In search of meaning and its representations for computational linguistics

Simon Dobnik[⋆], **Robin Cooper**[◇], **Adam Ek**[⋆], **Bill Noble**[⋆], **Staffan Larsson**[◇], **Nikolai Ilinykh**[⋆], **Vladislav Maraev**[⋆] **and Vidya Somashekarappa**[⋆*]
Department of Philosophy, Linguistics and Theory of Science
Centre for Linguistic Theory and Studies in Probability (CLASP)
University of Gothenburg, Sweden
[⋆] name.surname@gu.se [◇] name.surname@ling.gu.se

Abstract

In this paper we examine different meaning representations that are commonly used in different natural language applications today and discuss their limits, both in terms of the aspects of the natural language meaning they are modelling and in terms of the aspects of the application for which they are used.

1 Introduction

A crucial component to produce a "successful" NLP system is sufficiently expressive representations of meaning. We consider a sufficiently expressive meaning representation to be one that allows a system's output to be considered acceptable to native speakers given the task. In this paper we present several features of meaning and discuss how different methods of deriving meaning representations capture these features. This list is by no means exhaustive. It might be viewed as a first attempt to discuss ways of establishing a general methodology for evaluating meaning representations and characterising what kinds of applications they might be useful for.

2 Formal meaning representations

The rigour of the work on semantics by Richard Montague (Montague, 1973; Partee, 1976) inspired early work on computational semantics (perhaps the earliest was Friedman and Warren, 1978; Friedman et al., 1978). Two high-points of the literature on computational semantics based on Montague are Blackburn and Bos (2005), using logic programming, and van Eijck and Unger (2010), using functional programming. Montague's semantic techniques have also played an important role in semantic treatments using Combinatory Categorial Grammar (CCG, Bos et al., 2004).

One problem with Montague's treatment of semantics was that it was limited to the level of the sentence. It could not, for example, deal with cross-sentence anaphora such as *A dog$_i$ barked. It$_i$ was upset by the intruder*. This, among several other things, led to the development of Discourse Representation Theory (DRT, Kamp and Reyle, 1993; Kamp et al., 2011) and other variants of *dynamic* semantics such as Heim (1982) and Groenendijk and Stokhof (1991). Here "dynamic" is meant in the sense of treating semantic content as context change potential in order, among other things, to be able to pass referents from one sentence to a subsequent sentence in the discourse. This is a much less radical notion of dynamic interpretation than we discuss in Section 4, where the meaning associated with a word or phrase may change as a dialogue progresses. DRT has played an important role in computational semantics from early work on the Verbmobil project (Bos et al., 1996) to work by Johan Bos and others on the Groningen Meaning Bank[1] and the Parallel Meaning Bank[2].

One of the cornerstones of Montague's approach is **compositionality**, the ability to compute the meaning of phrases on the basis of the meanings of their immediate sub-constituents. Another central feature in the Montague tradition is the ability to derive conclusions based on **logical inference**, including logical inferences based on the semantics of logical constants such as *and*, *not* and logical quantifiers, and the ability to characterise additional axioms or "meaning postulates". Defeasible reasoning has been added to this kind of framework (e.g., Asher and Lascarides, 2003) and systems have been connected to theorem provers and model builders (Blackburn and Bos, 2005). The variants of dynamic semantics discussed above gave us the

[*] All authors contributed equally.

[1] https://gmb.let.rug.nl/
[2] https://pmb.let.rug.nl/

Proceedings of the 2022 CLASP Conference on (Dis)embodiment, pages 30–44
September 15–16, 2022.

ability to treat **discourse phenomena**. That is, phenomena occurring in texts or utterances of more than a single sentence, including cases of discourse anaphora. **Underspecified meaning representations** are single representations which cover several meanings in cases where there is systematic ambiguity. While there is some work on underspecification of meaning in the theoretical literature (Reyle, 1993), the most interest has been devoted to it in computational work based on formal semantics (such as Alshawi, 1992; Bos, 1996; Copestake et al., 2005). **Model theory** deals with representing the relationship between language and the world, in computational terms to database queries (Blackburn and Bos, 2005; van Eijck and Unger, 2010).

What we have sketched above might be called the classical canon of formal semantics as it relates to computational semantics. One of the features lacking in the classical canon includes **dialogue**. The notion that language is actually used in interaction between agents engaging in communication (and therefore going beyond the notion of discourse in texts discussed earlier) came quite late to formal semantics though there is now a significant body of theoretical work. Notions of dialogue semantics covering plan-based approaches to dialogue (Allen, 1988), questions under discussion (Ginzburg, 1994, 2012) and communicative grounding (Traum, 1994) became central in the literature on formal approaches to dialogue. This gave rise to the Information State Update approach to dialogue (Larsson and Traum, 2000; Larsson, 2002). TTR (a theory of types with records, Cooper, 2005, forthc) has played an important role in this. **Similarity of meaning** is another feature. In addition to meaning relations such as entailment there is a notion of words, phrases and sentences having similar meanings in various respects. In a formal meaning representation this can be represented, for example, by the use of record types in TTR. Yet another feature is **robust non-logical inference** which is represented, for example, in work on textual entailment now commonly referred to as Natural Language Inference (NLI). This is hard to square with the logic-based inference discussed above. Rather than representing something that follows logically, it corresponds to what conclusions people might draw from a given utterance or text. It is often reliant on background knowledge and is to a large extent defeasible. The work on topoi by Breitholtz (2020) and probabilistic TTR (Cooper et al., 2015) is suggestive of a computational approach to this. Finally, while model theory purports to relate language and the world it tells us little about how we relate our perception of the world and action in the world to the meaning of words and phrases which is known as **perceptual grounding**. Such issues become important, for example, if we want to put natural language on board a robot. This has become central to theories such as TTR (Cooper, forthc; Larsson, 2013; Dobnik et al., 2013) and Dynamic Syntax (Kempson et al., 2016). There is important formal work on **multimodal nature of communication**, for example (Lücking, 2016; Pustejovsky and Krishnaswamy, 2020).

Above we have mentioned examples of formal approaches which attempt to incorporate features which are not present in the classical canon. An alternative strategy is to try to incorporate features from the classical canon in non-formal approaches (e.g. Coecke et al., 2010) or to combine aspects of non-formal and formal approaches in a single framework (e.g. Erk and Herbelot, 2020).

3 Distributional meaning representations

Meaning as a function of its usage can be traced back to Wittgenstein (1953), but were popularised by Firth (1957). The idea at its core is that the meaning of a word is given by its context. Wittgenstein (1953) primarily speaks about meaning in relation to the world and real world activities while Firth (1957) speaks about language in relation to language. The second notion of meaning is the basis for distributional semantics. The notion that meaning in language can be found based on language context is related to the observation that if two words occur in the same context, their meaning is likely related.

