interest, gasoline, plenty, concert, helium, noise, locked, demolition, activity, payment, lose, great, literature, allow, bring, nephew, abstract, soul, paws, guardian, win, funny, might, expand, dreary, lover, tax, friends, skill, jail, put, uncle, ancient, joy, tough, tropical, happiness, boys, population, underground, understand, wander, stairs, abundance, value, idea, exhibition, cancer, choice, males, professor, reduce, mad, depth, hockey, discussion, flexible, compare, collect, appointment, exotic, think, seem, confidence, bad, steal, get, birds, dull, ceremony, abandoned, relaxed, sailing, industrial, lips, sunglasses, normal, surfers |
| change<br>person<br>causal agent | Misc | jellyfish, add, fact, journal, proof, paragraph, oak, liver, impression, danger, chipmunk, explosion, vision, chess, quote, rally, diet, prince, remain, receive, minister, sketch, sad, arrive, reef, task, college, leader, origami, stencil, planet, maid, champion, bold, chapter, army, actor, mallard, camp, sense, companion, formula, timber, conversation, warrior, pride, dew, theme, queen, vacation, comfort, age, self, mare, morning, redhead, mill, cold, celebration, reader, flood, phrase, era, cent, evening, zombie, partner, construct, know, violet, cellar, gut, august, manager, winner, copper, hard, autumn, mechanic, singer, month, tiles, bishop, poppy, miniature, festival, justice, attention, spider, blurred, children, listen, colour, animals, women, carnival, hound, girls, definition, triumph, hero, kids, peace, vitamin, week, dusk, dragonfly, job, web, wolf, sunrise, go, smart, author, president, quest, auto, graveyard, heavy, fashion, article, atmosphere, summer, flesh, restless, gather, emergency, cannon, suds, north, sell, vinegar, cute, world, pyramid, ford, handwriting, formal, wife, architecture, ..., wedding |

Table D.4: Members of the 40 clusters in $E_S$. Clusters are ordered by size.

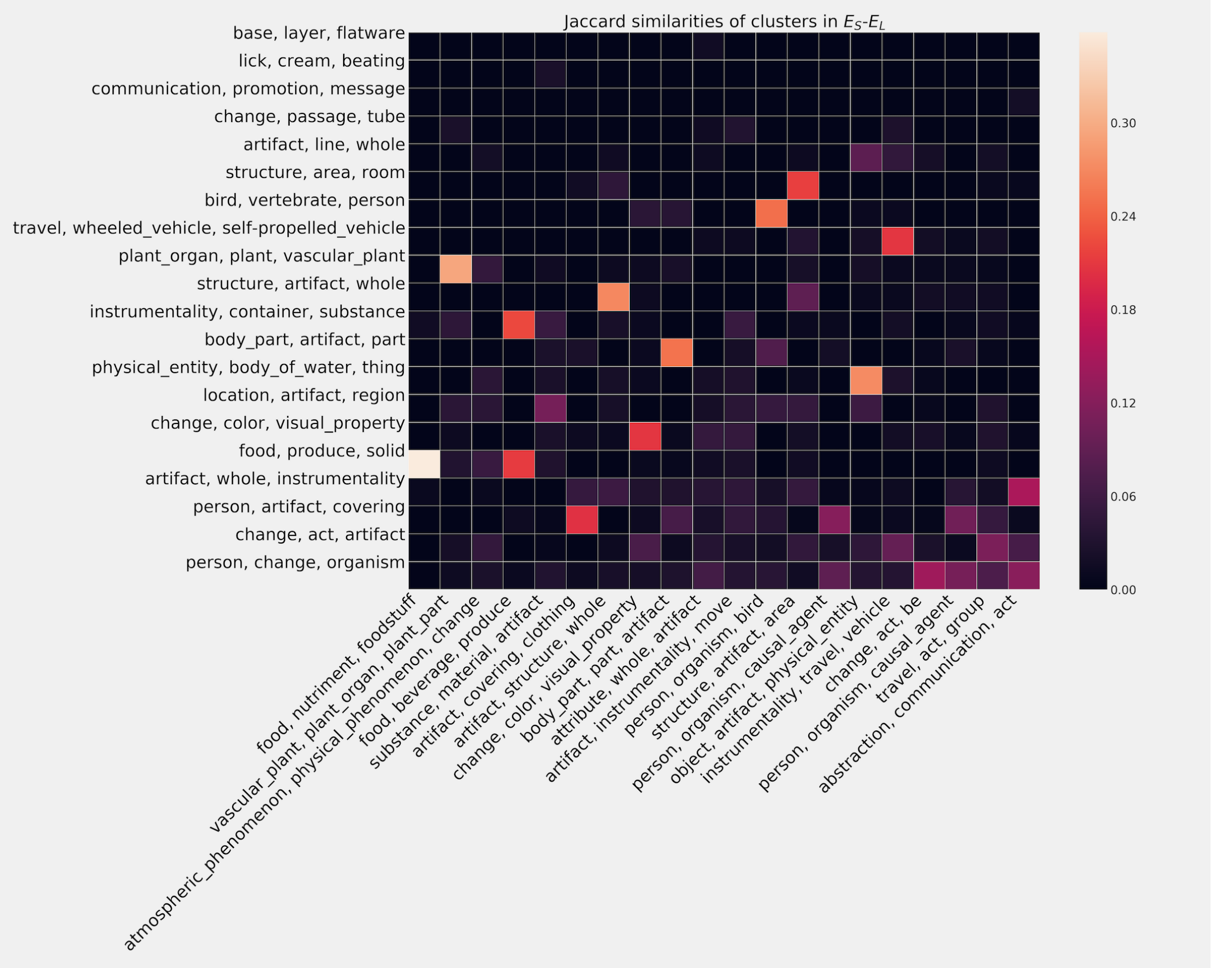


Figure D.1: Heatmap of Jaccard coefficients between K-means clusters of $E_S$ and $E_L$ ($y$ and $x$ axes respectively).

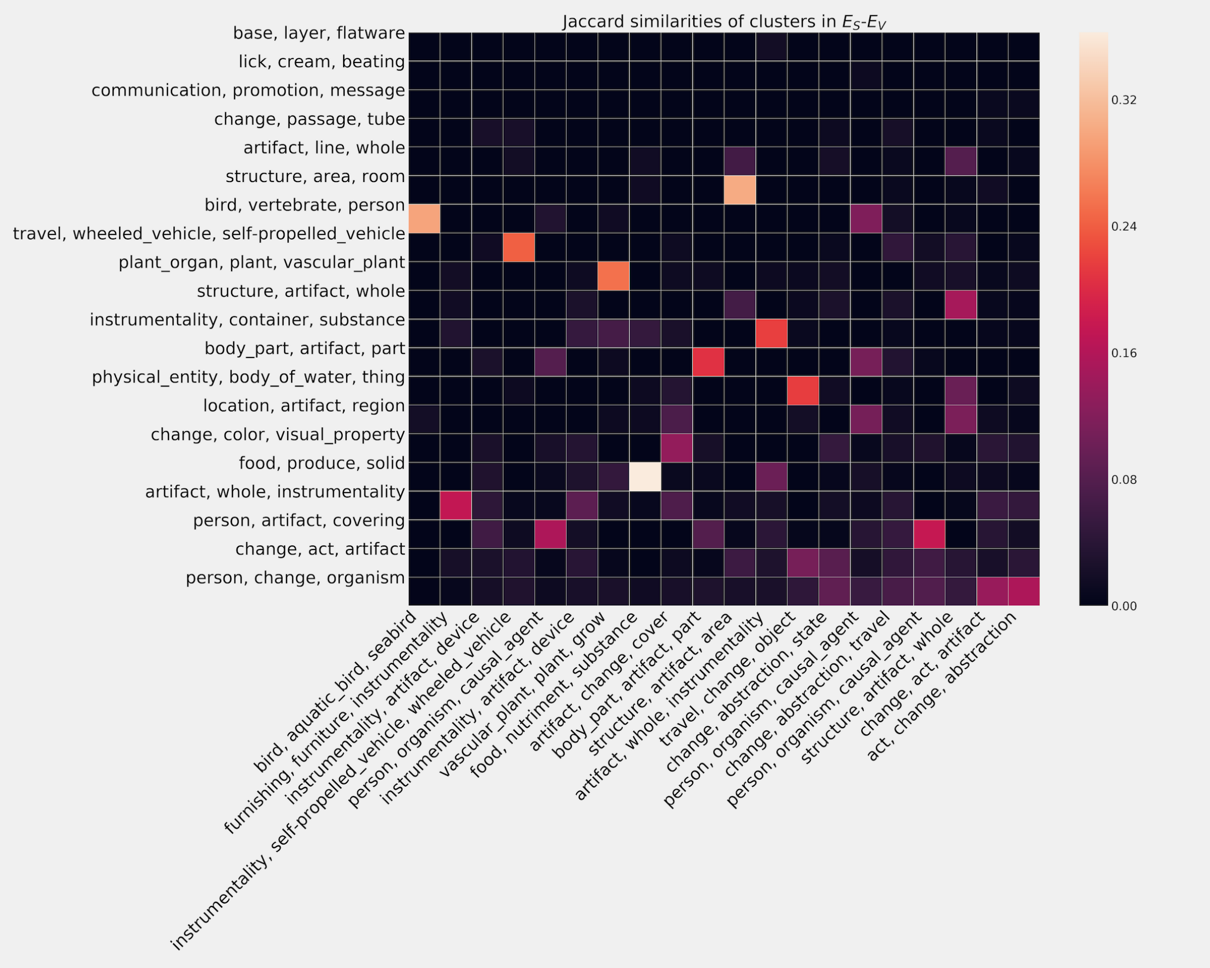


Figure D.2: Heatmap of Jaccard coefficients between K-means clusters of $E_S$ and $E_V$ ($y$ and $x$ axes respectively).

