Structural and Global Features for Comparing Semantic Representation Formalisms

Siyana Pavlova, Maxime Amblard, Bruno Guillaume
Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
{firstname.lastname}@loria.fr

Abstract

The area of designing semantic/meaning representations is a dynamic one with new formalisms and extensions being proposed continuously. It may be challenging for users of semantic representations to select the relevant formalism for their purpose or for newcomers to the field to select the features they want to represent in a new formalism. In this paper, we propose a set of structural and global features to consider when designing formalisms, and against which formalisms can be compared. We also propose a sample comparison of a number of existing formalisms across the selected features, complemented by a more entailment-oriented comparison on the semantic phenomena of the FraCaS corpus.

1 Introduction

Over the past decades, various semantic representation formalisms have emerged, focusing on different features of semantics. New formalisms and extensions are continuously developed, highlighting a dynamic field, but few works have been carried out on their comparison. Abend and Rappoport (2017) provide a high-level summary of semantic features and existing formalisms. Žabokrtský et al. (2020) provide an overview and comparison of eleven deep-syntactic graph-based formalisms, focusing largely on their formal graph features. Insights into the difference between encoding some semantic phenomena in different formalisms can also be found in empirical work based on rule-based (Hershcovich et al., 2020; Pavlova et al., 2022) and machine learning (Kuznetsov and Gurevych, 2020; Wu et al., 2021; Prange, 2022) techniques.

Our goal is to provide a theoretical overview of various features of semantics and what choices are available for including them in the design of a new semantic representation formalism. The set of features can also serve for comparing different formalisms. In this spirit, we present some existing formalisms[1] and compare them against the outlined features. For a more entailment-balanced view and an empirical comparison, we also compare these formalisms against Cooper et al. (1994)'s FraCaS corpus. We focus on sentence-level semantics, but provide a short discussion on multi-sentence awareness for semantic representation formalisms.

The rest of the paper is organised as follows: in §2, §3 and §4 we present some global and structural features to be taken into consideration when comparing or designing a semantic representation formalism. In §5, we briefly present the following formalisms: Conceptual Graphs (CG) (Sowa, 1984), Montague Semantics (MS) (Montague, 1970; Montague et al., 1970; Montague, 1973), Discourse Representation Theory (DRT) (Kamp and Reyle, 1993), Minimal Recursion Semantics (MRS) (Copestake et al., 2005), Abstract Meaning Representation (AMR) (Banarescu et al., 2013), Universal Conceptual Cognitive Annotation (UCCA) (Abend and Rappoport, 2013), Universal Decompositional Semantics (UDS) (White et al., 2016), and Uniform Meaning Representation (UMR) (Van Gysel et al., 2021). In §6, we compare the formalisms across the selected features, and FraCaS features.

2 Pre-Semantics Issues

In this section, we outline a few aspects that we consider to not be constituent parts of what a semantic representation formalism (hereupon referred to as "formalism") is as such, but put it in a more global perspective and are nonetheless important to consider when designing one.

Scalability. Semantic representation formalisms vary in terms of complexity and expressive power.

[1] The list is not exhaustive, though we have attempted to cover a wide range of families.

Proceedings of the 4th International Workshop on Designing Meaning Representations, pages 1–12
June, 2023. ©2023 Association for Computational Linguistics

While more complex ones may be more robust and encode a wider range of phenomena, complexity is negatively correlated to readability and therefore scalability. To be able to make use of various Machine Learning methods for parsing and generation, we need large amounts of manually (or at least semi-manually) annotated data. However, if the representation formalism is more complex, the skills required for annotation are more specialised. This makes the pool of potential annotators smaller and requires a longer and more in-depth training. A balance is necessary to ensure that a formalism covers a wide range of phenomena, yet it is not overly complex in order to keep the threshold for annotators relatively low. Alternatively, formalisms could propose lattices (Van Gysel et al., 2019)[2] for different phenomena, similarly to what UMR does. This would allow for a more coarse-grained annotation by less specialised annotators, and a more in-depth one by more specialised annotators.

This balance is also beneficial for analysis and comparison of different formalisms, as getting comfortable with reading and interpreting the representations is more straightforward.

Universality. The general intuition when talking about multi-linguality, is that meaning is preserved across languages. Thus, the semantic representation for a given text should be the same in all languages. In reality, many semantic representation formalisms are built upon the syntactic structure of sentences, which can differ greatly, especially between pairs of languages from distant families[3].

Furthermore, syntax-based semantic formalisms (and syntax-agnostic ones too) have historically been developed with well-resourced languages (mostly Indo-European, and in particular English) in mind. Thus, formalisms are likely skewed towards better representing phenomena that occur in those languages, and might even miss phenomena that do not appear in them.

Finally, similarly to scalability, using lattices could be beneficial for universality as the same semantic phenomenon may contain a more fine-grained set of categories in some languages than in others. Using lattices allows smoother annotation in different languages, while still keeping the possibility for cross-language comparison.

Unicity. Unicity addresses whether a formalism has a unique representation for a given meaning. In AMR inverting the direction of relations changes the focus and thus the meaning of the representation. On the other hand, if we take the formal logic representation of a sentence containing negation and conjunction, and apply De Morgan's Laws, we end up with a different, but logically equivalent representation. If equivalent representations are allowed for a certain formalism, it may be necessary to establish what constitutes the canonical form and how members of an equivalence class relate to it.

Flavor. We use the term "flavor" as used by Koller et al. (2019) in relation to the level of abstraction from surface form for graph-based semantic representations. Koller et al. (2019) define three levels: flavor 0 – a one to one correspondence between graph nodes and surface tokens; flavor 1 – all tokens are present as nodes, but there are additional nodes in the graph too; flavor 2 – not necessarily all tokens correspond to nodes, and there may be nodes that do not correspond to a specific token.