The two predominant approaches to constructing distributional meaning representations today is to use machine learning to construct distributed and contextualised word representations (Sahlgren, 2006; Mikolov et al., 2013a; Peters et al., 2018). In these approaches, the meaning of a word is encoded as a dense vector of real valued numbers. The values of the vector are obtained by training a neural network to perform some task, using a (possibly annotated) corpus. The task then helps guide the neural network to produce meaningful representations. Distributed word representations focus on building static representations of words given a corpus. Popular techniques for obtaining

these representations are BoW (Bag-of-Words) or SGNS (Skip Gram with Negative Sampling), popularised by (Mikolov et al., 2013a). The main trick to BoW and SGNS is to construct a training schema such that given a random meaning representation for the word x, the representation is transformed so it can be used to identify the word in a context[3], or can be used to to identify the context of the word. The BoW or SGNS meaning representations can then be used as a component in another system. Contextualised representations on the other hand build dynamic word representations, that is, a single word will have different vector representations in different contexts. These representations are typically informed by the output from a language model. Thus, to really exploit contextual representations effectively a sentence is needed when extracting the meaning representation. This is in contrast to the BoW and SGNS representations which are fixed after being constructed.

With distributed representations we may also **reason analogically** about words and combinations of concepts, e.g. "Russia" + "River" = "Volga" (Mikolov et al., 2013b). That is, we may construct complex meaning by combining simpler parts. By combining the representation for "Russia" and "river" we obtain some vector z which contains information about the contexts of both "Russia" and "river". By querying the vector space for words with a *similar* representation to that of z we find other words with similar context. The success of distributed meaning representations, both static and contextualised, can in part be attributed to the ability of a model to **predict similarity** between units of language. Because meaning is defined as the context in which words occurs, two vector representations can be compared and their similarity measured. (Conneau and Kiela, 2018; Vulic et al., 2020). This similarity can be explored in terms of words (Hill et al., 2015; Artetxe et al., 2016) and in term of sentences (Cer et al., 2017).

The ability to model similarity allows models to **discover relationships** between units of language. It allows models to transfer knowledge between languages. For example, unsupervised word translation can be done by aligning monolingual vector spaces. (Lample et al., 2018; Artetxe et al., 2018). Transformer models (Vaswani et al., 2017) have also enabled zero-shot and transfer learning, e.g. by learning English word representations and evaluating on a task in another language (Pires et al., 2019). The simplicity of static and contextualised meaning representations allows us to construct them for *any* unit of language, be it words, sentences (Conneau et al., 2018), documents (Lau and Baldwin, 2016) or languages (Östling and Tiedemann, 2017).

However, a word or a sentence may mean different things depending also on a larger context. For example a sentence in different domains will express different meanings even if the words are exactly the same. This presents a problem for distributed representations, as our observation of a word or sentence in the real world includes additional context from what we have recorded in the data. However, the effects of different domains may be counteracted by **domain adaptation** techniques (Jiang and Zhai, 2007).

Distributed representations enjoy success across a wide variety of NLP tasks. However, a consequence of automatically learning from a corpus results in some inherent shortcomings. A corpus is a snapshot of a subset of language and only captures language as it was used then and there. This means that the resulting meaning representations do not inherently capture language change (though they can used to study it, see Section 4). Additionally, the meaning representations are generally created from observing language in a corpora, not from language use in the world. A consequence of this is that distributional meaning representations don't capture the state-of-affairs in the world, i.e. the context in which the language was used. In practical terms this means that for tasks that depend on the state-of-affairs in the world, such as robot control, dialogue or image captioning, a system must incorporate this information somehow which is further explored in the remaining sections.

4 Dynamic meaning representations

To see how meaning is context dependent in (at least) two different ways we can make the distinction between *meaning potential* and *situated meaning* (Norén and Linell, 2007). The situated meaning of a word is its disambiguated and contextually enriched interpretation in a particular context of use. Meaning potential (or *lexical meaning*) is the system of affordances (Gibson, 1966; Gregoromichelaki et al., 2020) that a word offers for sense-making in a given context. In this conception, situated meaning is context dependent by construction, but we also claim that the meaning

[3]Context here is typically a n-gram containing x.

potential of a word depends on context of a certain kind. In particular, it depends on what is *common ground* (Stalnaker, 2002) between a speaker and their audience. At a linguistics conference, a speaker might use words like *token* or *modality*—words that would mean something completely different (or nothing at all) at a family dinner. The conference speaker expects to be understood *based on* their and their audience's joint membership in the computational linguistics community, where they (rightly or wrongly) consider certain specialised meanings to be common ground. The communities that serve as a basis for semantic common ground can be as broad as *speakers of Spanish* (grounding the "standard" Spanish lexicon), or as small as a nuclear family or close group of friends (grounding specialised meanings particular to that group of people) (Clark, 1996).

Recent work in NLP has demonstrated the value of modelling context, including sentential (Section 3) and multimodal context (Section 5) for representing situated meanings. The dynamic representations given by language models like BERT depend on the *local context* in which the word appears, but don't the context of the community or individual speakers involved. Little work has been done in NLP to explicitly incorporate *social context*, which provides the basis for semantic common ground. Recent work has shown that neural language models can be used to detect and analyse variation and change in post-hoc way (Del Tredici et al., 2019a; Giulianelli et al., 2020). This suggests that explicitly modelling social context may be a fruitful way forward.

In the following, we identify three kinds of social context that might be accounted for with dynamic meaning representations

Variation As demonstrated in the conference example, lexical meaning is community dependent. This doesn't necessarily mean that every NLP application needs to mimic the human ability to tailor our semantic representations to the different communities we belong to, but some applications may serve a broader set of users by doing so (Hovy, 2015). Consider, for example, an application that serves both the general public and experts in some domain. Even where variation is not explicitly modelled, it is an important factor to consider on a meta level. In practice, NLP models typically target the most prestigious, hegemonic dialect of a given language, due in part to biases in what training data is easily available on the internet (Bender et al., 2021). This results in applications that favour users who are more comfortable with the dominant language variety. Furthermore, many applications *assume* a single variety of a given language, when in fact the training data of the models they rely on is rather specific. The standard English BERT model, for example, is trained on a corpus of unpublished romance novels and encyclopedia articles, but is applied as if it represents English written large.

Alignment Semantic common ground is not only based on joint community membership—it can also be built up between particular agents through interaction. Experiments have shown that pairs of speakers develop shorter lexicalised referring expressions when they need to repeatedly identify a referent (Mills and Healey, 2008). Additions or modifications to existing common ground can take place implicitly (through *semantic accommodation*) or *meaning accommodation* (Larsson, 2010) or explicitly, as in *word meaning negotiation* (Myrendal, 2015).