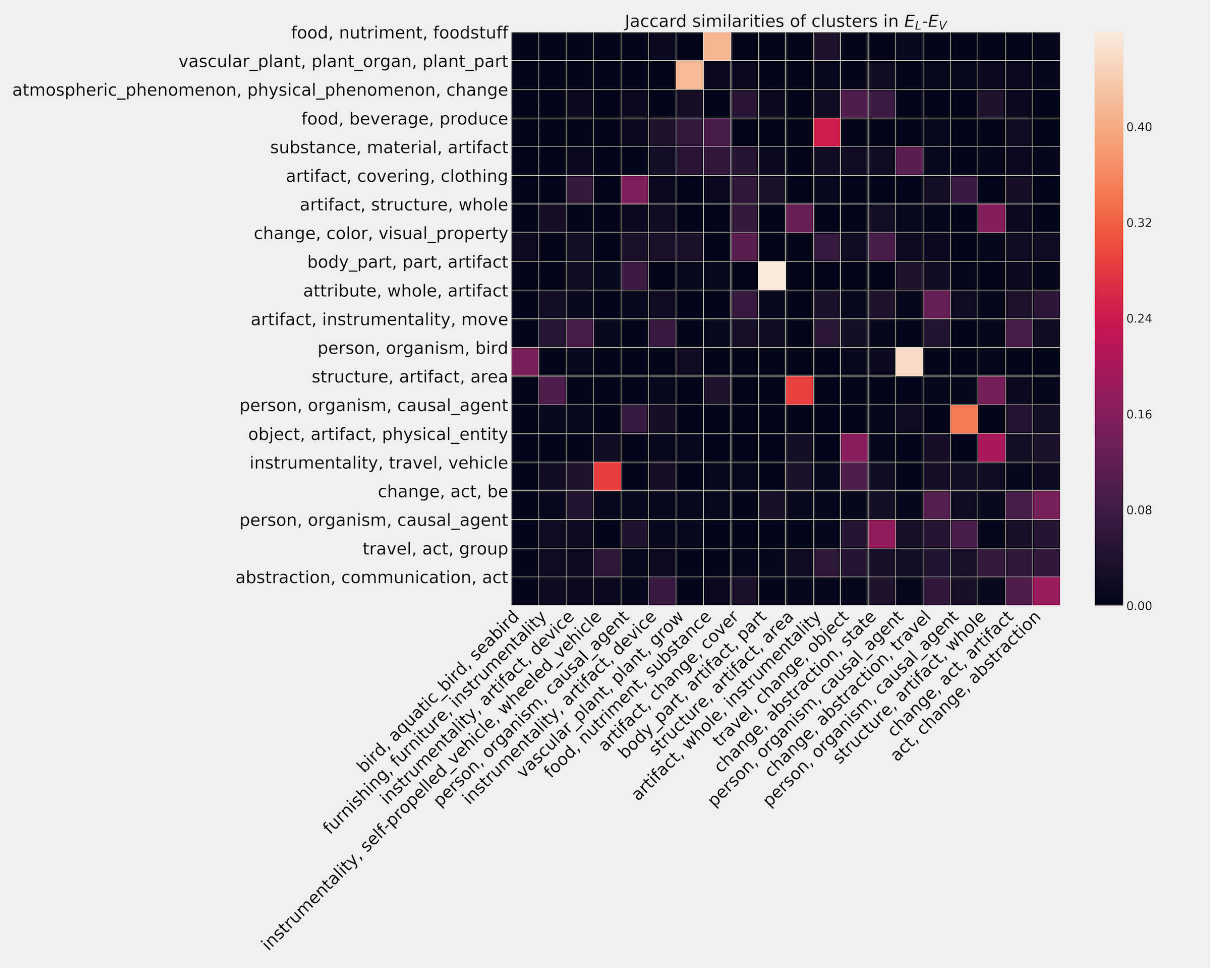


Figure D.3: Heatmap of Jaccard coefficients between K-means clusters of $E_L$ and $E_V$ ($y$ and $x$ axes respectively).

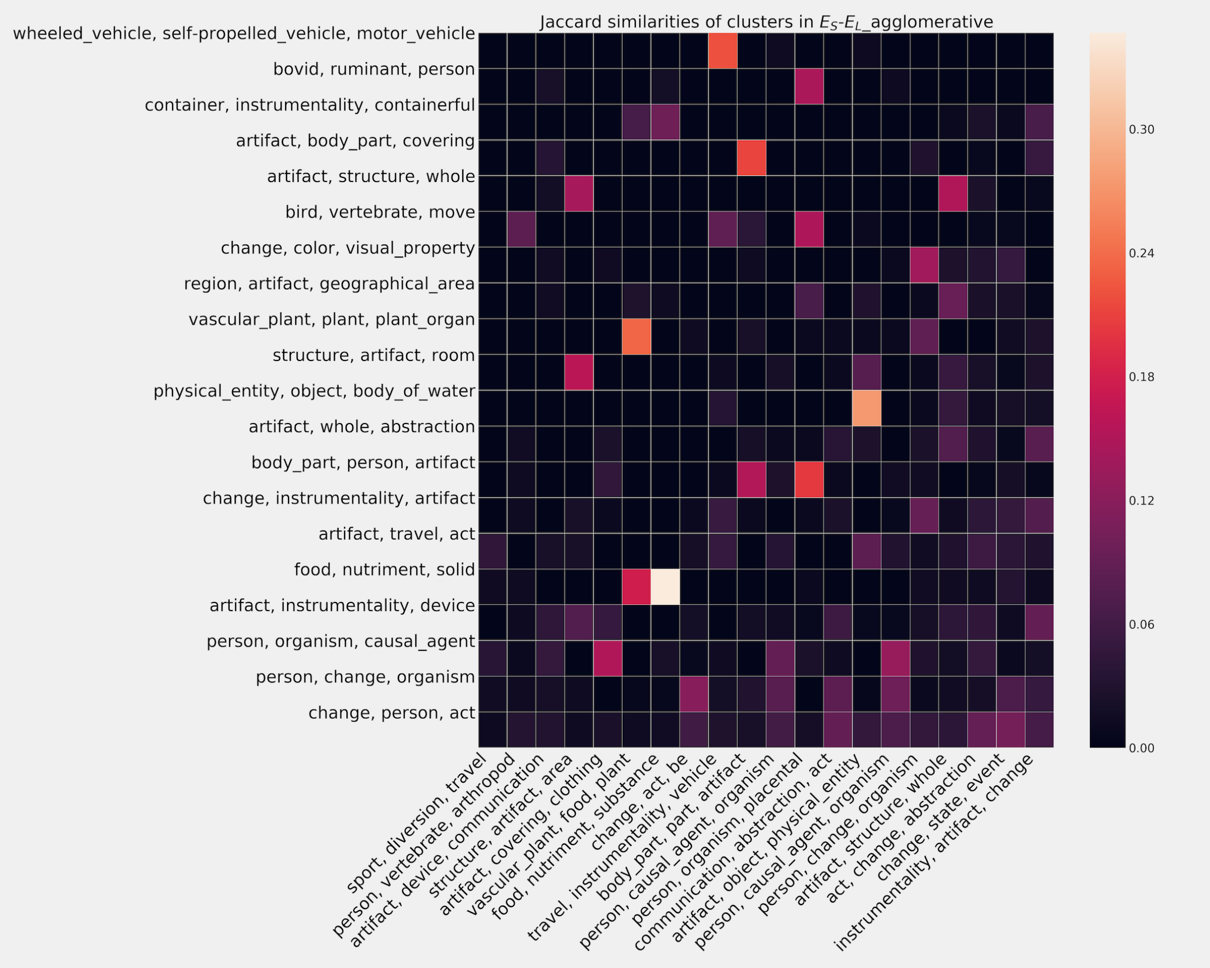


Figure D.4: Heatmap of Jaccard coefficients between Agglomerative clusters of $E_S$ and $E_L$ ($y$ and $x$ axes respectively).

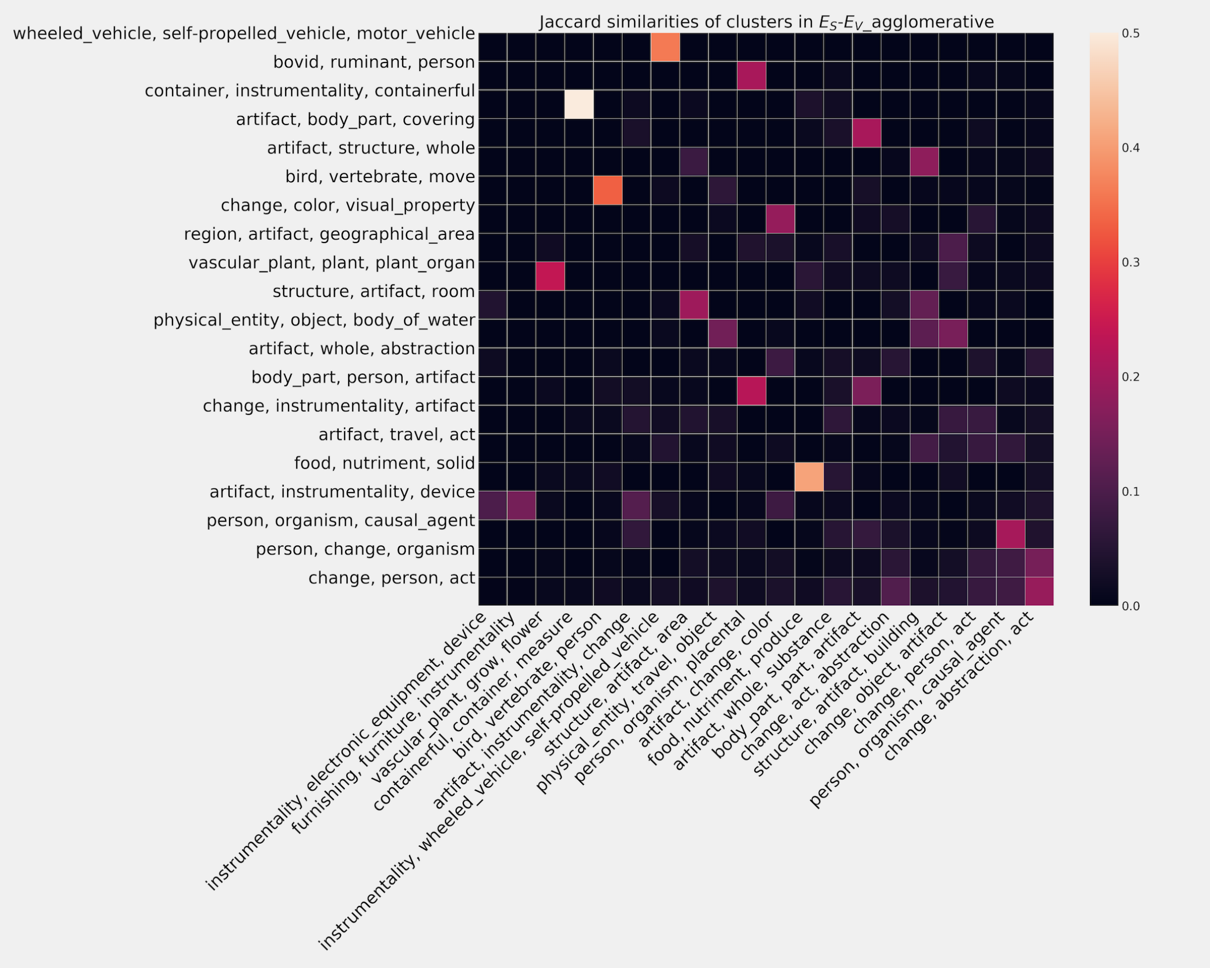


Figure D.5: Heatmap of Jaccard coefficients between Agglomerative clusters of $E_S$ and $E_V$ ($y$ and $x$ axes respectively).

