Cross-Lingual Semantic Annotation: Reconciling the Language-Specific and the Universal

Jens E. L. Van Gysel, Meagan Vigus, Pavlina Kalm,
Sook-kyung Lee, Michael Regan, William Croft
MSC 03 2130 Linguistics
1 University of New Mexico
Albuquerque NM 87131-0001, USA
{jelvangysel,mvigus,pavlinap,sklee,reganman,wcroft}@unm.edu

Abstract

Developers of cross-lingual semantic annotation schemes face a number of issues not encountered in monolingual annotation. This paper discusses four such issues, related to the establishment of annotation labels, and the treatment of languages with more fine-grained, more coarse-grained, and cross-cutting categories. We propose that a lattice-like architecture of the annotation categories can adequately handle all four issues, and at the same time remain both intuitive for annotators and faithful to typological insights. This position is supported by a brief annotation experiment.

1 Introduction

In recent years, the field of computational linguistics has become increasingly interested in annotation schemes with cross-lingual applicability (Ponti et al., 2018). For syntactic annotation, the Universal Dependencies scheme for grammatical relations between constituents (Nivre et al., 2016) is probably the best-known representative of this new tendency.

On the semantic side, various annotation schemes have been proposed for specific conceptual domains. The Abstract Meaning Representation project (Banarescu et al., 2013) aims to provide a language-neutral representation of argument structure, and was shown by Xue et al. (2014) to have potential in this direction. The Universal Conceptual Cognitive Annotation (Abend and Rappoport, 2013) has the same objective. Annotation schemes designed for cross-lingual application have also been proposed for such semantic domains as the meanings of discourse connectives (Zufferey and Degand, 2017), temporal information (Katz and Arosio, 2001; Pustejovsky et al., 2003), epistemicity (Lavid et al., 2016), modality in general (Nissim et al., 2013), and preposition-like senses (Saint-Dizier, 2006).

However, languages diverge widely in the semantic distinctions they conventionally express, and in the formal means they use to do so (Comrie, 1989; Croft, 2002). Therefore, devising a cross-lingual annotation scheme poses challenges that developers of language-specific schemes need not face. This paper discusses some crucial choices developers of cross-lingual semantic annotation schemes must make with regards to the granularity of linguistic categories. To a large extent, these apply to syntactic annotation as well. In particular, the following four issues need to be accounted for by any annotation scheme with cross-linguistic ambitions:

1. What are the values of the basic labels of the semantic annotation scheme, i.e. which distinctions are annotators expected to make?
2. How are languages with more coarse-grained semantic distinctions accommodated?
3. How are languages with more fine-grained semantic distinctions accommodated?
4. How are languages with distinctions that cross-cut the categories distinguished in the base level annotation scheme treated?

Section 2 of this paper discusses these issues in more detail, exemplifying each of them with data from a range of semantic domains and a range of languages, and section 3 provides a brief overview of how previous cross-lingual annotation schemes have treated them. In section 4, we survey a wider range of possible solutions for these challenges, each with their advantages and drawbacks, and make an argument in favour of establishing a lattice-like structure of hierarchically organized, typologically motivated categories. We also propose a set of guidelines for annotators on which levels of this lattice to use. Section 5 presents an exploratory cross-lingual annotation exercise using such an architecture.

Proceedings of the First International Workshop on Designing Meaning Representations, pages 1–14
Florence, Italy, August 1st, 2019

2 Issues in Cross-Lingual Annotation

When devising an annotation scheme for a semantic domain, one must carve up this region of conceptual space into discrete subregions. For a monolingual scheme, one can straightforwardly base these annotation values on distinctions overtly made in the language. One is likely to run into trouble, however, trying to apply such monolingual categories to a wider sample of languages.

For example, Zufferey and Degand (2017) and Zufferey et al. (2012) have shown that the English-based feature set for the semantics of discourse connectives used by the Penn Discourse Tree Bank (Prasad et al., 2008) needed to be refined when applying it to closely related languages such as French, German, Dutch and Italian. Divergences are expected to be even larger when applying a monolingual scheme to genetically unrelated languages. This section discusses how one can devise a principled cross-linguistic set of labels, and make allowances for languages that do not fit it.