There is some hope for developing models that dynamically update their meaning representations based on interaction with other agents. Larsson and Myrendal (2017) suggest an inventory of semantic update functions that could be applied to formal meaning representations based on the results of an explicit word meaning negotiation. Dynamic Interpretation Theory (Bunt, 2000) offers a way of representing meaning as change to the conversational context, including social context, and has been incorporated in the implementation of several dialogue managers (Keizer et al., 2011; Malchanau, 2019). On the distributional side, one- or few-shot learning may eventually allow models to generalise from a small number of novel uses by drawing on existing structure in the lexicon (Lake et al., 2019). One question that remains unexplored in both these cases is which updates to local (dialogue or partner-specific) semantic ground should be propagated to the agent's representation of the communal common ground (and to which community). This naturally bring up the issue of community-level semantic change.

Change How words change in meaning has long been an object of study for historical linguists (e.g., Paul, 1891; Blank, 1999). Historical change may not seem like a particularly important thing for NLP applications to model. After all, we can accommodate for changes over decades or cen-

turies by simply retraining models with more current data, but significant *semantic shift* can also take place over a much shorter timeline, especially in smaller speech communities (Eckert and McConnell-Ginet, 1992). The issue of semantic change also intersects with that of variation, since coinages and shifts in meaning that take place in one community can cause the lexical common ground to diverge from another community. Conversely, variants in one community may come to be adopted by another (possibly broader) community. The recent widespread use of distributional semantics to study semantic change suggests that distributional representations are capable of capturing change.[4] Diachronic distributional representations have been used to study semantic change on both a historic/language level (e.g., Dubossarsky et al., 2015; Hamilton et al., 2016) and on a short-term/community level (Rosenfeld and Erk, 2018; Del Tredici et al., 2019b; Noble et al., 2021). While social context is not often taken into account in meaning representations, ongoing research on semantic variation and change suggests that such dynamic representations are possible as extensions of the formal and distributional paradigms.

5 Grounded meaning representations

The meaning of words is not merely in our head. It is grounded in our surroundings and tied to our understanding of the world (Regier, 1996; Bender and Koller, 2020), particularly through visual perception (Mooney, 2008). Mapping language and vision to get **multi-modal** meaning representations imposes an important challenge for many real-world NLP applications, e.g. conversational agents. Such agents typically learn by finding statistical relations and often lack causal reasoning about the world (Agarwal et al., 2020) and common-sense knowledge (Hwang et al., 2021). This section describes how different modalities are typically integrated to get a meaning representation for **language-and-vision (L&V)** tasks and what is still missing in the respective **information fusion** techniques.

Historically, modelling of situated language has been influenced by ideas from language technology, computer vision and robotics, where a combination of top-down rule-based language systems was connected with Bayesian models or other kinds of classifiers of action and perception (Kruijff et al., 2007; Dobnik, 2009; Tellex et al., 2011; Mitchell et al., 2012). In these approaches, most of the focus was on how to ground words or phrases in representations of perception and action through classification. Another reason for this hybrid approach has also been that such models are partially interpretable. Therefore, they have been a preferred choice in critical robotic applications where security is an issue. The compositionality of semantic representations in these systems is ensured by using semantic grammars, while perceptual representations such as SLAM maps (Dissanayake et al., 2001) or detected visual features (Lowe, 1999) provide a model for interpreting linguistic semantic representations. Deep learning, where linguistic and perceptual features are learned in an interdependent manner rather than engineered, has proven to be greatly helpful for the task of image captioning (Vinyals et al., 2015; Anderson et al., 2018a; Bernardi et al., 2016) and referring expression generation (Kazemzadeh et al., 2014).

A more in-depth analysis of how meaning is represented in these models is required. Ghanimifard and Dobnik (2017) show that a neural language model can learn compositionality by grounding an element in the spatial phrase in some perceptual representation. In terms of methodology for understanding what type of meaning is captured by the model, attention (Xu et al., 2015; Lu et al., 2017) has been successfully used. Lu et al. (2016) have shown that co-attending to image and question leads to a better understanding of the regions and words the model is focused on the most. Ilinykh and Dobnik (2020) demonstrate that attention can struggle with fusing multi-modal information into a single meaning representation based on the human evaluation of generated image paragraphs. This is because the nature of visual and linguistic features and the model's structure significantly impact what representations can be learned when using an attention mechanism. Examining attention shows that attention can correctly attend to objects, but once it is tasked to generate relations (such as prepositional spatial relations and verbs), attention visually disappears as these relations are non-identifiable in the visual features utilised by the model. This leads several researchers to include specifically geometric information in image captioning models (Sadeghi et al., 2015; Ramisa et al., 2015). On the other hand, it has also been shown that a lot of meaning can be extracted solely from word dis-

[4] See (Tahmasebi et al., 2018), (Tang, 2018), and (Kutuzov et al., 2018) for recent surveys.

tributions. Choi (2020) demonstrates how linguistic descriptions encode common-sense knowledge which can be applied to several tasks while Dobnik and Kelleher (2013); Dobnik et al. (2018) demonstrate that word distributions are an important part of the semantics of spatial relations.

Interactive set-ups such as visual question answering (VQA) (Antol et al., 2015; de Vries et al., 2017) or visual dialogue (Das et al., 2017) make first attempts in modelling multi-modal meaning in multi-turn interaction. However, such set-ups are asymmetric in terms of each interlocutor's roles, which leads to homogeneous question-answer pairs with rigid word meaning. *Conversational games* have been proposed as set-ups in which the meaning of utterances is agreed upon in a collaborative setting (Dobnik and Storckenfeldt, 2018). These settings allow for modelling meaning coordination of grounded perceptual classifiers (Larsson, 2013) and phenomena such as clarification requests. Several corpora of perceptual dialogue exist where conversational partners need to leverage dialogue and visual information to achieve mutual understanding of a scene, for example MeetUp! (Ilinykh et al., 2019), PhotoBook (Haber et al., 2019) and Cups (Dobnik et al., 2020).