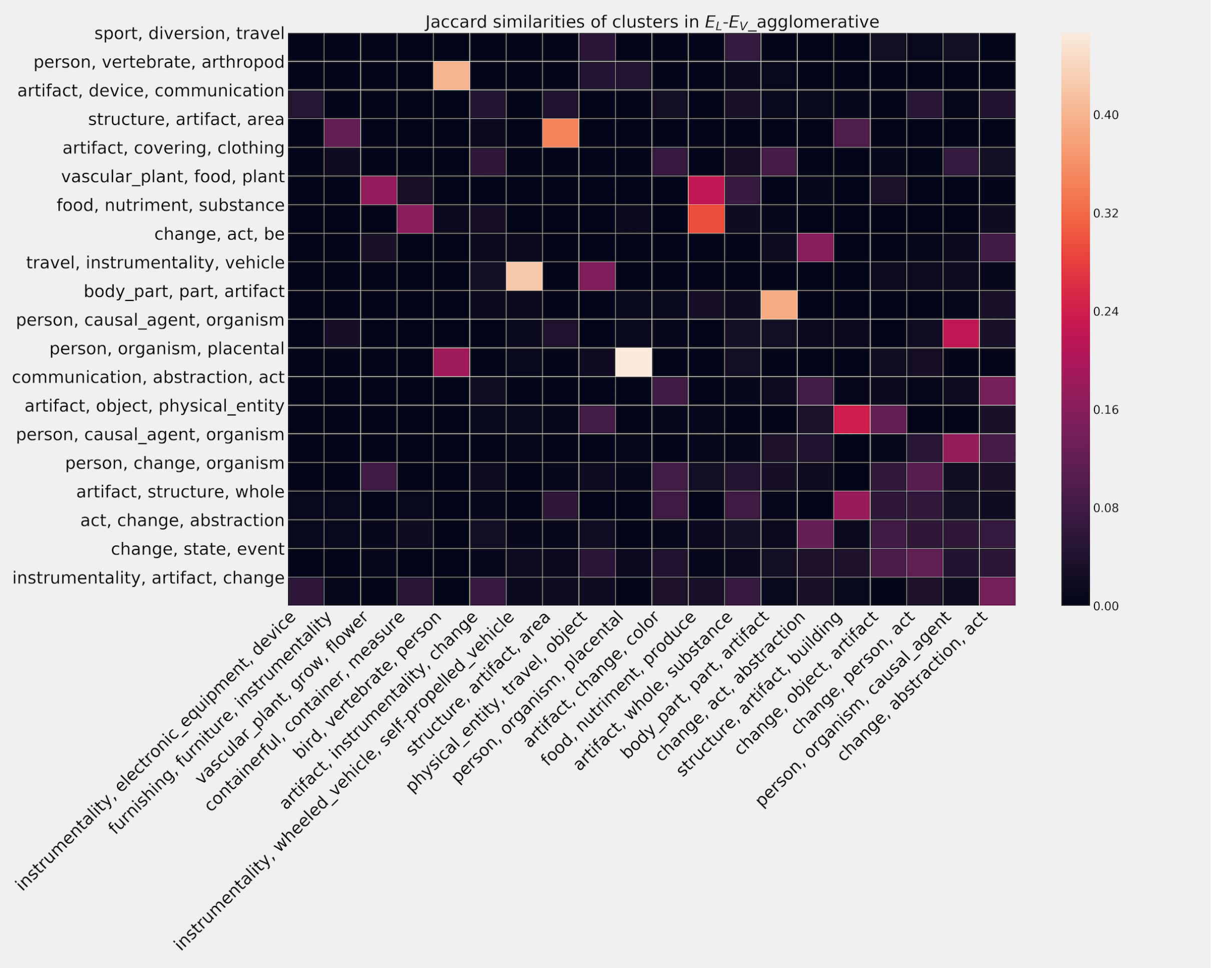


Figure D.6: Heatmap of Jaccard coefficients between Agglomerative clusters of $E_L$ and $E_V$ ($y$ and $x$ axes respectively).

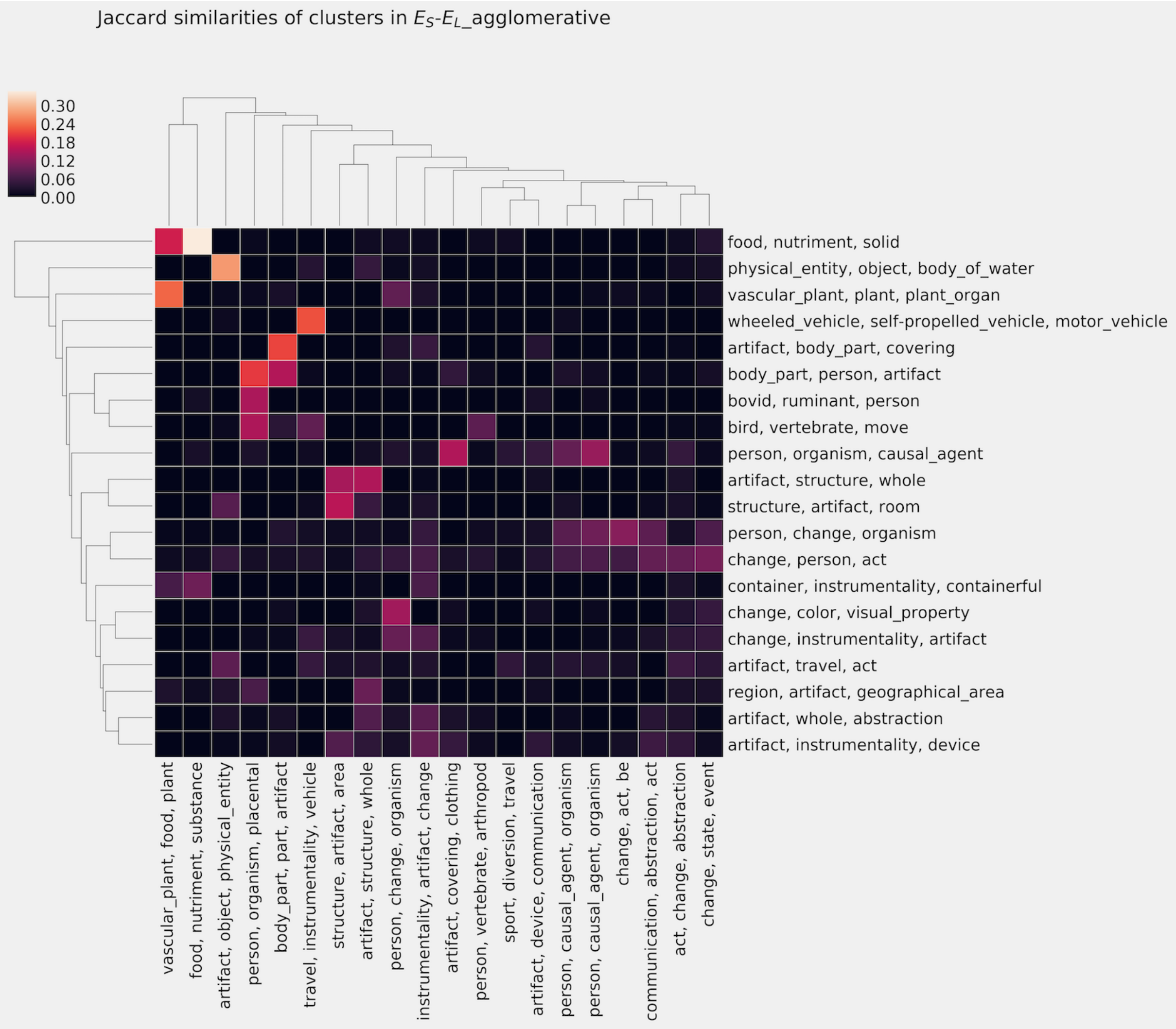


Figure D.7: Cluster map of Jaccard coefficients between Agglomerative clusters of $E_S$ and $E_L$ ($y$ and $x$ axes respectively).

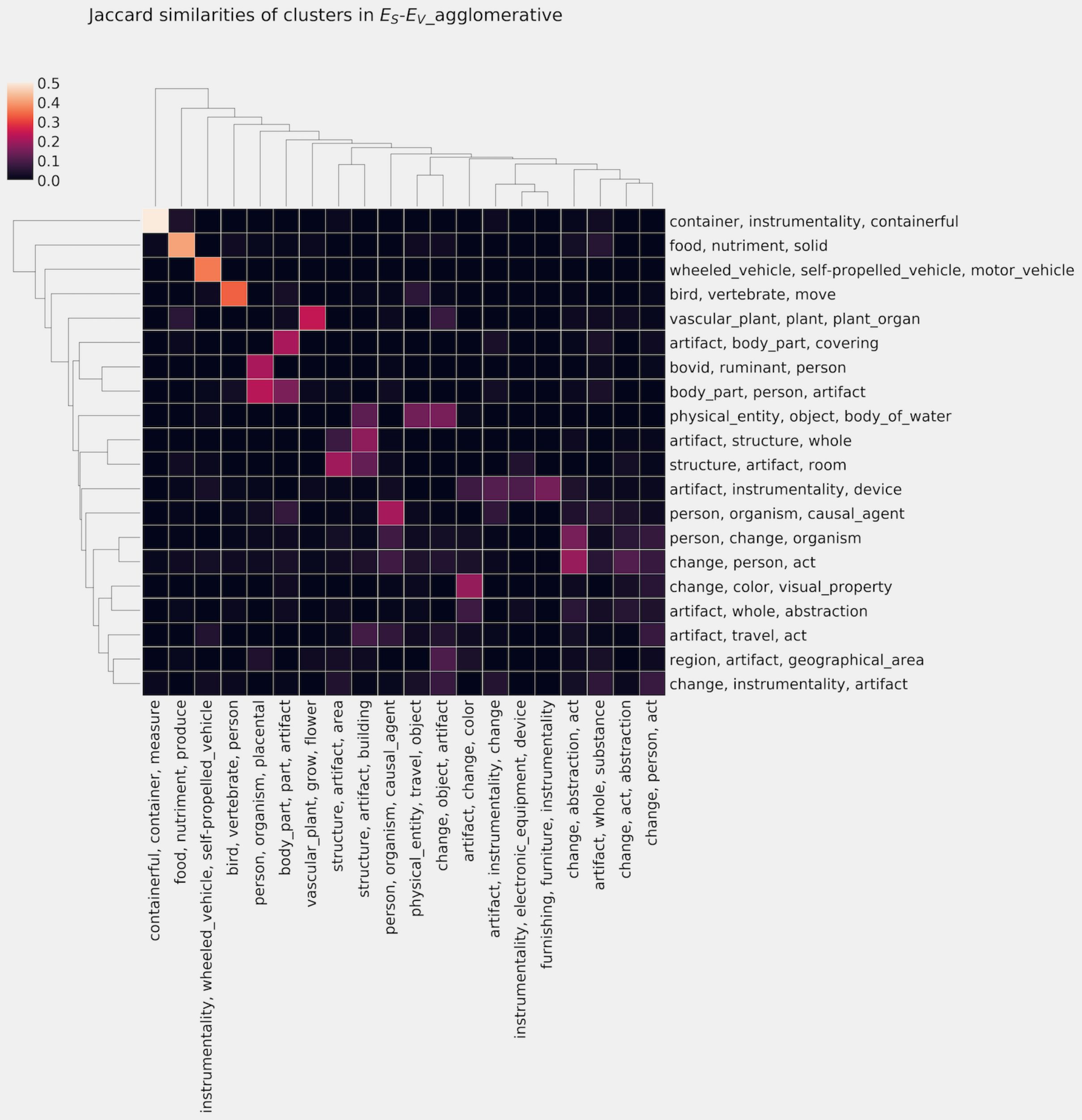


Figure D.8: Cluster map of Jaccard coefficients between Agglomerative clusters of $E_S$ and $E_V$ ($y$ and $x$ axes respectively).