2.1 Establishing the Categories

We propose two heuristics to help one decide on a subdivision of a semantic domain with maximal cross-linguistic applicability. Firstly, choosing semantic categories distinguished by the majority of languages in the world naturally makes the labels of the annotation scheme widely applicable.

For example, Boye (2012) finds that the typologically most common way in which languages subdivide the conceptual domain of epistemic strength, defined as "judgements about the factual status of a proposition" (Palmer, 2001), is a three-way distinction between full support (certainty about the reality status of an event), partial support (less than certain knowledge about the reality status of an event), and neutral support (non-commitment as to the reality status of an event).[1]

Similarly, in the domain of entity quantification, a simple singular vs. non-singular distinction is highly common in the languages of the world (Corbett, 2000). In a cross-lingual annotation scheme for these semantic domains, choosing [FULL, PARTIAL, NEUTRAL] and [SINGULAR, NON-SINGULAR] as basic annotation categories allows most languages to be felicitously analyzed.

A second, practical rather than theoretical, criterion for establishing the main annotation categories is the ease of making the semantic distinctions regardless of the language of annotation. When developers assert that their chosen categories are cross-linguistically applicable, they implicitly argue that they are interpretable even for speakers of languages which do not make them. They also need to provide sufficiently clear guidelines for annotators of many if not all languages to successfully implement them. In the temporal domain, for instance, this would be an argument for an annotation scheme to adopt distinctions between [PAST, PRESENT, FUTURE]. Such categories are both highly salient in our real-world experience, and can be defined in a non-ambiguous way. Therefore, even though some languages (such as Mandarin) lack grammaticalized means to express these categories, one can reasonably assume that annotators will be able to annotate sentences for past, present, or future time reference based on contextual information.

2.2 More Coarse-Grained Distinctions

Not all languages will make the semantic distinctions chosen by the developers as the base values for a conceptual domain. One way in which languages can diverge from them is by lumping together distinctions, i.e. dividing up this region of conceptual space in a more coarse-grained way.

In the domain of modality, for instance, Boye (2012) finds languages that use more coarse-grained distinctions than [FULL, PARTIAL, NEUTRAL]. Southern Nambiquara lumps together partial and neutral support, making a two-way distinction within verbal suffixes (Boye, 2012, p. 99). This two-way distinction corresponds to full ("Declarative") vs. non-full ("Dubitative") epistemic strength. In the temporal domain, Hua shows a Future vs. Non-Future distinction, lumping together past and present (Haiman, 1980), as do many other languages. One may want the annotation scheme to allow for flexibility beyond the use of the base categories to accommodate such languages.

2.3 More Fine-Grained Distinctions

Languages can also subdivide conceptual space in more specific ways than the chosen annotation categories. In the number domain, for instance, more

[1] In keeping with general typological practice, semantic concepts are capitalized in the text when they are language-specific, and are written with a lower-case first letter when they have cross-linguistic reference. Labels for annotation categories are represented in small caps.

fine-grained distinctions within the non-singular region of conceptual space can be made. Languages may distinguish sets of two entities from sets of more than two entities (Dual vs. Plural, Upper Sorbian); sets of two entities, sets of three entities and sets of more than three entities (Dual vs. Trial vs. Plural, Larike); or "small" sets of entities from "large" sets of entities (Paucal vs. Plural, Bayso, Corbett 2000, chapter 2). In the domain of modality, Limbu (Sino-Tibetan) subdivides the Partial category into Weak Partial and Strong Partial support (Boye, 2012).

These cases do not necessarily form problems for an annotation scheme. Since the more fine-grained categories discussed here are all neatly categorized as subdivisions of the chosen basic annotation categories, annotators are expected to be able to identify the correct category label without problems. Nevertheless, in order to preserve as much information as possible, it may be desirable to provide annotators with a way to use more fine-grained categories made in their language instead of (or in addition to) the pre-established category values.

2.4 Cross-Cutting Distinctions

The largest challenge to cross-lingual annotation schemes is posed by languages which divide semantic space in ways that cross-cut, or overlap with, the pre-established categories. This will inevitably be the case in semantic domains that form a continuum which has to be carved up into discrete values for the annotation labels. Examples of such categories can once again be found in the modality and number domains.