Examining L&V models and representations they learn points to several significant and interesting challenges. The first relates to the structure of both datasets and models. Many corpora contain prototypical scenes where the model can primarily optimise on the information from the language model to generate an answer without even looking at the image (Cadene et al., 2019). Secondly, information captured by a language model is more compact and expressive than patterns of visual and geometric features. Thirdly, common-sense and visual information are not enough (Lake et al., 2017; Bisk et al., 2020; Tenenbaum, 2020): we also rely on mental simulation of the scene's physics to estimate, for example, from the appearance and position of a person's body that they are making a jump on their skateboard rather than they are falling over a fire hydrant. Such representations are necessary for modelling *embodied agents* (Anderson et al., 2018b; Das et al., 2018; Kottur et al., 2018). Fourthly, adding more modalities and representations puts new requirements on inference procedures and more sophisticated models of attention (Lavie et al., 2004) that weighs to what degree such features are relevant in a particular context. In recent years we have seen work along these lines implemented with a transformer architecture (Lu et al., 2019; Su et al., 2020; Herdade et al., 2019). However, the interpretability of how individual parts (self-attentions) of large-scale models process information from different modalities is still an open question (Ilinykh and Dobnik, 2022).

6 Meaning expressed with our body

Meanings can result in bodily reactions and, conversely, they can be expressed with our bodies, for example non-verbal vocalisations, gaze and gestures.

Emotions Meanings perceived from the environment can change our emotional states and be expressed in bodily reactions: evaluating events as intrinsically unpleasant may result in gaze aversion, pupillary constriction and some of the other components (Scherer, 2009). On the other hand, our emotional states can be expressed and the expressions can be adjusted by emotional components, such as mood (Marsella et al., 2010).

Over the last years *appraisal theories* became the leading theories of emotions (for overview, see Oatley and Johnson-Laird, 2014). These theories posit that emotion arises from a person's interpretation of their relationship with the environment or *appraisal*. The key idea behind cognitive theories is that emotions not only reflect physical states of the agents but also emotions are judgements, depending on the current state of the affairs (depending on a person, significance/urgency of the event etc.). In our view, linguistic events can as well enter the calculation of appraisal on the level of information-state of the agent which can be modelled by formal theories. For instance, following Oatley and Johnson-Laird (2014) we can distinguish emotions as either free-floating or requiring an object such as a linguistic entity, entity in the environment or a part of agent's information-state (e.g., obstruction of the agent's goal can lead to anger or irritation, and, vice versa, agent's sadness can lead to the search for a new plan). Several attempts implement emotional appraisal in text and speech (e.g., Alm, 2012), and within the agent models (e.g., Marsella et al., 2010).

Non-verbal vocalisations Non-verbal vocalisations, such as laughter, are ubiquitous in our everyday interactions. In the British National Corpus laughter is a quite frequent signal regardless of gender and age—the spoken dialogue part of the

British National Corpus contains approximately one laughter event every 14 utterances. In the Switchboard Dialogue Act corpus non-verbally vocalised dialogue acts (whole utterances marked as non-verbal) constitute 1.7% of all dialogue acts and laughter tokens make up 0.5% of all the tokens that occur in the corpus.

Despite a distinct bodily reaction (laughter causes tensions and relaxations of our bodies), it is believed that we laugh in a very different sense from sneezing or coughing (Prusak, 2006). Many scholars agree that laughter is not involuntary but we laugh for a reason, *about* something and that laughter perfomrs a social function (Mehu, 2011). It is associated with senses of closeness and affiliation, establishing social bonding and smoothing away discomfort. For example, tickling not only requires the presence of the other but also it is more likely if subjects have close relationships (Harris, 1999).

Therefore, the meaning of laughter ought to be represented so that an artificial agent can understand it and react to it accordingly (Maraev et al., 2018; Mazzocconi et al., 2021). Mazzocconi (2019) presents a function-based taxonomy of laughter, distinguishing functions such as indication of pleasant incongruity or smoothing the discomfort in conversation. Ginzburg et al. (2020) propose a formal account of laughter within the information-state of dialogue participants which includes scaling up to non-verbal social signals such as smiling, sighing, eye rolling and frowning.

Gaze Gaze has many functions. It can dictate attention, intentions, and serves to give communicative cues in interaction (Somashekarappa et al., 2021). Gaze following can infer objects that people are looking at. While we scan a visual scene, our brain stores fixation sequences in memory and reactivates them when visualising the scene later in the absence of any perceptual information (Brandt and Stark, 1997). Scan-path theory illustrations indicate that meaning representations on scanned areas depend on the semantics of sentences (Bochynska and Laeng, 2015). Semantic eye fixations supports the view of mental imagery that is flexible and creative. Being grounded in previous experiences, by selecting a past episode we are able to generalise the past information to novel images that share features (Martarelli et al., 2017). Spatial representations associated with different semantic categories launch eye movements during retrieval (Spivey et al., 2000).

For dialogue participants gaze patterns represent resources to track their stances. Interlocutors engage in mutual gaze while producing agreeing assessments (Haddington, 2006). Gaze shifts follow sequentially a problematic stance and are followed by a divergent stance by the person who produced the gaze shift. Gaze patterns are not meaningful per se but acquire interpretation within their linguistic and interactional contexts.

Eye movement patterns, EEG signals and brain imaging are some of the techniques that have been used to augment traditional qualitative text-based features. Temporal course and flexibility of the speaker's eye gaze can be used to disambiguate referring expressions in spontaneous dialogue. Eye-tracking data from reading experiments provide structured information with fine-grained temporal resolution which closely follows sequential structure of speech and is related to the cognitive workload of speech processing (Barrett and Hollenstein, 2020). Deep convolutional neural networks have been used to classify text to gaze using eye movements. Their performance has improved when human readers were tackling semantic challenges (Mishra and Bhattacharyya, 2018). For multimodal and multi-party interaction in both social and referential scenarios, (Somashekarappa et al., 2020) calls for categorical representation of gaze patterns.

Gestures Gestures are hand and body movements that help to convey information (Kita and Özyürek, 2003). The observational, experimental, behavioural and neuro-cognitive evidence indicates that language and gestures are linked both during comprehension and production (Wilkins, 2006; Willems et al., 2007). Speech and gestures are semantically and temporally coordinated and therefore involved in co-production of meaning. Gestures convey meaning through iconicity and spacial proximity providing information that is not necessarily expressed in speech.

While shaping of gestures is related to conceptual and semantic aspects of the accompanying speech, gestures cannot be unambiguously interpreted by naïve listeners (Hadar and Pinchas-Zamir, 2004). However, Morett et al. (2020) showed that the semantic relationship between representational gestures vs their lexical affiliates and language is evaluated similarly. Mentions of referents for the first time in a discourse are often accompanied by

gestures. For example, Debreslioska and Gullberg (2020) report that the "entity" gesture accompanies referents expressed by indefinite nominals. As referents are introduced in clauses, inferable referents referred to by definite nominals are identified by the contrasting "action" gestures. Head movements are produced to give feedback (Petukhova and Bunt, 2009) and it is possible to identify a specific pattern for a specific movement and that movements can be easily measured and their extent can be quantified (Allwood and Cerrato, 2003).