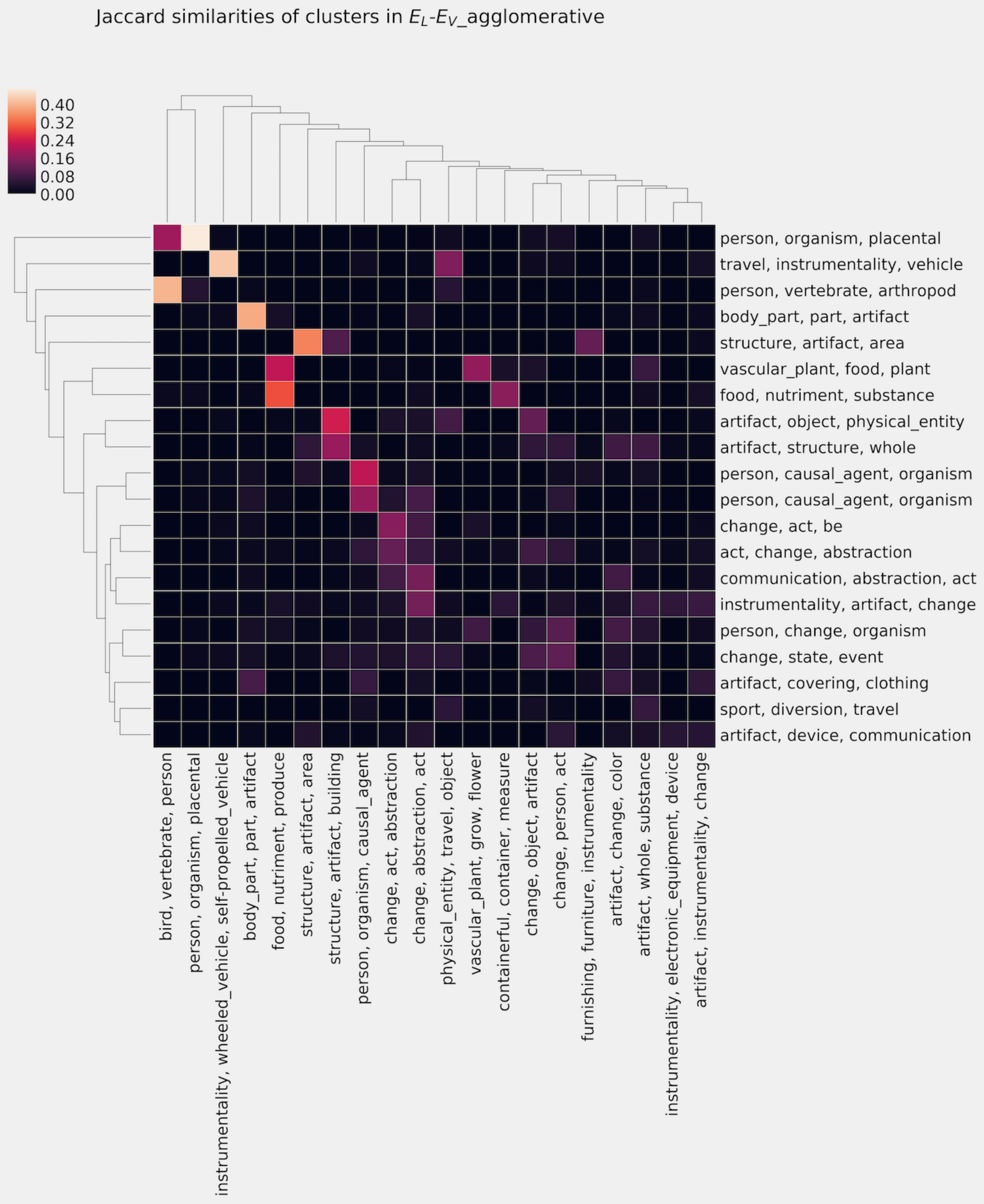


Figure D.9: Cluster map of Jaccard coefficients between Agglomerative clusters of $E_L$ and $E_V$ ($y$ and $x$ axes respectively).

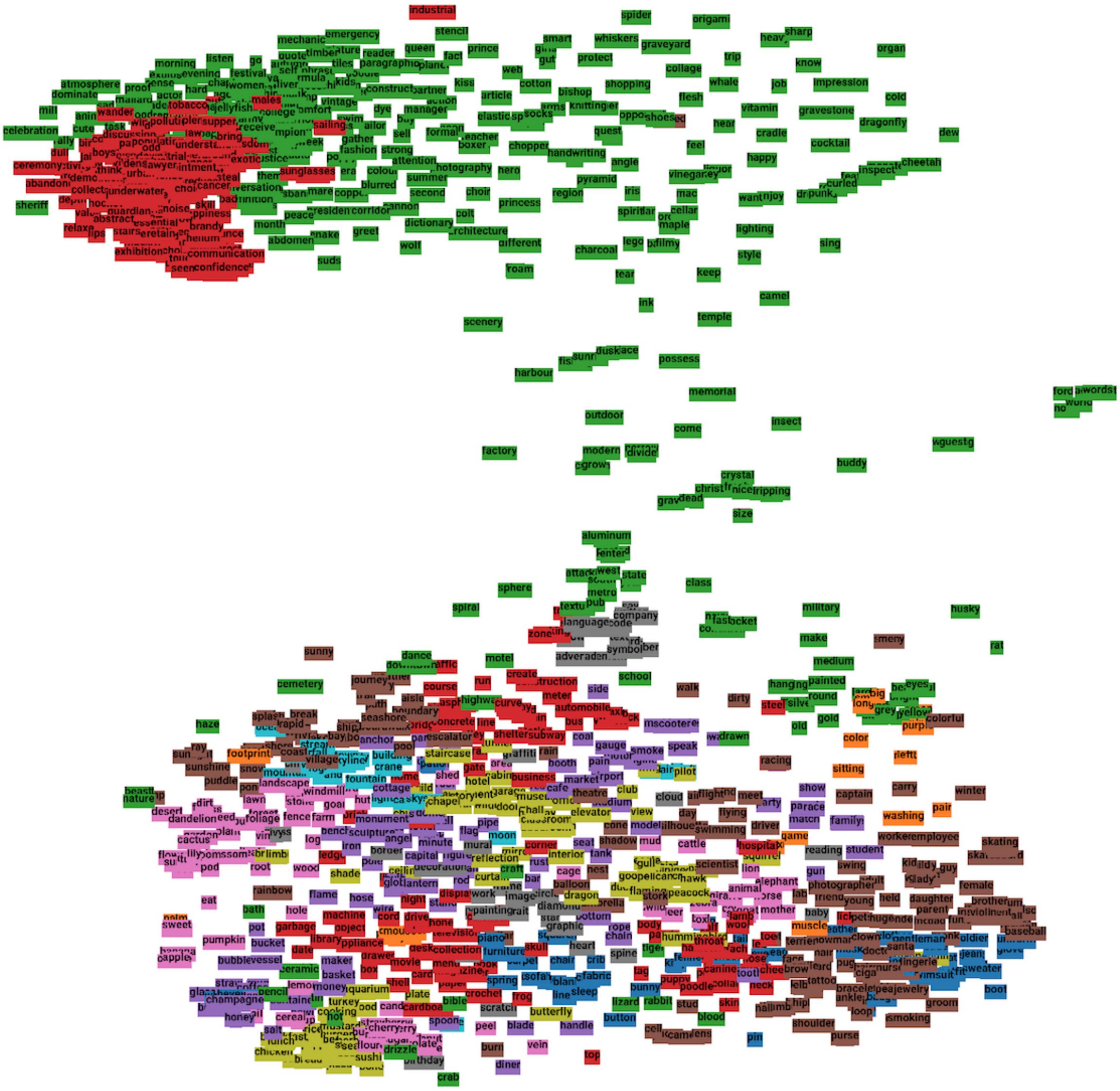
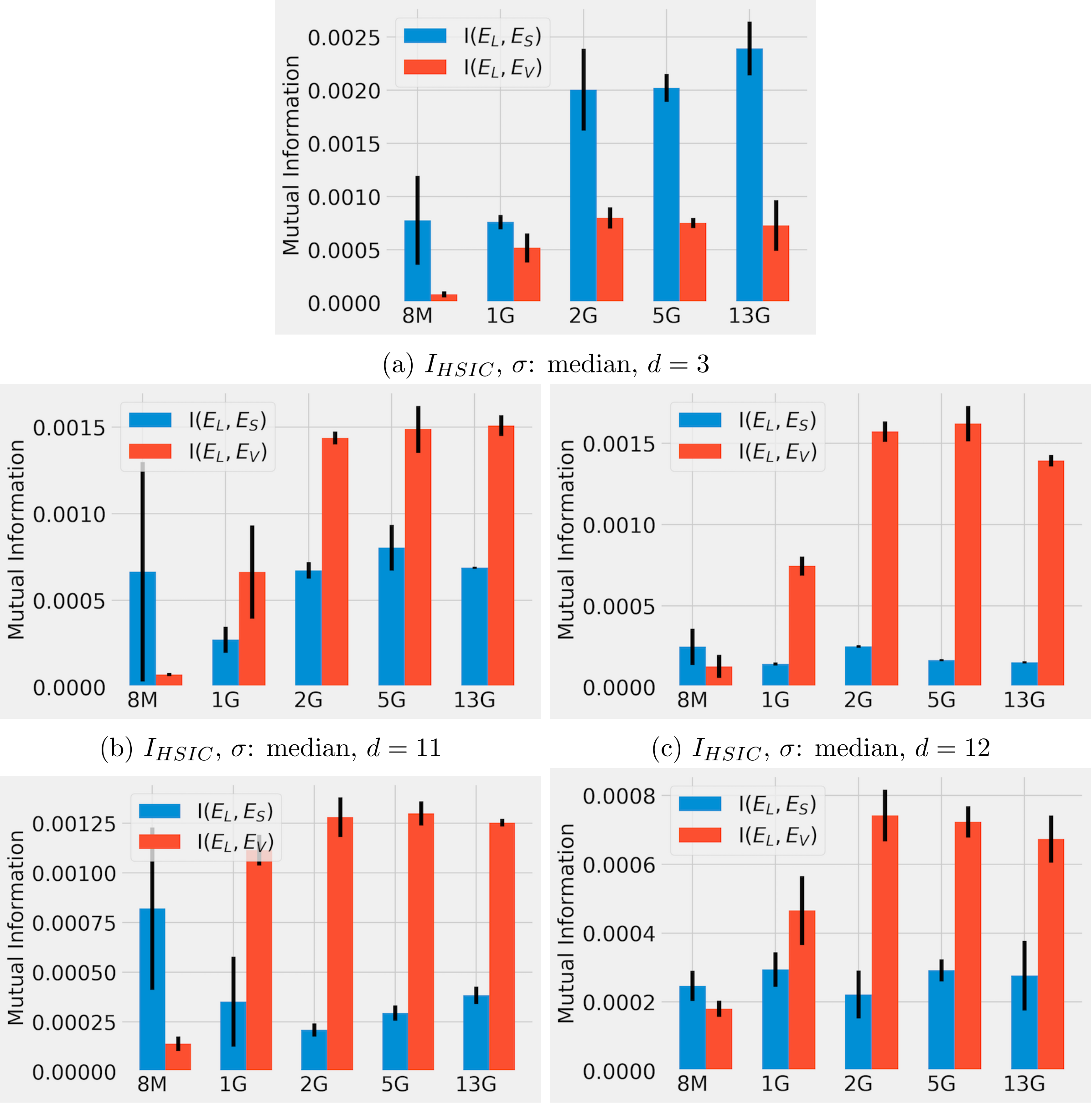


Figure D.10: T-SNE plot of $E_S$ with 40 cluster labels obtained by K-means clustering. TSNE perplexity = 52.