Use of lexical resources. Some formalisms rely on lexical resources for predicate and concept senses, or argument structure of predicates. This works well for languages where these resources already exist and are well-developed. However, for languages where this is not the case, there may be the need to produce them in parallel with producing annotations for the formalism, like the creators of UMR propose (Van Gysel et al., 2021). While viable, this makes the process longer and more complex and should be taken into consideration for the design. It is also tied to the *Universality* aspect: in order to enable cross-language comparison, for formalisms that do use lexical resources, there needs to be a link between said resources for multiple languages. While efforts exist in this direction (Bond and Foster, 2013; Bond et al., 2020), for most languages, this link is not there yet. Ideally, when creating datasets for a new language, the linking to other languages can be created in the process too. This, again, entails more effort, but we believe the cost of that is worth the benefits of having a more complete resource.

3 Semantic Features

In this section we discuss aspects of semantics that constitute what a semantic representation is.

Predicate-argument structure. The most

[2] E.g. *number* can be coarsely annotated as *singular* or *non-singular*, while the latter can be further broken down into *paucal* and *plural* where the distinction exists.

[3] If we assume that it is possible for two representations to have the same meaning, then the different representations stemming from different underlying syntactic structures in different languages would be less of an issue.

prominent feature of many formalisms is that they are centered around the predicate-argument structure of the events occurring in a sentence. Events are usually represented as predicates that take a certain number and kind of arguments. The relation between a predicate and an argument is expressed via a semantic role, which can be predicate specific (in the spirit of PropBank (Palmer et al., 2005)) or from a generic closed set (like VerbNet (Kipper et al., 2008)), with varying granularity.

Practical issues here arise from the fact that different formalisms use different lexical resources, making comparison and transformation more challenging. For English, work has been done to align (Palmer, 2009) and continue to improve the alignment (Stowe et al., 2021) of these resources. However, English is one of few languages where lexical resources are comparatively well developed. Thus, the use of language-specific frames for predicates comes with the cost of developing such resources. This is an argument against their use and for adopting methods that do not encode such senses, making the *Universality* point more easily attainable, similar to what UCCA does (Abend and Rappoport, 2013). That may, however, make the formalism less expressive.

Temporality. Temporal information deals with when an event occurs. We consider two aspects of this - when it occurrs relative to other events in the text, and when it occurred relative to the moment of speaking. Temporal information can be encoded in a variety of ways – via grammatical tense, from the lexicon with certain adverbs, or specific words or phrases, or may even be implicit. Combined with the fact that different languages have a stronger preference for some approaches over others, the task of encoding it is challenging. Formalisms need to decide whether temporal information will be encoded at all, and whether all kinds, that is, whether grammatical tense will be considered or only information present on the surface.

Aspect. Complementary to grammatical tense, grammatical aspect expresses how an event develops over time – whether it is one-time, whether it is continuous, whether it ended or is still ongoing. Here, again, formalisms have a choice – whether to encode aspect, and which features of it.

Spatial information. As Abend and Rappoport (2017) point out, spatial information in semantics is considered mainly for domains such as geographical information systems and robotics navigation. From a more theoretical perspective, we consider the resolution/interpretation of location-related deictics (*here*/*there*) and demonstrative pronouns to be an important aspect of the representation.

Encoding spatial information is especially relevant for sign languages, where its semantics is richer. For example, the handshape can express a distinction in an object's shape (e.g. curved or flat object) (Supalla, 1986), and the orientation of the handshape can express an object's orientation (Brozdowski et al., 2019).

Reification. Reification in semantics is the process of transforming events, actions and concepts so that they are expressed with (quantifiable) variables. This facilitates the translation of the so transformed representation into first-order logic and is therefore an important consideration if we want to use a formalism for logical inference.

Scope. The scope of semantic operators (such as those of quantification or negation) shows to which entities or events that operator applies. Some formalisms choose to not encode scope at all, making consistent logical inference impossible.

Scope does not directly relate to word order, which gives rise to *scope ambiguity* – a single sentence containing more than one scope operator can be interpreted in more than one way depending on how the operators combine. In case of scope ambiguity, the question for formalisms is whether to force a specific interpretation or to leave the representation underspecified. The latter allows that restrictions are added at a later stage when the correct interpretation becomes obvious from the context.

Negation. Negation, similarly to many of the other phenomena, can be expressed in different ways – overtly as a separate token, or as a morpheme of a token, presenting a challenge of whether to encode the two in the same way. We believe that meaning-wise, they should be equivalent and semantic representations should abstract away from the difference between the two. This is especially important when we consider logical inference and scope. Indeed, this is what many of the formalisms in section 5 do. There are some exceptions, notably UCCA, where, for example, the phrases "not clear" and "unclear" would be encoded differently despite having the same meaning.

Modality. Modality is used to express the reality of an event: realis – whether is it actually realised, or irrealis – whether it is a possibility or necessity. Modal expressions are often expressed on the surface as modal auxiliaries, adverbs or adjectives. They get special treatment for most formalisms, be it as specially dedicated predicates (as in AMR), or operators between boxes (as in some realisations of DRT).

Modal expressions are also categorised in *flavors* (different from the flavor we discussed in section 2), showing how the possibility under discussion is linked to reality. *Epistemic* flavor covers possibilities based on some knowledge or belief, while *deontic* flavor expresses that the possibility is in accordance with what is required in reality.

Evidentiality is a phenomenon that encodes the type of evidence the speaker has for a statement. For example, one may differentiate between the speaker having direct (e.g. visual) or reportative (e.g. having heard about it and merely repeating what they have been told) evidence. In most languages this is expressed lexically with specific phrases (such as "reportedly" in English). However, in about a quarter of the world's languages, these differences are expressed grammatically (Aikhenvald, 2004), which formalisms do not address.