Fixing gesture functions, integrating different modalities and determining their composite meanings is challenging. For artificial agents multimodal output planning is crucial and timing must be explicitly represented. Lücking (2016) takes a qualitative formal approach from a type-theoretic perspective, representing iconic gestures in TTR and linking them with linguistic predicates. (Pustejovsky and Krishnaswamy, 2020) take a hybrid approach linking qualitative representations in VoxML with machine learning classification.

7 Conclusions

We surveyed formal, distributional, interactive, multi-modal and body-related representations of meaning used in computational linguistics. They are able to deal with compositionality, underspecification, similarity of meaning, inference and provide an interpretation of expressions but in very different ways, capturing very different kinds of meaning. These aspects can be broadly categorised into (i) aspects that are related to the construction of linguistic forms and (ii) aspects concerned with the interpretations and understanding the world and human activity in it.

Current mainstream computational linguistics is a practical field which is not working toward a uniform model of human language but focuses on several sub-tasks which, although related, are frequently considered in isolation. For example, natural language understanding and natural language generation frequently use entirely different approaches and representations even when the linguistic context is the same, for example texts or image captions. Solutions are provided given the practical goals and limitations of each task. Secondly, the solutions are also limited by what linguistic information can be feasibly collected for this task and by our understanding of human language and behaviour (or its lack-of) as witnessed by ongoing work in linguistics and psychology.

If our goal is to translate documents or answer general fact-based questions then a reasonable performance can be achieved even if the system is able to ground representations only indirectly in linguistic contexts over situations rather than situations themselves. However, for a situated robot semantic grounding in texts, although relevant, is not enough as it has to connect language with its environment that it accesses with its sensors and actuators. For example, word embeddings for *left* and *right* will tell us that they are similar relations but also that they have slightly different selectional preferences for objects that they relate.

Humans rely on different aspects of meaning for different kinds of descriptions and contexts and it may be perfectly fine for our task-specific computational models to only use some dimensions and in an indirect way. What is wrong to claim, however, is that any of these models have reached human-like intelligence. Humans can re-evaluate linguistic descriptions against different dimensions of meaning and this is something that our systems are not capable of. Work on the cognitive notion of attention informs us how aspects of meaning representations are selected to disambiguate under-specified linguistic utterances by balancing information from different modalities.

A stronger connection between representations from different tasks is certainly desirable and important progress has been made for example in integration of formal grammars in situated agent systems or integration of vision and language representations learned by neural networks. However, the challenge remains precisely because "linguistic experiences" of our systems are limited by the narrow tasks and domains that they are specialised on. Transferring models across contexts of language use and aligning their representations is by no means straightforward as there may be very little overlap. At the same time we also expect that models learn generalisations that apply across contexts. In line with this we suggest that future work should focus on developing benchmarks that test different representations in different contexts. For this we need datasets of instances requiring some type of linguistic inference where instances are labelled by the context type and the type(s) of modality required for successful inference. We hope that this paper points to some of the aspects of representations that need to be taken into account.

Acknowledgements

The research reported in this paper was supported by a grant from the Swedish Research Council (VR project 2014-39) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg.

References

Vedika Agarwal, Rakshith Shetty, and Mario Fritz. 2020. Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9690–9698.

James Allen. 1988. *Natural language understanding*. Benjamin-Cummings Publishing Co., Inc.

Jens Allwood and Loredana Cerrato. 2003. A study of gestural feedback expressions. In *First nordic symposium on multimodal communication*, pages 7–22. Copenhagen.

Cecilia Ovesdotter Alm. 2012. The role of affect in the computational modeling of *Natural* language. *Lang. Linguistics Compass*, 6(7):416–430.

Hiyan Alshawi, editor. 1992. *The Core Language Engine*. MIT Press.

P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. 2018a. Bottom-up and top-down attention for image captioning and visual question answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6077–6086.

P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, and A. van den Hengel. 2018b. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3674–3683.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2289–2294. The Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 789–798. Association for Computational Linguistics.

N Asher and A Lascarides. 2003. *Logics of conversation*. Cambridge University Press.

Maria Barrett and Nora Hollenstein. 2020. Sequence labelling and sequence classification with gaze: Novel uses of eye-tracking data for natural language processing. *Language and Linguistics Compass*, 14.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 610–623, New York, NY, USA. Association for Computing Machinery.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *J. Artif. Int. Res.*, 55(1):409–442.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience grounds language. *arXiv*, arXiv:2004.10151 [cs.CL].

Patrick Blackburn and Johan Bos. 2005. *Representation and Inference for Natural Language: A First Course in Computational Semantics*. CSLI Studies in Computational Linguistics. CSLI Publications, Stanford.

A. Christian Blank. 1999. Why do new meanings occur? A cognitive typology of the motivations for lexical semantic change. In Andreas Blank and Peter Koch, editors, *Historical Semantics and Cognition*. De Gruyter Mouton.

Agata Bochynska and Bruno Laeng. 2015. Tracking down the path of memory: eye scanpaths facilitate retrieval of visuospatial information. *Cognitive processing*, 16 Suppl 1.

J. Bos. 1996. Predicate logic unplugged. In *Proceedings of the Tenth Amsterdam Colloquium*, pages 133–143, Amsterdam. ILLC/Department of Philosophy, University of Amsterdam.

Johan Bos, Stephen Clark, Mark Steedman, James R Curran, and Julia Hockenmaier. 2004. Wide-coverage semantic representations from a CCG parser. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 1240–1246.

Johan Bos, Björn Gambäck, Christian Lieske, Yoshiki Mori, Manfred Pinkal, and Karsten Worm. 1996. Compositional semantics in Verbmobil. *arXiv preprint cmp-lg/9607031*.

Stephan Brandt and Lawrence Stark. 1997. Spontaneous eye movements during visual imagery reflect the content of the visual scene. *Journal of cognitive neuroscience*, 9:27–38.

Ellen Breitholtz. 2020. *Enthymemes in Dialogue*. Brill.

Harry Bunt. 2000. Dialogue pragmatics and context specification. In Harry Bunt and William Black, editors, *Abduction, belief and context in dialogue: studies in computational pragmatics*, pages 81–150. John Benjamins Publishing, Amsterdam/Philadelphia.

Remi Cadene, Corentin Dancette, Matthieu Cord, and Devi Parikh. 2019. RUBi: Reducing unimodal biases for visual question answering. In *NeurIPS*, pages 841–852.

Daniel M. Cer, Mona T. Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity - multilingual and cross-lingual focused evaluation. *CoRR*, abs/1708.00055.

Yejin Choi. 2020. Intuitive reasoning as (un)supervised language generation. Seminar, Paul G. Allen School of Computer Science and Enginneering, University of Washington and Allen Institute for Artificial Intelligence, MIT Embodied Intelligence Seminar.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press.

Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen Clark. 2010. Mathematical foundations for a compositional distributional model of meaning. *arXiv preprint arXiv:1003.4394*.

Alexis Conneau and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA).

Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single \$&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2126–2136. Association for Computational Linguistics.

Robin Cooper. 2005. Records and record types in semantic theory. *Journal of Logic and Computation*, 15(2):99–112.

Robin Cooper. forthc. *From Perception to Communication: a Theory of Types for Action and Meaning*. Oxford University Press.

Robin Cooper, Simon Dobnik, Shalom Lappin, and Staffan Larsson. 2015. Probabilistic Type Theory and Natural Language Semantics. *Linguistic Issues in Language Technology*, 10(4):1–45.

Ann Copestake, Dan Flickinger, Carl Pollard, and Ivan A Sag. 2005. Minimal recursion semantics: An introduction. *Research on language and computation*, 3(2):281–332.

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. 2018. Embodied Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sandra Debreslioska and Marianne Gullberg. 2020. What's new? gestures accompany inferable rather than brand-new referents in discourse. *Frontiers in Psychology*, 11.

Marco Del Tredici, Raquel Fernández, and Gemma Boleda. 2019a. Short-term meaning shift: A distributional exploration. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2069–2075, Minneapolis, Minnesota. Association for Computational Linguistics.

Marco Del Tredici, Raquel Fernández, and Gemma Boleda. 2019b. Short-Term Meaning Shift: A Distributional Exploration. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, volume 1 (Long and Short Papers), pages 2069–2075, Minneapolis, Minnesota. Association for Computational Linguistics.

M. W. M. G Dissanayake, P. M. Newman, H. F. Durrant-Whyte, S. Clark, and M. Csorba. 2001. A solution to the simultaneous localization and map building (SLAM) problem. *IEEE Transactions on Robotic and Automation*, 17(3):229–241.

Simon Dobnik. 2009. *Teaching mobile robots to use spatial words*. Ph.D. thesis, University of Oxford: Faculty of Linguistics, Philology and Phonetics and The Queen's College, Oxford, United Kingdom.

Simon Dobnik, Robin Cooper, and Staffan Larsson. 2013. Modelling language, action, and perception in Type Theory with Records. In Denys Duchier and Yannick Parmentier, editors, *Constraint Solving and Language Processing: 7th International Workshop, CSLP 2012, Orléans, France, September 13–14, 2012, Revised Selected Papers*, volume 8114 of *Lecture Notes in Computer Science*, pages 70–91. Springer Berlin Heidelberg.

Simon Dobnik, Mehdi Ghanimifard, and John D. Kelleher. 2018. Exploring the functional and geometric bias of spatial relations using neural language models. In *Proceedings of the First International Workshop on Spatial Language Understanding (SpLU 2018) at NAACL-HLT 2018*, pages 1–11, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Simon Dobnik and John D. Kelleher. 2013. Towards an automatic identification of functional and geometric spatial prepositions. In *Proceedings of PRE-CogSsci 2013: Production of referring expressions – bridging the gap between cognitive and computational approaches to reference*, pages 1–6, Berlin, Germany.

Simon Dobnik, John D. Kelleher, and Christine Howes. 2020. Local alignment of frame of reference assignment in English and Swedish dialogue. In *Spatial Cognition XII: Proceedings of the 12th International Conference, Spatial Cognition 2020, Riga, Latvia*, pages 251–267, Cham, Switzerland. Springer International Publishing.

Simon Dobnik and Axel Storckenfeldt. 2018. Categorisation of conversational games in free dialogue over spatial scenes. In *Proceedings of AixDial – Semdial 2018: The 22st Workshop on the Semantics and Pragmatics of Dialogue*, pages 1–3, Aix-en-Provence, France.

Haim Dubossarsky, Yulia Tsvetkov, Chris Dyer, and Eitan Grossman. 2015. A bottom up approach to category mapping and meaning change. *NetWordS 2015 Word Knowledge and Word Usage*, page 5.

Penelope Eckert and Sally McConnell-Ginet. 1992. Communities of practice: Where language, gender, and power all live. *Locating Power, Proceedings of the 1992 Berkeley Women and Language Conference*, pages 89–99.

Jan van Eijck and Christina Unger. 2010. *Computational Semantics with Functional Programming*. Cambridge University Press.

Katrin Erk and Aurelie Herbelot. 2020. How to marry a star: probabilistic constraints for meaning in context. *arXiv preprint arXiv:2009.07936*.

J. R. Firth. 1957. *Papers in Linguistics, 1934-1951*. Oxford University Press, London.

Joyce Friedman, Douglas B. Moran, and David S. Warren. 1978. Two Papers on Semantic Interpretation in Montague Grammar. *American Journal of Computational Linguistics*. Microfiche 74.

Joyce Friedman and David S Warren. 1978. A parsing method for Montague grammars. *Linguistics and Philosophy*, 2(3):347–372.

Mehdi Ghanimifard and Simon Dobnik. 2017. Learning to compose spatial relations with grounded neural language models. In *Proceedings of IWCS 2017: 12th International Conference on Computational Semantics*, pages 1–12, Montpellier, France. Association for Computational Linguistics.

James J Gibson. 1966. *The senses considered as perceptual systems*. Mifflin, New York [u.a.

Jonathan Ginzburg. 1994. An update semantics for dialogue. In *Proceedings of the 1st International Workshop on Computational Semantics*, Tilburg University. ITK Tilburg.

Jonathan Ginzburg. 2012. *The Interactive Stance: Meaning for Conversation*. Oxford University Press, Oxford.

Jonathan Ginzburg, Chiara Mazzocconi, and Ye Tian. 2020. Laughter as language. *Glossa: a journal of general linguistics*, 5(1).

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online. Association for Computational Linguistics.

Eleni Gregoromichelaki, Stergios Chatzikyriakidis, Arash Eshghi, Julian Hough, Christine Howes, Ruth Kempson, Jieun Kiaer, Matthew Purver, Mehrnoosh Sadrzadeh, and Graham White. 2020. Affordance Competition in Dialogue: The Case of Syntactic Universals. In *Proceedings of the 24th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*.

Jeroen Groenendijk and Martin Stokhof. 1991. Dynamic predicate logic. *Linguistics and Philosophy*, pages 39–100.

Janosch Haber, Tim Baumgärtner, Ece Takmaz, Lieke Gelderloos, Elia Bruni, and Raquel Fernández. 2019. The PhotoBook dataset: Building common ground through visually-grounded dialogue. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1895–1910, Florence, Italy. Association for Computational Linguistics.