# Appendix E

# Mutual Information of Semantic Spaces



(a) $I_{HSIC}$, $\sigma$: median, $d = 3$

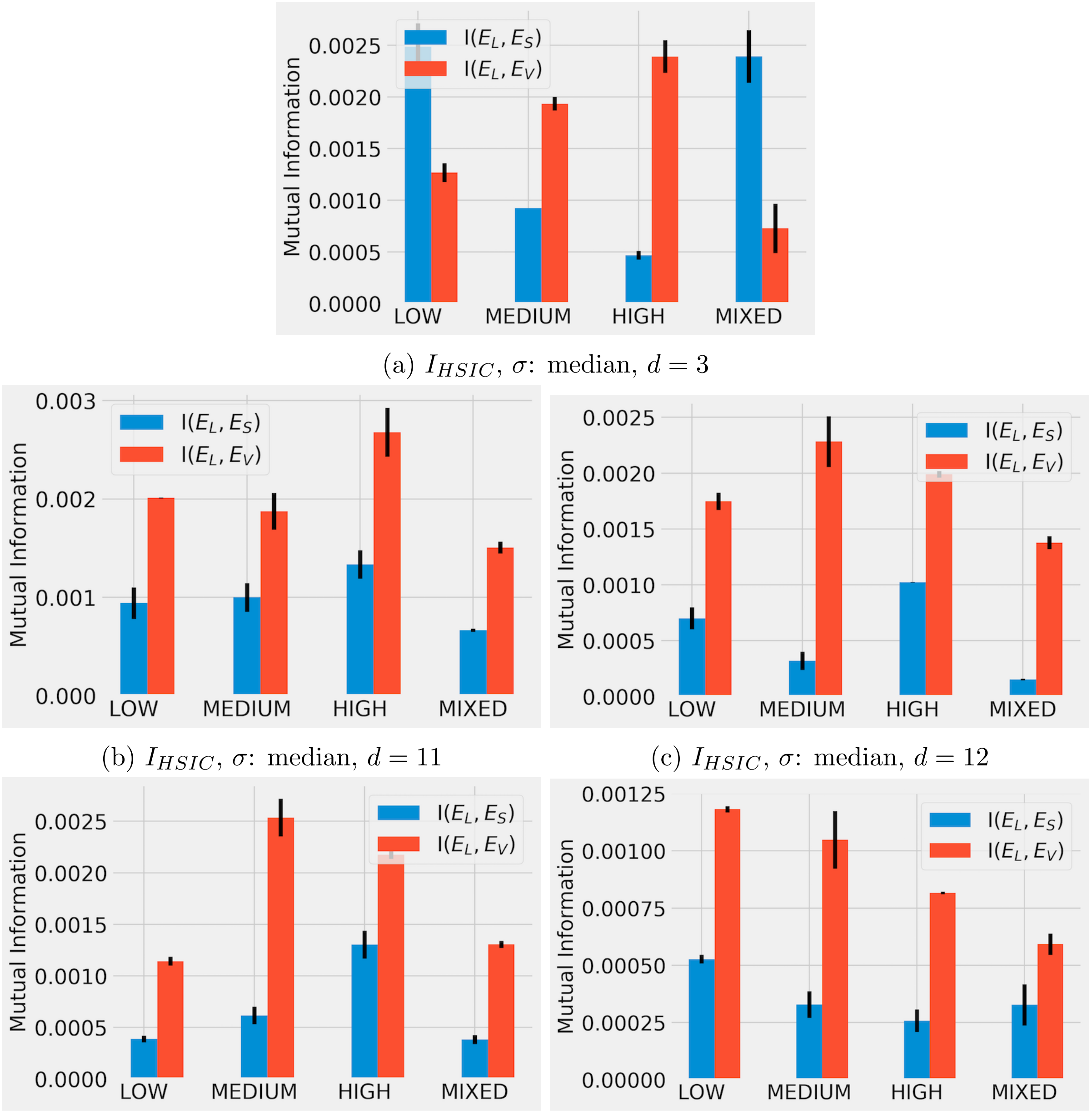
(b) $I_{HSIC}$, $\sigma$: median, $d = 11$

(c) $I_{HSIC}$, $\sigma$: median, $d = 12$

(d) $I_{HSIC}$, $\sigma$: median, $d = 13$

(e) $I_{HSIC}$, $\sigma$: median, $d = 50$

Figure E.1: Estimated Mutual Informations: $I(E_L, E_V)$ (red) and $I(E_L, E_S)$ (blue) for different corpus sizes.



(a) $I_{HSIC}$, $\sigma$: median, $d = 3$

(b) $I_{HSIC}$, $\sigma$: median, $d = 11$

(c) $I_{HSIC}$, $\sigma$: median, $d = 12$

(d) $I_{HSIC}$, $\sigma$: median, $d = 13$

(e) $I_{HSIC}$, $\sigma$: median, $d = 50$

Figure E.2: Estimated Mutual Informations: $I(E_L, E_V)$ (red) and $I(E_L, E_S)$ (blue) for different word frequency ranges.

# Appendix F

# Centroid Contexts

| Centroid | Wikipedia | VG |
|---|---|---|
| plate | tectonics, nazca, restrictor, farallon, subducts, license, cribriform, tectonic, subducting, eurasian | plate, lying_on_top_of, on, has, on_top_of, in |
| licking | dipmix, upsumida, “eager, mauwe, caṭa, glucorticoid, nameru, schlecken, audhumla, borametz | licking, cealing, tougue, tongue, girrafe, lioness |
| ad | hoc, din, libbed, valorem, hominem, libs, alcorcón, litem, infinitum, libbing | ad, of_different, re_an_ad, features_view, templeton, lining |
| rust | epique, cronartium, oleum, cohle, obritzberg, blister, belt, puccinia, windexed, colored | rust, stains_down, around_side_of, rusted_onto, on_fire, with_a_lot |
| railway | station, line, midland, bnsf, nearest, gauge, junction, western, rhaetian, stations | railway, detach, elevated_on_platform_over, passes_over_a, spliced_through, traditional |
| classroom | teachers, instruction, collaborize, embiggening, 이데아, space, 257327273628, borrowred, classbank, approach—she | classroom, discussing_in, standing_inside, sitting_inside, student, attending |
| hummingbird | amazilia, selasphorus, mellisuga, calypte, cynanthus, berylline, scintillant, orthorhyncus, eupherusa, chinned | hummingbird, eat_nectar_from, in_flight_below, flapping_its, flapping, windspan |
| cab | cutie, calloway, hansom, obradoiro, driver, taxi, muzen, signalling, abstracts, susbde | cab, on_hood_of, paintin, rding, back_window_of_taxi, driving_side_of |
| bloom | dooryard, algal, slieve, jazmine, harold, vlastnik, flowers, asmuth, twillerbuds, orlando | bloom, cherry_blossom_tree_in, in_full_summer, rose_has_fully, buttercup, at_a |
| chapel | sistine, hill, ease, carolina, methodist, calvary, chantry, unc, mortuary, brancacci | chapel, church, outside, home, trim, to |
| champagne | ardenne, aishihik, bottle, stakes, lanson, châlons, noëlla, beaugrand, crayeuse, 1172–1219 | champagne, iin, in_a_womans, carafe, cling_to, wrapped_around |
| jaw | moose, droppingly, lower, phossy, osteonecrosis, upper, dropping, ćehu’pa, palinal, wakamow | jaw, strong, searching, facing_opposite_of, of, wearing_no |
| rapid | transit, interborough, growth, wien, prototyping, intensification, bucureṣti, expansion, bus, industrialization | rapid, powering_through, maneuver_through, rides_of, crashing_over, goes_over |
| dandelion | taraxacum, burdock, c4h4n2os, cicòria, insubrico, vliegezwam, zangune, freelya, oduvanchik, paardebloem | dandelion, in_empty_spot_of, against_some, are_distributed_in, among, growing_in |
| bright | yellow, colors, red, orange, lights, colours, kellie, spots, sunshine, green | bright, light_green_ten, lit_red_arrow_pointing, sky_cloudy_but, street_light, tv_screen, yellow_painted_wood |
| drizzle | 毛毛雨, drazzle, chispear, babucha, ourdelta, , miggelen, dampy, fzdz | drizzle, a_donut, decorated_same_as, adorning, ont_he, topping |
| crochet | yarntainers, asperó, kolose, freeformations, networks—signaled, knitting, filet, tatting, marandati | crochet, clit, needlepoint, are_for, sitting_o_top_of, yarn |
| fun | poked, poking, pokes, poke, loving, lot, lovin, yidishn, fun, weäsell | are_having, are_having_great, fun, facing_away, planning, having |
| pain | neuropathic, abdominal, chronic, excruciating, chest, velayat, orofacial, myofascial, swelling, khiyaban | choppy_near, waving_from, sitting_back, sticking_up_out_of, travel_on |
| ant | dec, formicid, anoplolepis, arboreal, blue-black, etkenmen, solenopsis, leafcutter, genus | ant, showing_through, reflecting_off, are_behind, vase, are_on |