Logical inference. If we want to be able to use a semantic representation for reasoning, it is important that the formalism used permits logical inference. Not all formalisms are equally well equipped for this. For example, as Bos (2020) points out, with AMR, we are able to draw inferences, as long as there is no negation. That is, we can infer "it rained" from "it rained heavily", but we can also infer "it rained" from "it didn't rain". According to Bos (2020), this is due to negation in AMR being expressed as a predicate rather than an operator that takes scope. This highlights the importance of formalisms expressing scope-relevant phenomena (such a quantification and negation) in the appropriate way if we want to permit logical inference.

4 Semantics Interfaces

In this section we outline structural features of formalisms that are linked to their applications or to interfaces of semantics with syntax and pragmatics.

Generation and Analysis. When designing a new formalism, it is worth considering whether there are specific intended uses and applications for the formalism. Some tasks may rely more on parsing or on generation, so it is important to consider whether there are aspects that can be encoded into the design of the formalism to make parsing and/or generation more robust.

A lot of effort has gone into the parsing of text into various semantic representations as the amount of works on the topic suggests (Oepen et al., 2019, 2020). Challenges for parsing may come from the various types of ambiguities (e.g. lexical, scope) and, if we assume equivalent representations, which one to produce.

Similarly to parsing, generation from meaning representations has gathered much attention (Ribeiro et al., 2021; Hajdik et al., 2019). When keeping track of word order as part of the representation and without lemmatizing or otherwise modifying the original tokens, "generation" from semantic representation is straightforward for formalisms of flavor 1[4]. On the other hand, generation is a more interesting problem when working with flavor 2 formalisms, where the question is what to generate for a structure which may have more than one interpretation within the formalism.

Evaluation. For parsing, for most formalisms there are established methods for evaluating the produced representation against a gold one (Cai and Knight, 2013; Hershcovich et al., 2017; Oepen et al., 2014)[5]. Regardless, there are difficulties when using lexical resources and there is ongoing work on how to score closely related (but not perfectly overlapping) concepts in the representation (Opitz et al., 2020). Finally, if a formalism allows for multiple equivalent representations, similarity metrics will need to take this into account when evaluating a representation that is not in the canonical form for its equivalence class.

Evaluating generation is not straightforward when we consider that for some flavor 2 formalisms, many sentences have the same representation (e.g. AMR does not encode tense, so "I went to Paris" and "I will go to Paris" have the same representation). In such cases, it is necessary to consider whether it is enough to generate only one of the correct sentences in order to consider the process successful, or we need all the possible ones. Paraphrases pose a further issue, as they may have a (nearly) identical meaning to the original sentence, but look very different on the surface, with paraphrase evaluation being a subfield in its own right (Shen et al., 2022).

Compositionality. The meaning of a sentence

[4] Still, if the representation was automatically produced, the process may not be as direct.

[5] These metrics are also often used to compute inter-annotator agreement for manual annotation.

(or a phrase) is generally thought to be a function of the meanings of its composite parts. Historically, producing the semantic representation for a given sentence has passed through the syntactic one first, necessarily making compositionality a feature of the final representation. The Machine Learning revolution, however, has enabled the parsing of text directly into a semantic representation, rather than relying on the syntax-semantics interface, thus many of the newer formalisms have a choice to make about whether to preserve compositionality as a feature of the design.

A broader question is whether we consider the semantic representation to be only the final structure (e.g. graph or logical formula) that we obtain, or also the process of building that structure (in the spirit of derivation vs. derived trees for TAGs (Joshi et al., 1975)). If we take the latter view, then, necessarily, compositionality becomes a core aspect of the representation. We note, however, that this adds additional complexity to the annotation process, especially if the syntactic structure is not used as an underlying component.

Syntax-semantics interface. As mentioned above, semantic representation formalisms were built in such a way that the semantic structure of a sentence can be constructed from its syntactic one. Many of the newer formalisms are syntax-independent. While the first method may work for well-resourced languages with developed grammars, the latter one might be more beneficial for languages where these resources do not exist. This ties to the *Universality* point.

Multi-sentence. Many formalisms focus on representing sentences but do not necessarily employ means to go beyond the sentence boundary. The considerations we describe here can appear within a single sentence too, but are frequently seen when dealing with multiple sentences, namely anaphora and co-reference resolution, and the representation of discourse markers and relations.

When it comes to anaphora and co-reference, formalisms may choose to annotate the referents with the same variable, or with different ones. In the latter case, they may choose to employ a way to indicate that the variables refer to the same object or not do so. We note here the interesting case of AMR which includes a way, albeit somewhat superficial, to encode multiple sentences in the same representation. For referents occurring in the same sentence, AMR uses the same variable, but different ones when they occur in different sentences. Finally, similarly to scope ambiguity, formalisms need to take into consideration anaphoric ambiguity (in "John told Tom his brother left." it is ambiguous who "his" refers to) – whether to select one of the options, produce all different version, or leave the representation underspecified.

When treating discourse markers, formalisms have the choice to represent them in the same way as other relations, or give them a special status, thus adding a layer that sits on the boundary between semantics and pragmatics.

Questions. A distinction is usually made between Wh-questions and yes/no questions. For Wh-questions, a common approach is to maintain the structure of a declarative sentence and introduce a special concept or symbol (e.g. *amr-unknown* in AMR) to put in place of the entity or predicate that is being asked about. Yes/no questions usually need an additional relation to indicate that the whole statement is a question. It is interesting to note that in the case of DRS (at least in the version implemented in the Parallel Meaning Bank (Abzianidze et al., 2017), yes/no questions are ignored altogether and annotated in the same way as their declarative counterparts. This can be explained with the fact that DRT is designed to deal with discourse, as opposed to dialogue, and takes the stance that questions are only part of the latter.