Uri Hadar and Lian Pinchas-Zamir. 2004. The semantic specificity of gesture. *Journal of Language and Social Psychology - J LANG SOC PSYCHOL*, 23:204–214.

Pentti Haddington. 2006. The organization of gaze and assessments as resources for stance taking. *Text & Talk - TEXT TALK*, 26:281–328.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Christine R Harris. 1999. The mystery of ticklish laughter. *American Scientist*, 87(4):344.

Irene Heim. 1982. *The Semantics of Definite and Indefinite NPs*. Ph.D. thesis, University of Massachusetts at Amherst.

Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. 2019. Image captioning: Transforming objects into words. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Comput. Linguistics*, 41(4):665–695.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 752–762, Beijing, China. Association for Computational Linguistics.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(7):6384–6392.

Nikolai Ilinykh and Simon Dobnik. 2020. When an image tells a story: The role of visual and semantic information for generating paragraph descriptions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 338–348, Dublin, Ireland. Association for Computational Linguistics.

Nikolai Ilinykh and Simon Dobnik. 2022. Attention as grounding: Exploring textual and cross-modal attention on entities and relations in language-and-vision transformer. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4062–4073, Dublin, Ireland. Association for Computational Linguistics.

Nikolai Ilinykh, Sina Zarrieß, and David Schlangen. 2019. Meet up! a corpus of joint activity dialogues in a visual environment. In *Proceedings of the 23rd Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*, London, United Kingdom. SEMDIAL.

Jing Jiang and ChengXiang Zhai. 2007. Instance weighting for domain adaptation in NLP. In *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic*. The Association for Computational Linguistics.

Hans Kamp, Josef van Genabith, and Uwe Reyle. 2011. Discourse Representation Theory. In Dov Gabbay and Franz Guenthner, editors, *Handbook of Philosophical Logic*, volume 15. Springer Science+Business Media B.V. .

Hans Kamp and Uwe Reyle. 1993. *From Discourse to Logic*. Kluwer, Dordrecht.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L. Berg. 2014. Referit game: Referring to objects in photographs of natural scenes. In *EMNLP*.

Simon Keizer, Harry Bunt, and Volha Petukhova. 2011. Multidimensional Dialogue Management. In Antal van den Bosch and Gosse Bouma, editors, *Interactive Multi-modal Question-Answering*, Theory and Applications of Natural Language Processing, pages 57–86. Springer, Berlin, Heidelberg.

Ruth Kempson, Ronnie Cann, Eleni Gregoromichelaki, and Stergios Chatzikyriakidis. 2016. Language as Mechanisms for Interaction. *Theoretical Linguistics*, 42(3-4):203–276.

Sotaro Kita and Asli Özyürek. 2003. What does cross-linguistic variation in semantic co-ordination of speech and gesture reveal?: Evidence of an interface representation of spatial thinking and speaking. *Journal of Memory and Language*, 48:16–32.

Satwik Kottur, José M. F. Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. 2018. Visual coreference resolution in visual dialog using neural module networks.

Geert-Jan M. Kruijff, Hendrik Zender, Patric Jensfelt, and Henrik I. Christensen. 2007. Situated dialogue and spatial organization: what, where... and why? *International Journal of Advanced Robotic Systems*, 4(1):125–138.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: A survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Brenden M. Lake, Tal Linzen, and Marco Baroni. 2019. Human few-shot learning of compositional instructions. *arXiv:1901.04587 [cs]*.

Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. 2017. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Staffan Larsson. 2002. *Issue-based Dialogue Management*. Ph.D. thesis, University of Gothenburg.

Staffan Larsson. 2010. Accommodating innovative meaning in dialogue. In *Aspects of Semantics and Pragmatics of Dialogue. SemDial 2010, 14th Workshop on the Semantics and Pragmatics of Dialogue*, pages 83–90.

Staffan Larsson. 2013. Formal semantics for perceptual classification. *Journal of Logic and Computation*, 25(2):335–369.

Staffan Larsson and Jenny Myrendal. 2017. Dialogue Acts and Updates for Semantic Coordination. In *SEMDIAL 2017 (SaarDial) Workshop on the Semantics and Pragmatics of Dialogue*, pages 52–59. ISCA.

Staffan Larsson and David R Traum. 2000. Information state and dialogue management in the trindi dialogue move engine toolkit. *Natural language engineering*, 6(3-4):323–340.

Jey Han Lau and Timothy Baldwin. 2016. An empirical evaluation of doc2vec with practical insights into document embedding generation. In *Proceedings of the 1st Workshop on Representation Learning for NLP, Rep4NLP@ACL 2016, Berlin, Germany, August 11, 2016*, pages 78–86. Association for Computational Linguistics.

Nilli Lavie, Aleksandra Hirst, Jan W de Fockert, and Essi Viding. 2004. Load theory of selective attention and cognitive control. *Journal of Experimental Psychology: General*, 133(3):339–354.

David G Lowe. 1999. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. IEEE.

J. Lu, C. Xiong, D. Parikh, and R. Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3242–3250.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 289–297, Red Hook, NY, USA. Curran Associates Inc.

Andy Lücking. 2016. Modeling co-verbal gesture perception in type theory with records. In *Computer Science and Information Systems (FedCSIS), 2016 Federated Conference on*, pages 383–392. IEEE.

A. Lücking. 2016. Modeling co-verbal gesture perception in type theory with records. In *2016 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pages 383–392.

Andrei Malchanau. 2019. *Cognitive Architecture of Multimodal Multidimensional Dialogue Management*. Ph.D. thesis, Saarland University, Saarbrücken.

Vladislav Maraev, Chiara Mazzocconi, Christine Howes, and Jonathan Ginzburg. 2018. Integrating laughter into spoken dialogue systems: preliminary analysis and suggested programme. In *Proceedings of the FAIM/ISCA Workshop on Artificial Intelligence for Multimodal Human Robot Interaction*, pages 9–14.

Stacy Marsella, Jonathan Gratch, Paolo Petta, et al. 2010. Computational models of emotion. *A Blueprint for Affective Computing-A sourcebook and manual*, 11(1):21–46.

Corinna Martarelli, Sandra Chiquet, Bruno Laeng, and Fred Mast. 2017. Using space to represent categories: insights from gaze position. *Psychological Research*, 81.

Chiara Mazzocconi. 2019. *Laughter in interaction: semantics, pragmatics and child development*. Ph.D. thesis, Université de Paris.

Chiara Mazzocconi, Vladislav Maraev, Vidya Somashekarappa, and Christine Howes. 2021. Looking for laughs: Gaze interaction with laughter pragmatics and coordination. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, ICMI '21, page 636–644, New York, NY, USA. Association for Computing Machinery.