| | | |
|---|---|---|
| bus | routes, service, terminal, services, intercity, stops, stop, lines, shuttle, rapid | bus, on_front_of, on_side_of, in, wearing, lit_up_on |
| giraffe | giraffa, reticulated, masai, gogolick, erdmännchen, four–horned, geraneous, giraffae, mtundu, qəribə | giraffe, has, of, on, wearing, man |
| glass | stained, windows, magnifying, borosilicate, leaded, ionomer, window, menagerie, panes, beads | glass, wine, wearing, on, liquid, half_full_of |
| hand | right, sleight, grenades, left, hand, cranked, grenade, claps, gloved, upper | hand, holding, held_in, on, in_mans, man |
| window | transfer, openings, transom, glass, palladian, oriel, lancet, sills, sash, panes | window, built_into, wearing, man, on_side_of, on |
| plane | crash, projective, focal, euclidean, crashed, hyperbolic, inclined, astral, transpyloric, crashes | plane, flying_in, wearing, man, on_side_of, of |
| white | sox, black, supremacist, house, 0, creamy, collar, supremacists, tailed, stripes | white, colored, grouped, rustic, tile_on_a, blossem |
| grass | marram, roots, splendour, günter, molinia, tussock, rosmalen, yelloweyed, cogon, tussocks | grass, eating, grazing_on, standing_in, has, grazing_in |
| tree | trunks, banyan, noisily, bodhi, christmas, beerbohm, fig, porcupine, frog, lined | tree, growing_on, behind, leave, on, man |
| room | dining, locker, dressing, waiting, billiard, temperature, schoolhouse, boiler, hotel, romper | room, in, in_other, on, in_corner_of, smiling_in |
| water | polo, drinking, supply, potable, fresh, sanitation, brackish, vapor, shallow, soluble | water, in, swimming_in, on, floating_in, wading_through |
| wall | street, hadrian, curtain, dodd–frank, antonine, retaining, hangings, berlin, qibla, cell | wall, hanging_on, against, has, hung_on, man |
| dog | bonzo, naughty, hound, sled, mad, whelks, junkyard, stray, shaggy | dog, on, man, in, chasing, has |
| sky | perfectv, madrean, big, gocheok, sports, tg24, blue, coast—sea, news | in, sky, hanging_in, flying_in, cloud, on |
| train | kmph, passenger, wagon, halts, derailed, station, freight, express, southbound, services | train, on_front_of, in, wearing, man, waiting_for |
| table | tennis, lists, periodic, shows, following, summarises, sortable, summarizes, lookup, hash | table, sitting_at, on_top_of, has, in, at |
| man | spider, isle, young, mega, yōshū, beenie, iron, tt, old, pac | wearing, on, man, in, wears, holding |
| living | couples, someone, alone, together, 18, families, poverty, daylights, quarters, people | living, island, docked_near, plant, horizon, top |
| say | goodbye, needless, anything, yes, goodbyes, darndest, goes, sources, went, hello | say, atna, hilton, adopt, allerton, easden, moda |
| seagull | 60033, chekhov, fluoxytine, , arcadina, treplev, en160, hamatam, kaarrku, merikajakas | seagull, floating_with, over_and_in, beside_of_a, has_webbed, about_to_dive_in |
| furniture | upholstered, antique, fittings, store, homestores, designer, maker, widdicomb, depository, harrods | furniture, sniffing_under, occupying, has_a_reflection_in, has_shadows, lid_hidden_by, matched_with |
| inspect | cerny2012, ensenade, evasion—arrives, maniktolla, shutter—the, ofstin, sokotí, ×supervises, goods..., skagerrak | inspect, zebra, hoof, branch, mane, train |
| accident | car, automobile, fatal, motorcycle, occurred, boating, freak, traffic, tragic, investigation | accident, scurrying_around, arrived_at_an, was_in_an, whit, wearing_one |
| woman | wonder, young, suffrage, first, pregnant, bionic, named, elderly, man, beautiful | woman, wearing, on, holding, man, in |

| | | |
|---|---|---|
| object | oriented, neptunian, inanimate, limerent, thorne–żytkow, substellar, xmlhttprequest, permanence, datasource, remotable | object, on_bottom_of_an, on_bottom_of_a, supporting_a, in_front_of_an, in_to |
| weed | noxious, forestin, jimson, thurlow, whacker, invasive, cabomba, sot, teyck, forrestin | weed, are_growing_in, obscures, growing_through, growing_in, grow_out_of |
| sushi | sashimi, nigiri, nudoki, tempura, benkay, cremiere, dishes—list, funtoast, itacho | sushi, rolls, avacado, fishes, are_partially, are_below |
| texture | leathery, velvety, crunchy, porphyritic, mapping, crumbly, shaders, chewy, papery, waxy | texture, angle_creates, ha_different, a_clear, tells, casts_on |
| gut | microbiota, wrenching, microbiome, volkseigenes, riechst, stjepko, aiderbichl, alkalay, wrenchingly, byeolsin | gut, stringy, pumpkin, wering, all, this |
| museum | art, stedelijk, ashmolean, kunsthistorisches, hirshhorn, metropolitan, whitney, guggenheim, boijmans | parked_outside_of, museum, mounted_to_side_of, of_airplane, parked_outside, parked_in_front_of |
| compare | cmurek, d'anello, itimatres, kolskys, quoteboxes, yadhaikenu, disposition—to, favorably, bubenspitzle, hetairistria | compare, kid, phone |
| island | rhode, staten, coney, long, baffin, whidbey, canvey, mackinac, rikers, vancouver | island, living, rimmed_in, on_front_part_of, opened_next_to, building_into |
| theme | song, ending, opening, tune, anatolic, recurring, armeniac, park, thracesian, parks | theme, gorilla, opens_on, depicting, stands_at, spatula |
| animals | cruelty, plants, furry, wild, domesticated, nonhuman, impaling, humans, stuffed | animals, stuffed, cubicle, insdie, sit, brown |
| think | tank, don't, tanks, "i, shudder, it's, said, saying, dance, bozos | think, grey, trunk |
| monk | thelonious, buddhist, benedictine, tonsured, bretton, fryston, xuanzang, huifan, 李欽, canatella | monk, grows_sparsely_near, helping_a, navigating, have_on, monastery |
| village | administrative, municipality, administrated, howmeh, population, greenwich, locality, makeup, pomerania, kielce | village, at_foot_of_a, overlooks, traveling_through, traveling, harbour |
| learn | shocked, scikit, surprised, ready, perfyct, horrified, students, basics, opportunity, dismayed | learn, ski, child, for, book, on |
| activity | volcanic, thunderstorm, economic, enzymatic, sexual, seismic, progestogenic, extravehicular, fumarolic, estrogenic | activity, attending, 3, watching, around, spectator |
| phone | mobile, call, calls, cell, hacking, windows, pals, cellular, phreaks, booth | phone, talking_on, holding, using, looking_at, talking_on_a |
| drawer | wbstartup, desk, doorbands, caryad, cigareet, gräffs, h9a00, jhangeer, panphobic, trinchante | drawer, face_in, built_underneath, dresser, handle, knob |
| poppy | opium, papaver, jhakra, eschscholzia, delevingne, breadseed, sovia, pipopapo, seeds, pricklyhead | poppy, cluster_of_pink, stamen, iris, lapel, gentleman |
| wine | sparkling, grape, tasting, cellar, cellars, mulled, grapes, jancis, tastings, bread | wine, pouring, glass, drinking, tasting, half_filled |
| swan | whooper, nautor, serinda, silvertones, leda, odile, tuonela, coscoroba, lake, pukā | swan, swimming_above_a, swimming_on, swimming_upon, swimming_in, wading_on |
| blanket | nonpartisan, ejecta, bog, statementstein, uasb, bingo, primary, blackfriargate, 200–600mm, asrolig | blanket, sewn_into, adorning, color_in, draped_over, wrapped_up_in |
| bike | mountain, racks, lanes, dockless, trails, paths, path, citi, orienteer, ride | bike, riding, locked_to, in, riding_a, chained_to |
| beach | palm, myrtle, delray, vero, daytona, pompano, redondo, pebble, volleyball, long | beach, at, walking_on, breaking_on, playing_on, standing_on |

| | | |
|---|---|---|
| calf | roping, fatted, cow, golden, astovl, injury, savatsa, eday, vitlo, milkcow | calf, nursing_from, nursing_in, nursing, has_head_on, and_street_with |
| yellow | jackets, fever, pale, greenish, bright, perch, bellied, flowers, orange, peril | yellow, green_face_of, colored_sauce_on, track_has, yellow_ground, colored_doughnut |
| cooking | utensils, pots, cleaning, vinyl, pot, ‘koken, oil, dotch, baking, sewing | cooking, grill, pan, hotdog, oil, steak |
| horn | hardart, cape, rimmed, kimley, kaniehtiio, blinkey, africa, tiio, flugel, trevor | horn, curving_away_from, horns_on_giraffe, has, has_curved, grow_from |
| nail | tooth, biter, biting, çakırhan, mazitovich, yakupov, coffin, polishes, bitingly, salons | nail, used_in, impaled_on, holding_together, hung_up, has_thumb |
| buddy | holly, ebsen, desylva, defranco, lazier, landel, valastro, stoodios, cianci, roemer | buddy, best, animals_in, three, stands_by, stand |
| haze | daizee, evot, transboundary, purple, dingleberry, bagilgul, defringe, hameshumash, raykeea, unidentifiable–uniform | haze, over_top, are_higher_then, are_very_far_from, blob, in_distance |
| protest | peaceful, resigned, marches, liulitun, mechaat, protestiram, rosenstrasse, strike, nonviolent, izik | protest, for, sign, road, holding, on |
| sleep | apnea, rem, deprivation, nrem, obstructive, dreamless, polyphasic, apnoea, disorders, bruxism | sleep, with_a_blue, with_green_shirt, covered_with_a, has_his_head, shoes_are_on |
| clothes | swaddling, oxxford, washing, civilian, dryers, plain, borlo, dryer, amarilly, nagchaumpa | clothes, on_clos, stores, have_pockets_in, are_piled_on_top_of, strown_on |
| butter | peanut, margarine, cocoa, bread, clarified, cheese, melted, jelly, unsalted, murumuru | butter, large_near, partly_outside, sauteeing_in, margarine_on, in_a_small |
| flower | buds, heads, antlike, travellin, spikes, pasque, lotus, anthos, beds, stalks | flower, vase, blooming_in, blooming_inside, in, in_mid_of |
| rain | torrential, singin, heavy, pouring, soaked, snowfall, snow, hatful, shine, freezing | rain, falling_on, getting_wet_by, towns, walking_on, taking_shelter_from |
| coffee | shop, roasters, beans, shops, plantations, arabica, decaffeinated, starbucks, roaster, tea | coffee, cup, mug, poured_from, of_steaming, blowing_across |
| cow | clarabelle, dung, milk, calf, hydrodamalis, parsnip, hocked, reined, mad, kowemerk | cow, on, standing_in, laying_in, milking, being_shown_at |
| wig | wam, jagbags, wamania, blonde, wag, sheinhardt, blond, apologise—it, hairpiece—a, lacefront | wig, in_clown, wears, guiding, not_wearing_a, pulled_by |
| tower | conning, eiffel, bell, clock, hamlets, martello, babel, spasskaya, druaga, spire | tower, housing_a, along_front_of, at_top_of, clock, containing_a |
| blue | jays, ribbon, öyster, jackets, heelers, collar, devils, riband, bombers, ridge | blue, clear, and_white_ocean, cloudless, air_on, brand_name_rusty_on_its |
| skin | irritation, rashes, grafts, lesions, pigmentation, irritations, goldbeater, mucous, nonmelanoma, tanned | skin, hanging_h, lady_light, appearing_on, mating_with, on_cat |
| flexible | compacting, sigmoidoscopy, fiberoptic, bendsome, conformationally, adaptable, plankostenrechnung, liquidtight, linkers, heliac | flexible, green_rim, to_catch_a, gren, in_dogs, frisbee |
| bag | duffel, plastic, duffle, bidita, avinabo, punching, sleeping, grab, colostomy, airsickness | bag, carrying, carrying_a, carries, beyween, placed_inside |
| bird | passerine, migratory, caged, sanctuary, watchers, watching, topley, species, prey, furnariidae | bird, perched_on, flying_in, flying_over, beak, flying_ahead_of |
| kitchen | sink, soup, hell, utensils, dining, hell’s, bathroom, pantry, scullery, laundry | kitchen, in, face_in, prepared_in, working_in, with_interior |
| father | succeeded, died, footsteps, adoptive, biological, death, law, inherited, son | father, leaning_over_to_touch, taking_a_picture_in, taking_a_self_in, and_son_learning_to_ice, cleans_sons |