5 Semantic formalisms

In this section, we describe existing formalisms and their core features, with a more exhaustive comparison in section 6. We strongly believe in the benefits of data-driven analysis and comparison. Therefore, if existent beyond a toy-corpus size, we also point to existing datasets.

Various extensions have been proposed for many of the formalisms. However, we do not know, for every extension, how it combines with the other ones and whether it does not interfere with the properties we explore. For example, adding scope interferes with compositionality. Thus, for this study we work with the original formalism, unless the extensions have been combined in a standalone one (as is the case with UMR).

Montague Semantics (MS) (Montague et al., 1970; Montague, 1970, 1973) introduced mathematical methods, namely higher-order predicate logic and lambda calculus, to semantics. Its core features are the use of model theoretic semantics, and compositionality.

Conceptual Graphs (CG) (Sowa, 1984) are based on semantic networks and C.S. Peirce's existential graphs. Aside from natural language semantics, CGs have also been influential in knowledge representation. CGs' most apparent difference from modern semantic graphs is that they encode all events, entities, and relations as nodes, whereas edges are unmarked. Original CGs do not encode scope, but later versions provide that, along with ways to deal with temporal and modal logic (Sowa, 2003, 2006) and work has been done to combine CGs with generalized quantifiers (Cao, 2001).

Discourse Representation Theory (DRT) (Kamp and Reyle, 1993) is a "dynamic semantics" formalism, i.e. the meaning of a sentence is considered with respect to its potential to update context. It was designed to deal with anaphora and tense, but has since evolved to treat other semantic aspects such as presupposition, and propositional attitudes. DRT expressions are called Discourse Representation Structures (DRS). They are usually presented as nested boxes, but those can be transformed into graphs (Abzianidze et al., 2020). The Parallel Mearning Bank (PMB) (Abzianidze et al., 2017) is a large DRS corpus with gold annotations in English, German, Italian and Dutch.

Minimal Recursion Semantics (MRS) (Copestake et al., 2005) is a formalism from the Head-driven Phrase Structure Grammar (HPSG) (Pollard and Sag, 1994) family. As such, it has a strong link to syntax, but is also meant to be universal. MRS annotates a large range of phenomena, but is a rather complex formalism for annotators without linguistic knowledge. A distinguishing feature is its underspecifiability, which allows for encoding scope ambiguity. A medium-sized parallel dataset has been annotated for 15 languages[6]

Abstract Meaning Representation (AMR) (Banarescu et al., 2013) is meant to be a simple formalism to increase the ease of annotation. This is achieved by ignoring features such as tense, plurality and definiteness. AMR's core is the predicate-argument structure of events, with additional non-core roles specified for predicate-independent relations. For English, AMR relies on PropBank (Palmer et al., 2005) for predicate senses and semantic roles. A multitude of extensions have been proposed for AMR for various aspects such as tense (Donatelli et al., 2018), scope (Pustejovsky et al., 2019; Bos, 2020), spatial information (Bonn et al., 2020), multi-sentence information (O'Gorman et al., 2018). Despite its being designed with English in mind, AMR has also been used for Chinese, Czech, and Korean, among others. Larger corpora are available for English under a paid license, smaller ones are freely available[7].

Universal Conceptual Cognitive Annotation (UCCA) (Abend and Rappoport, 2013) is likewise designed to be simple for annotators. UCCA's *Foundational Layer (FL)* uses a set of 14 broad semantic role categories (e.g. `Participant`, `Adverbial`) and does not rely on lexical resources. The latter point makes it easier to adopt for multiple languages. Extension layers exist for UCCA that deal with semantic roles (Shalev et al., 2019; Prange et al., 2019a), co-reference (Prange et al., 2019b) and implicit arguments (Cui and Hershcovich, 2020). There are datasets for the FL for English, German, French, Hebrew and Russian.[8]

Universal Decompositional Semantics (UDS) (White et al., 2016) adds a number of semantic layers on top of the syntactic Universal Dependencies (UD)[9]. UDS follows the principle of decomposition, e.g. for semantic roles, they take Dowty (1991)'s view on decomposing notions such as `Agent` into finer properties like `volition` and `awareness`, allowing a single predicate to be assigned multiple of these categories. The currently existing layers address semantic roles; irrealis vs realis distinction on events; predicate senses and entity types; genericity; and duration and relative order of events. Annotated datasets are available for English[10].

Uniform Meaning Representation (UMR) (Van Gysel et al., 2021) is a proposal that extends AMR with aspect, temporal information, scope, co-reference and modal dependencies. UMR takes into account the morphosyntactic differences between languages and, to the best of our knowledge, is the first formalism to propose concrete steps on how to proceed with annotation for low-resource languages.

[6] `https://github.com/delph-in/docs/wiki/MatrixMrsTestSuite`

[7] `https://amr.isi.edu/download.html`

[8] `https://github.com/UniversalConceptualCognitiveAnnotation`

[9] `https://universaldependencies.org/`

[10] `http://decomp.io/data/`

6 Comparison and Discussion

In this section, we compare the formalisms from section 5 across the features outlined in §2, §3 and §4, as well as the phenomena covered by the FraCaS corpus (Cooper et al., 1994).

6.1 Feature Comparison

Table 1 provides an overview of how the frameworks compare across the features described in §2, §3 and §4 with the following exceptions: we add rows for dataset size and for the number of languages in which data is available, as these can be indicative of *Scalability* and *Universality* respectively; we add a row to show whether a formalism leaves the representation underspecified or not in case of *scope ambiguity*; we consider *Generation* and *Analysis* separately; we omit *Evaluation*, because while it is an important aspect to talk about, evaluation metrics are not formalism specific (e.g. Smatch (Cai and Knight, 2013) is typically associated with AMR, but can be used to evaluate any graph-based formalism).