Marc Mehu. 2011. Smiling and laughter in naturally occurring dyadic interactions: Relationship to conversation, body contacts, and displacement activities. *Human Ethology Bulletin*, 26(1):10–28.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 3111–3119.

Gregory Mills and Pat Healey. 2008. Semantic negotiation in dialogue: The mechanisms of alignment. *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, pages 46–53.

Abhijit Mishra and Pushpak Bhattacharyya. 2018. *Scanpath Complexity: Modeling Reading/Annotation Effort Using Gaze Information: An Investigation Based on Eye-tracking*, pages 77–98. Springer, Singapore.

Margaret Mitchell, Xufeng Han, Jesse Dodge, Alyssa Mensch, Amit Goyal, Alex Berg, Kota Yamaguchi, Tamara Berg, Karl Stratos, and Hal Daumé III. 2012. Midge: Generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 747–756. Association for Computational Linguistics.

Richard Montague. 1973. The Proper Treatment of Quantification in Ordinary English. In Jaakko Hintikka, Julius Moravcsik, and Patrick Suppes, editors, *Approaches to Natural Language: Proceedings of the 1970 Stanford Workshop on Grammar and Semantics*, pages 247–270. D. Reidel Publishing Company, Dordrecht.

Raymond J. Mooney. 2008. Learning to connect language and perception. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 3*, AAAI'08, page 1598–1601. AAAI Press.

Laura Morett, Sarah Hughes Berheim, and Raymond Bulger. 2020. Semantic relationships between representational gestures and their lexical affiliates are evaluated similarly for speech and text. *Frontiers in Psychology*, 11.

Jenny Myrendal. 2015. *Word Meaning Negotiation in Online Discussion Forum Communication*. PhD Thesis, University of Gothenburg, University of Gothenburg.

Bill Noble, Asad Sayeed, Raquel Fernández, and Staffan Larsson. 2021. Semantic shift in social networks. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 26–37, Online. Association for Computational Linguistics.

Kerstin Norén and Per Linell. 2007. Meaning potentials and the interaction between lexis and contexts: An empirical substantiation. *Pragmatics*, 17(3):387–416.

Keith Oatley and P.N. Johnson-Laird. 2014. Cognitive approaches to emotions. *Trends in Cognitive Sciences*, 18(3):134–140.

Robert Östling and Jörg Tiedemann. 2017. Continuous multilinguality with language vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 644–649, Valencia, Spain. Association for Computational Linguistics.

Barbara H. Partee, editor. 1976. *Montague Grammar*. Academic Press.

Hermann Paul. 1891. *Principles of the History of Language*. London ; New York : Longmans, Green.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics.

Volha Petukhova and Harry Bunt. 2009. Grounding by nodding. In *Proceedings of GESPIN, Conference on Gestures and Speech in Interaction, Poznán*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Bernard G Prusak. 2006. The science of laughter: Helmuth plessner's laughing and crying revisited. *Continental philosophy review*, 38:41–69.

James Pustejovsky and Nikhil Krishnaswamy. 2020. Situated meaning in multimodal dialogue: Human-robot and human-computer interactions. *Revue TAL*, 61(3):17.

Arnau Ramisa, Josiah Wang, Ying Lu, Emmanuel Dellandrea, Francesc Moreno-Noguer, and Robert Gaizauskas. 2015. Combining geometric, textual and visual features for predicting prepositions in image descriptions. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 214–220, Lisbon, Portugal. Association for Computational Linguistics.

Terry Regier. 1996. *The human semantic potential spatial language and constrained connectionism*. Neural network modeling and connectionism. MIT Press, Cambridge.

Uwe Reyle. 1993. Dealing with ambiguities by underspecification: Construction, representation and deduction. *Journal of Semantics*, 10(2):123–179.

Alex Rosenfeld and Katrin Erk. 2018. Deep Neural Models of Semantic Shift. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 474–484, New Orleans, Louisiana. Association for Computational Linguistics.

Fereshteh Sadeghi, Santosh K Kumar Divvala, and Ali Farhadi. 2015. Viske: Visual knowledge extraction

and question answering by visual verification of relation phrases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1456–1464.

Magnus Sahlgren. 2006. *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Ph.D. thesis, Institutionen för lingvistik.

Klaus R Scherer. 2009. The dynamic architecture of emotion: Evidence for the component process model. *Cognition and emotion*, 23(7):1307–1351.

Vidya Somashekarappa, Christine Howes, and Asad Sayeed. 2020. An annotation approach for social and referential gaze in dialogue. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 759–765, Marseille, France. European Language Resources Association.

Vidya Somashekarappa, Christine Howes, and Asad Sayeed. 2021. A deep gaze into social and referential interaction. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 43(43).

Michael Spivey, Daniel Richardson, Melinda Tyler, and Ezekiel E Young. 2000. Eye movements during comprehension of spoken descriptions. In *Proceedings of the 22nd Annual Meeting of the Cognitive Science Society*.

Robert Stalnaker. 2002. Common Ground. *Linguistics and Philosophy*, 25(5-6):701–721.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2018. Survey of Computational Approaches to Diachronic Conceptual Change. *arXiv:1811.06278 [cs]*.

Xuri Tang. 2018. A state-of-the-art of semantic change computation. *Natural Language Engineering*, 24(5):649–676.

Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew Walter, Ashis Banerjee, Seth Teller, and Nicholas Roy. 2011. Understanding natural language commands for robotic navigation and mobile manipulation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 25(1).

Joshua B. Tenenbaum. 2020. Cognitive and computational building blocks for morehuman-like language in machines. Acl 2020 keynote, Center for Brains, Minds and Machines, MIT.

David R Traum. 1994. A computational theory of grounding in natural language conversation. Technical report, Rochester Univ NY Dept of Computer Science.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator.

Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. Guesswhat?! visual object discovery through multi-modal dialogue.

Ivan Vulic, Simon Baker, Edoardo Maria Ponti, Ulla Petti, Ira Leviant, Kelly Wing, Olga Majewska, Eden Bar, Matt Malone, Thierry Poibeau, Roi Reichart, and Anna Korhonen. 2020. Multi-simlex: A large-scale evaluation of multilingual and cross-lingual lexical semantic similarity. *CoRR*, abs/2003.04866.

David Wilkins. 2006. Adam kendon (2004). gesture: Visible action as utterance. *Gesture*, 6.

Roel Willems, Asli Özyürek, and Peter Hagoort. 2007. When language meets action: The neural integration of gesture and speech. *Cerebral cortex (New York, N.Y. : 1991)*, 17:2322–33.

Ludwig Wittgenstein. 1953. *Philosophical Investigations*. Basil Blackwell, Oxford.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2048–2057, Lille, France. PMLR.