| | | |
|---|---|---|
| shore | dinah, lake, north, eastern, batteries, geordie, bombardment, pauly, jersey | shore, breaking_on, washing_on, coming_to, coming_in_to, crashing_on |
| vehicle | motor, launch, registration, electric, mav, wheeled, reentry, rov, utility, uav | vehicle, parked_alongside_of, parked_alongside, parked_on, parking_on_side_of, are_parked_alongside_of |
| bring | back, helped, together, would, horizon, forth, could, attention, able, attempt | bring, says, wall, on_a, has_a, has |
| smile | smiley, carnt, vojdanov, スマイル, « », solami, toothy, dk, sonríe, face | smile, exposing, revealing, exposes, in_smiling, expressing |
| time | first, full, long, extra, spend, spent, real, consuming, slot, around | time, having_great, counts, scene_during_day, to_tell, tells |
| fact | despite, due, checking, spite, finding, complicated, matter, checker, compounded, evidenced | fact, listed_on, with_some, ingredient, jug, bottle |
| football | league, team, club, college, player, victorian, national, american, professional, coach | football, chases, to_hinder, trying_to_save, player_in, experiencing_a |
| fish | finned, cyprinid, wildlife, hatchery, bony, mardy, freshwater, anadromous, demersal, shellfish | fish, mole_on_a, writhes_inside, has_a_gray, caught_with, about_to_be_fed_to |
| film | festival, directed, cannes, drama, feature, comedy, sundance, documentary, horror, thriller | film, taken_with, are_being, taped_to, being, playing_on_a |
| arms | coat, coats, embargo, municipality's, ammunition, legs, pursuivant, canting, small, gules | arms, outstretched, bare, fold, skater, skateboarder |
| plant | flowering, power, pathogen, desalination, figwort, ornamental, host, hardiness, herbaceous, bottling | plant, growing_in, growing_on, growing_up_a, pot, leaf |
| food | fast, beverage, drug, drink, shortages, supplies, insecurity, neivethanam, agri, staple | food, added_on, sauteing_in, wrapped_inside, placed, cut_in_to |
| make | sure, amends, way, failed, would, easier, decisions, wanted, room, sense | make, exchange, fishpond, blower, construct, are_not |
| army | salvation, liberation, u, british, states, corps, potomac, red, kwantung, officer | patch_for, army, getting_out_of, calendar, on_back_of, green |
| body | governing, whorl, student, snatchers, sanctioning, politic, ecliptic, cremated, human, lifeless | body, blown_up_in, boat_in, drifting_in, away_from_thier, bent_away_from |
| school | high, elementary, secondary, grammar, district, primary, boarding, middle, preparatory, districts | school, bus_that, front, second, stenciled_on, yellow |
| forest | nottingham, wake, teutoburg, epping, boreal, montane, sclerophyll, jarrah, lawn, deciduous | forest, road_in, tipped_over_in, stand_above, form_a_distict_line, a_ground_in |
| new | york, zealand, jersey, orleans, hampshire, guinea, papua, brunswick, yorker, testament | new, urban, beard_and, generation_wide_screen_smart, model_white, aspire |
| city | york, kansas, council, makeup, mexico, limits, quezon, oklahoma, holby, residing | city, shining_in, building_in_a, in, wandering, in_asian |
| people | surname, notable, 000, per, republic, young, indigenous, disabilities, employed, aboriginal | people, are_enjoying, has, on, watching, are_watching |
| family | moth, cerambycidae, mollusk, beetle, crambidae, geometridae, erebidae, noctuidae, tortricidae, size | family, seated_around, stand_with, clearly, where_are, on_ground_for |
| house | representatives, commons, lords, manor, opera, white, publishing, delegates, speaker, random | house, on_faade_of, cemented_on, in_front_of, crossing_over_to, adorning |
| year | old, following, contract, olds, every, per, fiscal, rookie, next, previous | year, pub, 1893, states, age, was_taken_in |

| | | |
|---|---|---|
| party | communist, democratic, labour, liberal, conservative, janata, socialist, republican, political, labor | party, at_a_birthday, crying_at, sneaking_up_on, having, are_at_a |
| company | parent, brewing, insurance, worshipful, steamship, manufacturing, publishing, holding, production, founded | company, of_photography, are_enjoying_each_others, calls, boeing, that_owns |

Table F.1: Context words of cluster centroids with the 10 highest $\chi^2$ score.

| **Centroid** | **Wikipedia** | **Visual Genome** |
|---|---|---|
| plate | tectonics, license, river, nazca, restrictor, tectonic, umpire, eurasian, farallon, home | on, plate, on_top_of, on_a, lying_on_top_of, with |
| licking | county, river, mauwe, grooming, fork, cața, lips, wounds, dipmix, upsumida | licking, tongue, giraffe, cat, girrafe, cealing |
| ad | hoc, century, din, libbed, valorem, libs, alcorcón, lib, hominem, infinitum | ad, on, of_different, lining, for, has |
| rust | belt, colored, fungi, blister, cronartium, epique, coloured, fungus, oleum, cohle | rust, on, has, stains_down, on_fire, around_side_of |
| railway | station, line, western, midland, nearest, stations, company, junction, gauge, eastern | railway, detach, traditional, beside, elevated_on_platform_over, passes_over_a |
| classroom | teachers, instruction, space, language, building, outside, future, 0, every | classroom, in, standing_inside, sitting_inside, discussing_in, student |
| hummingbird | amazilia, selasphorus, throated, mellisuga, chinned, calypte, cynanthus, hawkmoth, species, scintillant | hummingbird, flapping, eat_nectar_from, flapping_its, has, in_flight_below |
| cab | cutie, calloway, hansom, driver, taxi, obradoiro, signalling, death, abstracts | cab, on, has, on_hood_of, driving_on, of |
| bloom | algal, harold, flowers, slieve, dooryard, orlando, jazmine, claire, leopold, full | bloom, cherry_blossom_tree_in, in_full_summer, rose_has_fully, on, in |
| chapel | hill, sistine, carolina, ease, methodist, built, dedicated, calvary, st, chantry | chapel, on, outside, church, of, to |
| champagne | ardenne, stakes, bottle, aishihik, en, châlons, brie, lanson, reims | champagne, iin, in_a_womans, carafe, glass, wrapped_around |
| jaw | moose, lower, upper, droppingly, broken, dropping, phossy, osteonecrosis, muscles, saskatchewan | jaw, of, has, of_a, strong, on |
| rapid | transit, growth, wien, expansion, bus, interborough, intensification, prototyping, succession, bucurești | rapid, powering_through, crashing_over, goes_over, maneuver_through, rides_of |
| dandelion | taraxacum, burdock, tribe, wine, superfine, plants, dxc, barleycup, geralt, krindle | dandelion, in, growing_in, in_empty_spot_of, among, against_some |
| bright | yellow, red, colors, orange, lights, green, colours, light, eyes, blue | bright, sleep_on, blue, eyes, green, browñ |
| drizzle | 毛毛雨, drazzle, freezing, chispear, babucha, ourdelta, , miggelen, rain, gearman | drizzle, a_donut, adorning, ont_he, decorated_same_as, on |
| crochet | knitting, hook, filet, yarntainers, stitches, asperó, kolose, tatting, knit | crochet, are_for, needlepoint, clit, under, are_on |
| fun | poked, poking, pokes, poke, loving, lot, much, making, make, fun | fun, are_having, are_having_great, facing_away, having, planning |
| pain | abdominal, neuropathic, chronic, chest, suffering, excruciating, relief, swelling, back, severe | choppy_near, waving_from, sitting_back, sticking_up_out_of, travel_on |
| ant | dec, species, genus, subfamily, man, arboreal, adam, fire, tupolev | ant, showing_through, reflecting_off, are_behind, vase, are_on |
| bus | routes, service, services, terminal, station, stop, lines, stops, rapid, route | on, bus, has, on_side_of, on_front_of, of |
| giraffe | giraffa, reticulated, masai, rothschild, zebra, melman, amb, gogolick, nubian | has, giraffe, of, on, behind, has_a |
| glass | stained, windows, window, magnifying, looking, leaded, beads, menagerie, borosilicate, bottles | glass, on, wearing, has, in, with |
| hand | right, left, hand, grenades, one, sleight, side, upper, grenade, combat | hand, has, holding, in, on, of |
| window | transfer, glass, openings, transom, rear, lancet, palladian, frames, oriel, arched | on, window, has, build, on_a, on_side_of |

| | | |
|---|---|---|
| plane | crash, projective, focal, crashed, euclidean, inclined, hyperbolic, astral, crashes, complex | on, plane, has, of, on_side_of, flying_in |
| white | black, sox, house, 0, supremacist, collar, red, tailed, blue, creamy | white, on, colored, of, black, small |
| grass | roots, courts, marram, splendour, günter, outdoor, tussock, natural, perennial, tall | grass, on, in, eating, standing_in, has |
| tree | trunks, christmas, oak, banyan, lined, frog, planting, fig, joshua, palm | tree, behind, on, in, near, has |
| room | dining, locker, dressing, waiting, temperature, living, hotel, reading, make, drawing | room, in, has, inside_of, in_a, in_corner_of |
| water | polo, supply, drinking, fresh, potable, quality, shallow, census, sanitation, resources | in, water, near, on, swimming_in, by |
| wall | street, hadrian, curtain, berlin, retaining, cell, paintings, outer, brick, stone | wall, on, hanging_on, against, in, behind |
| dog | hound, naughty, mad, hot, sled, bonzo, pet, breeds, stray | dog, has, on, of, with, has_a |
| sky | big, sports, news, blue, night, perfectv, conference, madrean, survey | in, sky, cloud, flying_in, hanging_in, above |
| train | station, passenger, wagon, express, services, freight, kmph, service, derailed, halts | on, train, has, of, on_front_of, on_side_of |
| table | tennis, following, shows, lists, periodic, round, mid, summarizes, summarises, bottom | on, table, on_top_of, sitting_at, sitting_on, at |
| man | spider, isle, young, old, mega, iron, named, one, match, tag | wearing, man, has, on, holding, of |
| living | couples, someone, alone, together, 18, families, people, poverty, room, conditions | living, island, docked_near, plant, horizon, top |
| say | goodbye, needless, anything, went, sources, goes, yes, would, something, never | say, atna, letter, word, calmette, centre |
| seagull | chekhov, 60033, livingston, fluoxytine, , supermarine, treplev, zarechnaya, arcadina, chekhov's | seagull, over_and_in, floating_with, beside_of_a, flies_over, flying_above |
| furniture | store, antique, designer, maker, pieces, fittings, upholstered, factory, makers, design | furniture, on, occupying, sniffing_under, cusion, has_a_reflection_in |
| inspect | organise, skagerrak, proceedings, visually, damage, goods..., unpiggable, cerny2012, ensenade, evasion—arrives | inspect, zebra, hoof, branch, of, train |
| accident | car, automobile, fatal, motorcycle, occurred, traffic, investigation, boating, freak, cause | accident, arrived_at_an, scurrying_around, was_in_an, whit, say |
| woman | young, wonder, first, suffrage, named, man, old, pregnant, beautiful, american | wearing, woman, has, holding, on, wears |
| object | oriented, neptunian, subject, inanimate, indirect, direct, verb, relational, affection | object, on, on_bottom_of_an, in, on_bottom_of_a, has |
| weed | noxious, thurlow, jimson, invasive, forestin, whacker, control, sot, invasion, alligator | weed, growing_in, are_growing_in, in, growing_through, growing_next_to |
| sushi | sashimi, restaurant, nigiri, tempura, chef, bar, nudoki, yo, restaurants | sushi, on, rolls, avacado, near, are_below |
| texture | mapping, leathery, velvety, color, crunchy, waxy, shaders, papery, porphyritic, chewy | texture, angle_creates, ha_different, a_clear, tells, has |
| gut | microbiota, wrenching, microbiome, flora, stjepko, volkseigenes, strings, riechst, wrenchingly, microflora | gut, pumpkin, stringy, wering, all, has |
| museum | art, national, metropolitan, modern, natural, fine, british, whitney, history | parked_outside_of, museum, parked_in_front_of, parked_outside, mounted_to_side_of, of_airplane |
| compare | favorably, used, favourably, swap, difficult, contrast, different, tables, prices, results | compare, kid, phone |
| island | rhode, staten, long, coney, edward, vancouver, platform, mare, baffin, whidbey | island, living, rimmed_in, in, on, on_front_part_of |