Table 1 should be read as follows: for most features we indicate whether a formalism encodes it (✔) or not (✗). Dataset size is divided into three categories: toy (< 100 sentences for any language), medium (between 100 and $1,000$ sentences for at least one language), large ($> 1,000$ sentences for at least one language). For predicate-argument structure, we indicate whether semantic roles are predicate-specific or generic. For *Temporal* and *Evidentiality* we distinguish three categories: (0) not encoded with a dedicated structure / relation type; (1) encoded with a dedicated structure, but only if present on the surface, and not when grammatical; (2) encoded in all cases. For *Negation* and *Modality*, we distinguish three categories: (0) not encoded; (1) encoded, but without scope; (2) encoded with scope. For *Questions*, we distinguish between: (0) not encoded at all, (1) only wh-questions are encoded, or (2) all questions are encoded.

From the table, we can see that MRS is the most expressive formalism across the chosen features. However, this comes at the cost of it being complex to annotate, making it scale poorly. Similarly, MS and CG require some specialised knowledge for annotation and do not scale well, but while MS is close to MRS in terms of expressive power, CG lags behind. Original DRT, likewise, requires some specialised knowledge for annotation. However, recent work on simplifying the representation (Bos, 2021) and the existence of a large corpus (Abzianidze et al., 2017) lead us to consider DRT scalable. On the scalable side are also the newer formalisms, which have been designed for ease of annotation. However, for AMR, UCCA and UDS this means that they are not well-equipped to encode many of the semantic features we consider. For AMR and UCCA, extensions exist to address some of these issues. UDS, being a layered formalism, with new layers being added continuously, also has the potential to address the missing aspects. Finally, we take a look at UMR, which incorporates many of the proposed extensions of AMR, while preserving the latter's features. As we can see from the table, UMR is almost as expressive as MRS and DRT, while remaining syntax-independent, which its creators consider to be a strong point for scalability.

Looking across the features, we can notice that all formalisms can be used for *Generation* and *Analysis*. However, they all lack tools to deal with spatial information, especially the kind that is present in sign languages. Similarly, grammatically-expressed evidentiality is not annotated by any formalism. This opens a broader discussion regarding the encoding of features which are expressed only grammatically. We notice that surface information tends to be encoded, while for certain phenomena the grammatical side is ignored altogether. Thus, there is the risk of under-representing grammatical phenomena that are more prevalent in low-resource languages, but not in the well-resourced languages used as the basis for the design of formalisms.

6.2 FraCaS Comparison

FraCaS (Cooper et al., 1994) is a corpus of 346 textual inference problems, each consisting of one to five premises and a hypothesis. For each example, it is indicated whether it is true, false or unknown that the hypothesis follows from the premises. The problems are split into nine categories, relating to semantics (leftmost column of Table 2), however their distribution is not uniform. Some work on evaluating formalisms against FraCaS can be found in (Abzianidze, 2016; Haruta et al., 2019).

In Table 2, we provide a high-level comparison across the phenomena present in the FraCaS corpus. The table shows whether a formalism should be able to encode all (✔), at least half but not all (0.5), or less than half (✗) of the examples for a phenomenon. We want to highlight that this is

	MS	CG	DRT	MRS	AMR	UCCA	UDS	UMR
Scalability	✗	✗	✔[*]	✗	✔	✔	✔	✔[†]
Datasets (size)	toy	toy	large	medium	large	large	large	toy[†]
Universality	✔	✔	✔	✔	✗[‡]	✔	✔	✔[†]
Datasets (# languages)	-	-	4	>10	>6	6	1	-[†]
Unicity	✗	✗	✗	✗	✔	✔	✔	✗
Flavor	2	2	2	2	2	1	1	2
Lexical Resources	✗	✗	✔	✗	✔	✗	✗	✔
Pred-arg	generic	generic	generic	generic	specific	generic	generic	specific
Temporality	0	0	2	2	1	1	2[¶]	2
Aspect	✗	✗	✗	✔	✗	✗	✗	✔
Spatial	✗	✗	✗	✗	✗	✗	✗	✗
Reification	✔	✗	✔	✔	✔	✗	✗	✔
Scope	✔	✗	✔	✔	✗	✗	✗	✔
Scope ambiguity	✗	✗	✗	✔	✗	✗	✗	✗
Negation	2	1	2	2	1	1	1	2
Modality	2	1	2	2	1	1	1	2
Evidentiality	1	1	1	1	1	1	1	1
Logical Inference	✔	✗	✔	✔	✗	✗	✗	✔
Generation	✔	✔	✔	✔	✔	✔	✔	✔
Analysis	✔	✔	✔	✔	✔	✔	✔	✔
Compositionality	✔	✔	✗	✔	✔	✔	✔	✗
SSI	✔	✗	✔	✔	✗	✗	✔	✗
Multi-sentence	✗	✔	✔	✗	✗	✗	✗	✗
Questions	0	0	1	2	2	0	0	2

Table 1: Feature comparison. ✔ = yes, ✗ = no.
Temporal, *Evidentiality*: 1 = only surface, but not grammatical; 2 = yes.
Negation, *Modality*: 1 = encoded, but without scope; 2 = encoded with scope.
Questions: 0 = no special way to encode; 1 = only wh-questions encoded; 2 = all types of questions encoded.
[*] Original DRT requires some specialised knowledge, but given the recent proposal for simplification (Bos, 2021) and the existence of a large annotated corpus, we consider it scalable;
[†] UMR is designed with scalability and universality in mind, but it is a young formalism and both aspects remain to be verified;
[‡] AMR does not claim to be universal, but corpora have been made available in a variety of languages.
[¶] UDS encodes duration and relative occurrence of events, but does not specify when an event occured relative to the moment of utterance.

		MS	CG	DRT	MRS	AMR	UCCA	UDS	UMR
Quantifiers	23%	✔	✗	✔	✔	✗	✗	✗	0.5
Plurals	10%	✔	✗	✗	✔	✗	✗	✗	✗
Anaphora	8%	✗	✔	✔	✔	✗	✗	✗	✗
Ellipsis	16%	✗	✔	✔	✔	✗	✗	✗	✗
Adjectives	7%	✗	✗	✗	✗	✗	✗	✗	✗
Comparatives	9%	0.5	✗	0.5	0.5	✗	✗	✗	✗
Temporal	22%	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
Verbs	2%	✗	✗	✗	0.5	✗	✗	0.5	0.5
Attitudes	4%	✔	✔	✗	✗	✗	✗	✔	✗
Total	100%	$\geq 52.5\%$	$\geq 39\%$	$\geq 62.5\%$	$\geq 73.5\%$	$\geq 11\%$	$\geq 11\%$	$\geq 16\%$	$\geq 23.5\%$

Table 2: Coarse-grained FraCaS comparison. ✔ = the formalism should be able to cover all examples for that feature; 0.5 = the formalism should be able to cover at least half, but not all examples for that features; ✗ = the formalism can cover less than half of the examples for that feature.

a very coarse-grained comparison[11] meant to balance the one in subsection 6.1 in providing a more entailment-based view and an empirical comparison of the formalisms. It should serve as a starting point for a more detailed, sentence-by-sentence comparison on this and other corpora.

We have made a few assumptions before our decision-making process. Where multi-sentence capabilities are relevant, namely for *Anaphora* and *Ellipsis*, we have taken the formalisms' ability to encode that into account. For categories where the focus is not on multi-sentence capabilities, we assume a conjunction of the premises to give a fairer chance to formalisms that only deal with single sentences. Table 2 is split in two parts, the lower one highlighting the sections where lexical information is necessary to resolve some of the examples. In such cases, we have taken the conservative view that formalisms are unable to encode the example. If we assume that with the help of the lexical resource the hypothesis can be deemed true, false or unknown, then the estimates for the lower part of the table would be higher. In what follows, we highlight the challenging areas in each category.

Quantifiers. For full coverage here, a formalism should be able to deal with scope, but also with definiteness, which is why while UMR covers scope, it cannot cover all examples in this section.

Plurals. While the majority of formalisms should be able to cover many of the *Conjoined Noun Phrases* examples, most will struggle with some of the bigger subsections, namely *Bare Plurals* and *Definite Plurals* due to inability to encode distinctions in definiteness.

Anaphora. While most formalisms can cover intra-sentential anaphora, for full coverage, they need to be able to also deal with inter-sentential one, which constitutes the larger part of this section.

Ellipsis. Similarly, if the ellipsis is in the same sentence, most formalisms perform well. However, since most examples in this section use multiple premises, only the formalisms that can deal with multiple sentences can get to full coverage.

Adjectives. Examples in this category rely heavily on lexical information (e.g. "former" implying that the phrase it is modifying is not necessarily up-to-date) and even some world knowledge (knowing that a "small elephant" is larger than a "large mouse"). Thus, none of the formalisms are equipped to deal with the majority of examples.

Comparatives. Two main difficulties arise here: similarly to *Adjectives*, lexical information is necessary for some examples, meaning none of the formalisms can reach full coverage. Furthermore, a large portion of the examples use quantification, making the formalisms that do not encode quantifiers well unable to cover even half of the examples.

Temporal. While some FraCaS temporal examples rely on tense or lexical semantics, for many there is temporal information present as separate surface tokens ("before", "for two years", "in 1991"). While most formalisms would be able to deal with these, many examples also include time spans which only UDS is explicitly equipped to encode. A few examples rely on lexical information as well ("started", "lasted", "was over" in example #259) which the formalisms will struggle with.

Verbs. For full coverage here, distinction between tenses, some lexical information, and capabilities to work with time spans are needed. Thus, none of the formalisms can encode all sentences.

Attitudes. To get a full coverage for this part, a formalism needs to either rely on lexical information (to distinguish between "managed to win" and "tried to win", for example) or employ specific ways to address epistemicity within its structure.

From Table 2, our general observation is that MRS, again, is the most expressive formalism, followed by DRT, while AMR, UCCA, UDS and UMR manage to fully encode only a few of the features. We remind the reader again that this is a very coarse-grained study. An in-depth sentence-by-sentence study is necessary to confirm our observations and provide an exact percentage of the FraCaS corpus by various formalisms.

7 Conclusion

In this paper we proposed a set of structural and global features to use when comparing semantic representation formalisms. We hope this set of features can be helpful for the community, both in the design of new formalisms and extensions, and in the selection of formalisms to use for specific tasks. The list of features is by no means complete, and extending it as well as the number of formalism can be the subject of future works. Similarly, we believe a more fine-grained study on the expressivity of formalisms with respect to the FraCaS corpus would be beneficial for the community.

[11] E.g. ✗ in the *Anaphora* row is different for UDS, which does not encode anaphora at all, and AMR, which encodes only intra-sentential anaphora, but still does not cover at least 50% of the *Anaphora* examples of FraCaS.

Acknowledgments

We would like to thank the anonymous reviewers for their feedback and comments. Part of this work has been funded by *Agence Nationale de la Recherche* (ANR, fr: National Research Agency), grant number ANR-20-THIA-0010-01.

References

Omri Abend and Ari Rappoport. 2013. Universal Conceptual Cognitive Annotation (UCCA). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 228–238, Sofia, Bulgaria. Association for Computational Linguistics.

Omri Abend and Ari Rappoport. 2017. The state of the art in semantic representation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 77–89, Vancouver, Canada. Association for Computational Linguistics.