| | | |
|---|---|---|
| theme | song, ending, opening, park, tune, recurring, main, parks, music, songs | theme, gorilla, depicting, opens_on, stands_at, has |
| animals | plants, wild, cruelty, humans, domesticated, furry, nonhuman, farm, domestic | animals, stuffed, cubicle, insdie, sit, brown |
| think | tank, tanks, don't, said, people, "i, saying, dance, it's, really | think, grey, trunk |
| monk | thelonious, buddhist, benedictine, tonsured, bretton, xuanzang, cistercian, mr, jain, meredith | monk, grows_sparsely_near, navigating, helping_a, wearing, holds |
| village | administrative, population, municipality, locality, small, makeup, greenwich, administrated, located, howmeh | village, overlooks, in, at_foot_of_a, traveling_through, traveling |
| learn | students, shocked, ready, surprised, opportunity, children, must, read, horrified, basics | learn, ski, child, for, book, on |
| activity | volcanic, economic, sexual, physical, thunderstorm, seismic, criminal, human, enzymatic, paranormal | activity, attending, 3, watching, around, in |
| phone | mobile, call, cell, calls, windows, hacking, number, numbers, cellular, booth | phone, holding, on, talking_on, has, using |
| drawer | desk, top, soccer, dresser, wbstartup, slides, crisper, painter, doorbands | drawer, on, in, has, handle, under |
| poppy | opium, papaver, seeds, seed, delevingne, jhakra, eschscholzia, cultivation, straw, remembrance | poppy, cluster_of_pink, stamen, iris, lapel, gentleman |
| wine | grape, sparkling, cellar, tasting, cellars, bread, grapes, red, region, spirits | wine, glass, pouring, of, in, drinking |
| swan | lake, river, black, whooper, nautor, leda, hunter, coastal, odile, districts | swan, swimming_in, makes, swimming_on, in, swimming_above_a |
| blanket | nonpartisan, ejecta, bog, primary, ban, wrapped, bingo, bogs, beach, uasb | blanket, on, has, bed, adorning, under |
| bike | mountain, lanes, racks, path, trails, paths, trail, ride, sharing, dirt | bike, on, riding, has, of, near |
| beach | palm, long, myrtle, florida, volleyball, daytona, boys, delray, miami, california | beach, on, at, walking_on, standing_on, playing_on |
| calf | roping, golden, injury, cow, fatted, muscle, strain, muscles, eday, astovl | calf, has, nursing_from, nursing, of, nursing_in |
| yellow | fever, jackets, pale, bright, orange, greenish, flowers, perch, card | yellow, on, painted, green_face_of, green, colored_sauce_on |
| cooking | utensils, oil, vinyl, pots, cleaning, pot, show, used, techniques, sewing | cooking, grill, pan, hotdog, pizza, food |
| horn | cape, africa, hardart, trevor, big, section, van, rimmed, golden, french | horn, has, on, of, giraffe, head |
| nail | tooth, biting, biter, coffin, polish, salon, records, salons, rusty, polishes | nail, on, has, of, has_a, used_in |
| buddy | holly, ebsen, guy, defranco, lazier, desylva, rich, roemer, collette, landel | buddy, best, animals_in, three, stand, in |
| haze | daizee, purple, transboundary, evot, dingleberry, smoke, booger, angel, jessika, fog | haze, over_top, are_higher_then, are_very_far_from, blob, below |
| protest | resigned, peaceful, movement, strike, marches, staged, rally, rallies, sit, violent | protest, for, sign, on, road, holding |
| sleep | apnea, rem, deprivation, nrem, obstructive, disorders, died, paralysis, dreamless, deep | sleep, with_a_blue, covered_with_a, with_green_shirt, shoes_are_on, wearing_a_blue |
| clothes | civilian, plain, washing, swaddling, wearing, wear, shoes, accessories, dryers, worn | clothes, wearing, on, on_clos, wears, stores |
| butter | peanut, bread, cocoa, margarine, cheese, clarified, melted, jelly, milk | butter, on, large_near, with, partly_outside, sauteeing_in |
| flower | buds, heads, spikes, lotus, beds, leaves, antlike, petals, stalks, garden | flower, in, on, vase, has, with |
| rain | heavy, torrential, singin, snow, forest, forests, fell, snowfall, shine, pouring | rain, walking_on, falling_on, walking_in, towns, getting_wet_by |

| | | |
|---|---|---|
| coffee | shop, shops, beans, plantations, tea, roasters, table, starbucks, house | coffee, cup, mug, in, of, filled_with |
| cow | milk, dung, clarabelle, calf, mad, henry, slaughter, parsnip, pasture, milking | cow, has, of, in, on, standing_in |
| wig | blonde, wam, wag, blond, wearing, jagbags, wamania, brunette, wore, mask | wig, in_clown, wearing, wears, has, wearing_a |
| tower | conning, bell, clock, eiffel, hamlets, built, london, water, observation, babel | tower, on, has, clock, on_top_of, in |
| blue | jays, ribbon, jackets, devils, collar, bombers, ridge, dark, white, toronto | blue, clear, on, wearing, and_white, in |
| skin | irritation, lesions, color, rashes, grafts, cancer, pigmentation, diseases, mucous, graft | skin, has, of, appearing_on, on, hanging_h |
| flexible | compacting, enough, sigmoidoscopy, thin, fuel, scheduling, highly, adaptable, stalk, plastic | flexible, frisbee, green_rim, to_catch_a, gren, in_dogs |
| bag | plastic, duffel, sleeping, punching, paper, mixed, grab, duffle, containing, bidita | bag, carrying, on, in, holding, has |
| bird | species, passerine, sanctuary, migratory, watching, prey, family, important, caged | bird, has, on, of, in, flying_in |
| kitchen | hell, sink, soup, dining, utensils, bathroom, room, garden, pantry, laundry | kitchen, in, in_a, inside_of, working_in, cabinet |
| father | died, succeeded, death, law, son, biological, footsteps, inherited, adoptive | father, taking_a_picture_in, taking_a_self_in, leaning_over_to_touch, rolling_a, walking_down_a |
| shore | north, lake, eastern, dinah, batteries, south, western, jersey, southern | shore, on, breaking_on, washing_on, coming_to, crashing_on |
| vehicle | motor, launch, electric, registration, aerial, utility, wheeled, armored, armoured, reentry | vehicle, on, of, parked_on, has, parked_alongside_of |
| bring | back, would, together, helped, could, able, attention, order, attempt, forth | bring, says, wall, on_a, has_a, has |
| smile | smiley, face, dk, smile, operation, make, lisa, carnt, toothy, frown | smile, exposing, revealing, has, on, face |
| time | first, full, long, spent, around, real, extra, second, short, spend | time, having_great, counts, on, tells, shows |
| fact | despite, due, finding, matter, spite, checking, many, complicated, refers, attributed | fact, listed_on, with_some, ingredient, jug, has |
| football | league, team, club, college, national, american, player, professional, coach, victorian | football, chases, playing, trying_to_save, to_hinder, placing |
| fish | finned, wildlife, cyprinid, freshwater, bony, hatchery, species, mardy, shellfish | fish, in, on, served_on, mole_on_a, aquarium |
| film | festival, directed, drama, feature, comedy, cannes, documentary, international, horror, short | film, taken_with, are_being, taped_to, being, in |
| arms | coat, coats, small, embargo, ammunition, legs, municipality's, bear, gules, dealer | arms, skater, skateboarder, outstretched, bare, fold |
| plant | flowering, power, species, family, host, pathogen, ornamental, manufacturing, treatment | plant, on, in, growing_in, growing_on, pot |
| food | fast, drug, drink, beverage, supplies, shortages, processing, safety, source, agriculture | food, on, in, on_top_of, plate, with |
| make | sure, would, way, failed, amends, could, wanted, order, room, able | make, exchange, being, splash, fishpond, construct |
| army | u, states, british, red, corps, liberation, air, officer, salvation, us | patch_for, army, getting_out_of, on_back_of, green, calendar |
| body | governing, whorl, student, human, weight, dead, parts, length, sanctioning, main | body, of, has, on, in, of_a |
| school | high, elementary, secondary, district, grammar, primary, middle, law, districts, public | school, bus_that, on, front, bus, yellow |
| forest | nottingham, wake, national, service, montane, epping, lawn, boreal, teutoburg, lowland | forest, in, in_a, tree, behind, filled_with |

| | | |
|---|---|---|
| new | york, zealand, jersey, orleans, hampshire, guinea, south, mexico, brunswick, papua | new, urban, aspire, beard_and, generation_wide_screen_smart, model_white |
| city | york, kansas, council, makeup, mexico, limits, population, centre, capital, oklahoma | city, in, in_a, shining_in, build, building_in_a |
| people | notable, surname, 000, per, young, republic, many, living, million, employed | people, on, in, watching, walking_on, are_enjoying |
| family | moth, beetle, cerambycidae, mollusk, size, average, income, crambidae, geometridae, erebidae | family, seated_around, stand_with, having, on_ground_for, sitting_around |
| house | representatives, commons, lords, white, opera, manor, publishing, built, historic, delegates | house, on, in_front_of, has, behind, near |
| year | old, following, contract, every, per, one, next, later, previous | year, pub, states, 1893, age, was_taken_in |
| party | communist, democratic, labour, liberal, conservative, republican, socialist, political, janata, labor | party, having, in, dance, at_a_birthday, crying_at |
| company | parent, production, founded, insurance, publishing, holding, manufacturing, india, brewing, theatre | company, of_photography, are_enjoying_each_others, calls, blender, boeing |

Table F.2: Context words of cluster centroids with the 10 highest PMI$^3$ score.