Lasha Abzianidze. 2016. Natural solution to FraCaS entailment problems. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 64–74, Berlin, Germany. Association for Computational Linguistics.

Lasha Abzianidze, Johannes Bjerva, Kilian Evang, Hessel Haagsma, Rik van Noord, Pierre Ludmann, Duc-Duy Nguyen, and Johan Bos. 2017. The Parallel Meaning Bank: Towards a multilingual corpus of translations annotated with compositional meaning representations. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 242–247, Valencia, Spain. Association for Computational Linguistics.

Lasha Abzianidze, Johan Bos, and Stephan Oepen. 2020. DRS at MRP 2020: Dressing up discourse representation structures as graphs. In *Proceedings of the CoNLL 2020 Shared Task: Cross-Framework Meaning Representation Parsing*, pages 23–32, Online. Association for Computational Linguistics.

Alexandra Y Aikhenvald. 2004. *Evidentiality*. OUP Oxford.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, pages 178–186.

Francis Bond, Luis Morgado da Costa, Michael Wayne Goodman, John Philip McCrae, and Ahti Lohk. 2020. Some issues with building a multilingual Wordnet. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3189–3197, Marseille, France. European Language Resources Association.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362, Sofia, Bulgaria. Association for Computational Linguistics.

Julia Bonn, Martha Palmer, Zheng Cai, and Kristin Wright-Bettner. 2020. Spatial AMR: Expanded spatial annotation in the context of a grounded Minecraft corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4883–4892, Marseille, France. European Language Resources Association.

Johan Bos. 2020. Separating argument structure from logical structure in AMR. In *Proceedings of the Second International Workshop on Designing Meaning Representations*, pages 13–20, Barcelona Spain (online). Association for Computational Linguistics.

Johan Bos. 2021. Variable-free discourse representation structures. *Semantics Archive*.

Chris Brozdowski, Kristen Secora, and Karen Emmorey. 2019. Assessing the comprehension of spatial perspectives in asl classifier constructions. *The Journal of Deaf Studies and Deaf Education*, 24(3):214–222.

Shu Cai and Kevin Knight. 2013. Smatch: an evaluation metric for semantic feature structures. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 748–752, Sofia, Bulgaria. Association for Computational Linguistics.

Tru H Cao. 2001. Generalized quantifiers and conceptual graphs. In *Conceptual Structures: Broadening the Base: 9th International Conference on Conceptual Structures, ICCS 2001 Stanford, CA, USA, July 30–August 3, 2001 Proceedings 9*, pages 87–100. Springer.

Robin Cooper, Richard Crouch, Jan Van Eijck, Chris Fox, Josef Van Genabith, Jan Jaspers, Hans Kamp, Manfred Pinkal, Massimo Poesio, Stephen Pulman, et al. 1994. Fracas–a framework for computational semantics. *Deliverable D6*.

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

Ruixiang Cui and Daniel Hershcovich. 2020. Refining implicit argument annotation for UCCA. In *Proceedings of the Second International Workshop on Designing Meaning Representations*, pages 41–52, Barcelona Spain (online). Association for Computational Linguistics.

Lucia Donatelli, Michael Regan, William Croft, and Nathan Schneider. 2018. Annotation of tense and aspect semantics for sentential AMR. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 96–108, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

David Dowty. 1991. Thematic proto-roles and argument selection. *language*, 67(3):547–619.

Valerie Hajdik, Jan Buys, Michael Wayne Goodman, and Emily M. Bender. 2019. Neural text generation from rich semantic representations. 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 2259–2266, Minneapolis, Minnesota. Association for Computational Linguistics.

Izumi Haruta, Koji Mineshima, and Daisuke Bekki. 2019. A ccg-based compositional semantics and inference system for comparatives. *arXiv preprint arXiv:1910.00930*.

Daniel Hershcovich, Omri Abend, and Ari Rappoport. 2017. A transition-based directed acyclic graph parser for UCCA. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1127–1138, Vancouver, Canada. Association for Computational Linguistics.

Daniel Hershcovich, Nathan Schneider, Dotan Dvir, Jakob Prange, Miryam de Lhoneux, and Omri Abend. 2020. Comparison by conversion: Reverse-engineering UCCA from syntax and lexical semantics. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2947–2966, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Aravind K Joshi, Leon S Levy, and Masako Takahashi. 1975. Tree adjunct grammars. *Journal of computer and system sciences*, 10(1):136–163.

Hans Kamp and Uwe Reyle. 1993. *From discourse to logic: Introduction to modeltheoretic semantics of natural language, formal logic and discourse representation theory*. Dordrecht. Kluwer.

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2008. A large-scale classification of english verbs. *Language Resources and Evaluation*, 42:21–40.

Alexander Koller, Stephan Oepen, and Weiwei Sun. 2019. Graph-based meaning representations: Design and processing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 6–11, Florence, Italy. Association for Computational Linguistics.

Ilia Kuznetsov and Iryna Gurevych. 2020. A matter of framing: The impact of linguistic formalism on probing results. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 171–182, Online. Association for Computational Linguistics.

Richard Montague. 1970. English as a formal language.

Richard Montague. 1973. The proper treatment of quantification in ordinary english. In *Approaches to natural language: Proceedings of the 1970 Stanford workshop on grammar and semantics*, pages 221–242. Springer.

Richard Montague et al. 1970. Universal grammar. *1974*, pages 222–46.

Stephan Oepen, Omri Abend, Lasha Abzianidze, Johan Bos, Jan Hajic, Daniel Hershcovich, Bin Li, Tim O'Gorman, Nianwen Xue, and Daniel Zeman. 2020. MRP 2020: The second shared task on cross-framework and cross-lingual meaning representation parsing. In *Proceedings of the CoNLL 2020 Shared Task: Cross-Framework Meaning Representation Parsing*, pages 1–22, Online. Association for Computational Linguistics.

Stephan Oepen, Omri Abend, Jan Hajic, Daniel Hershcovich, Marco Kuhlmann, Tim O'Gorman, Nianwen Xue, Jayeol Chun, Milan Straka, and Zdenka Uresova. 2019. MRP 2019: Cross-framework meaning representation parsing. In *Proceedings of the Shared Task on Cross-Framework Meaning Representation Parsing at the 2019 Conference on Natural Language Learning*, pages 1–27, Hong Kong. Association for Computational Linguistics.

Stephan Oepen, Marco Kuhlmann, Yusuke Miyao, Daniel Zeman, Dan Flickinger, Jan Hajič, Angelina Ivanova, and Yi Zhang. 2014. SemEval 2014 task 8: Broad-coverage semantic dependency parsing. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 63–72, Dublin, Ireland. Association for Computational Linguistics.

Tim O'Gorman, Michael Regan, Kira Griffitt, Ulf Hermjakob, Kevin Knight, and Martha Palmer. 2018. AMR beyond the sentence: the multi-sentence AMR corpus. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3693–3702, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Juri Opitz, Letitia Parcalabescu, and Anette Frank. 2020. AMR similarity metrics from principles. *Transactions of the Association for Computational Linguistics*, 8:522–538.

Martha Palmer. 2009. Semlink: Linking propbank, verbnet and framenet. In *Proceedings of the generative lexicon conference*, pages 9–15. GenLex-09, Pisa, Italy.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106.

Siyana Pavlova, Maxime Amblard, and Bruno Guillaume. 2022. How much of UCCA can be predicted from AMR? In *Proceedings of the 18th Joint ACL - ISO Workshop on Interoperable Semantic Annotation within LREC2022*, pages 110–117, Marseille, France. European Language Resources Association.

Carl Pollard and Ivan A Sag. 1994. *Head-driven phrase structure grammar*. University of Chicago Press.

Jakob Prange. 2022. *Neuro-Symbolic Models for Constructing, Comparing, and Combining Syntactic and Semantic Representations*. Ph.D. thesis, Georgetown University.

Jakob Prange, Nathan Schneider, and Omri Abend. 2019a. Made for each other: Broad-coverage semantic structures meet preposition supersenses. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 174–185, Hong Kong, China. Association for Computational Linguistics.

Jakob Prange, Nathan Schneider, and Omri Abend. 2019b. Semantically constrained multilayer annotation: The case of coreference. In *Proceedings of the First International Workshop on Designing Meaning Representations*, pages 164–176, Florence, Italy. Association for Computational Linguistics.

James Pustejovsky, Ken Lai, and Nianwen Xue. 2019. Modeling quantification and scope in Abstract Meaning Representations. In *Proceedings of the First International Workshop on Designing Meaning Representations*, pages 28–33, Florence, Italy. Association for Computational Linguistics.

Leonardo F. R. Ribeiro, Martin Schmitt, Hinrich Schütze, and Iryna Gurevych. 2021. Investigating pretrained language models for graph-to-text generation. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 211–227, Online. Association for Computational Linguistics.

Adi Shalev, Jena D. Hwang, Nathan Schneider, Vivek Srikumar, Omri Abend, and Ari Rappoport. 2019. Preparing SNACS for subjects and objects. In *Proceedings of the First International Workshop on Designing Meaning Representations*, pages 141–147, Florence, Italy. Association for Computational Linguistics.

Lingfeng Shen, Lemao Liu, Haiyun Jiang, and Shuming Shi. 2022. On the evaluation metrics for paraphrase generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3178–3190.

John F Sowa. 1984. *Conceptual structures: information processing in mind and machine*. Addison-Wesley Longman Publishing Co., Inc.

John F Sowa. 2003. Laws, facts, and contexts: Foundations for multimodal reasoning. *Knowledge Contributors*, pages 145–184.

John F Sowa. 2006. Worlds, models and descriptions. *Studia Logica*, 84(2):323–360.

Kevin Stowe, Jenette Preciado, Kathryn Conger, Susan Windisch Brown, Ghazaleh Kazeminejad, James Gung, and Martha Palmer. 2021. SemLink 2.0: Chasing lexical resources. In *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*, pages 222–227, Groningen, The Netherlands (online). Association for Computational Linguistics.

Ted Supalla. 1986. The classifier system in american sign language. *Noun classes and categorization*, page 181.

Jens E. L. Van Gysel, Meagan Vigus, Pavlina Kalm, Sook-kyung Lee, Michael Regan, and William Croft. 2019. Cross-linguistic semantic annotation: Reconciling the language-specific and the universal. In *Proceedings of the First International Workshop on Designing Meaning Representations*, pages 1–14, Florence, Italy. Association for Computational Linguistics.

Jens EL Van Gysel, Meagan Vigus, Jayeol Chun, Kenneth Lai, Sarah Moeller, Jiarui Yao, Tim O'Gorman, Andrew Cowell, William Croft, Chu-Ren Huang, et al. 2021. Designing a uniform meaning representation for natural language processing. *KI-Künstliche Intelligenz*, 35(3-4):343–360.

Aaron Steven White, Drew Reisinger, Keisuke Sakaguchi, Tim Vieira, Sheng Zhang, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme. 2016. Universal decompositional semantics on universal dependencies. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1713–1723.

Zhaofeng Wu, Hao Peng, and Noah A. Smith. 2021. Infusing finetuning with semantic dependencies. *Transactions of the Association for Computational Linguistics*, 9:226–242.

Zdeněk Žabokrtský, Daniel Zeman, and Magda Ševčíková. 2020. Sentence meaning representations across languages: What can we learn from existing frameworks? *Computational Linguistics*, 46(3):605–665.