Boye (2012), based on data from Craig (1977), shows that Jacaltec distinguishes only Strong Support (*chubil*) and Weak Support (*tato*) in its complementizers. Strong Support corresponds to the cross-linguistic prototype of full support and strong partial support, while Weak Support corresponds to the cross-linguistic prototype of neutral support and weak partial support. In other words, these categories cross-cut the partial support category. For a sentence containing the Weak Support marker, an annotator who wishes to adhere to the proposed category labels must judge whether it falls under the NEUTRAL or PARTIAL category - a judgement they cannot make based on explicit evidence from the language.

Similarly, a small number of languages (e.g. Ainu, Eastern Pomo) make a Few vs. Many distinction in the number domain rather than a Singular vs. Non-Singular one (Veselinova, 2013). They have one category that refers to single referents or small groups (typically up to a maximum of three for Ainu), and a different one to refer to groups greater than this number - dividing up the semantic space in a different, rather than more fine-grained or more coarse-grained, way than the categories found in the majority of languages. In such situations, it is difficult to guide annotators on what to do when they encounter such an overlapping category.

3 Related Work

Previous cross-lingual annotation schemes have not often explicitly addressed the issues laid out in section 2. One scheme accounting for at least two of these issues is Zufferey and Degand's (2017) multilingual adaptation of the PDTB guidelines for discourse connectives. Establishing a hierarchical set of annotation labels based on a small sample of genetically related languages allows them to deal with more fine-grained and more coarse-grained distinctions. Individual annotators are allowed to freely choose values from any level in the hierarchy. When a language divides the semantic domain up in a more fine-grained way, annotators can simply choose values from lower levels of the hierarchy, while for languages with more coarse-grained categories, annotators can choose categories higher up in the structure.

When a given markable is either ambiguous between two pre-established categories, or semantically intermediate between them, they allow annotators to annotate the markable with two tags. Implicitly, this seems meant to solve the problem of cross-cutting categories outlined in 2.4. It does not, however, capture the typological insight that many semantic domains are internally structured and can be captured in semantic maps (Haspelmath, 2003). We know, for example, that in the domain of modality, it should be exceedingly rare if not impossible for a language to show a semantic category subsuming full and neutral support, but not partial support. Therefore, allowing annotators to freely combine annotation labels seems to be too unconstrained of a mechanism to deal with cross-linguistic variation in category boundaries.

Other cross-lingual annotation schemes (e.g. UCCA, Abend and Rappoport 2013; SSA, Grif-

fitt et al. 2018), aim to keep the scheme as intuitive as possible while maintaining cross-linguistic comparability. To this end, UCCA only provides highly schematic annotation categories on the order of [PARTICIPANT, TEMPORAL RELATION, EVENT]. These categories are so general that no language would have more coarse-grained categories. Because of their high level of abstraction, they are also so far apart in conceptual space that languages are unlikely to show overlapping categories. On the other hand, every language will have more fine-grained categories than provided in this scheme. These are not annotated in the base level UCCA, but left to additional annotation layers which researchers can develop for their own purposes.

Lavid et al. (2016) use a similar approach to Zufferey and Degand (2017). They provide a hierarchical structure with three levels of categories for annotating epistemicity, encouraging the use of the lowest levels. When in doubt between the lower-level categories, annotators can choose a higher-level category instead. Nissim et al.'s (2013) cross-lingual scheme for modality also allows annotators to choose coarse-grained categories if they are not confident judging an utterance as an instance of a lower-level category.

While this solution works for languages with coarse-grained categories, strict hierarchical architectures do not allow for easy annotation of overlapping categories. For example, while both these annotation schemes distinguish values for [CERTAINTY, PROBABILITY, POSSIBILITY], the immediately higher-level category is simply one of EPISTEMIC MODAL/FACTUALITY. There is no way to capture categories like those of Jacaltec where some cases of PROBABILITY group with CERTAINTY and others with POSSIBILITY.

4 Potential Solutions

We believe that the most promising architecture for a cross-lingual semantic annotation scheme is to structure the typologically motivated labels as a lattice with different levels, rather than a strict hierarchy. One level contains the categories originally chosen based on the criteria set out in 2.1. This level is designated as the "base level": annotators are encouraged to use categories from this level as the default. The higher and lower levels, respectively, contain equally typologically motivated coarser-grained and finer-grained categories, which can be used when called for by certain applications or certain language-specific categorizations. Such lattices capture the idea that many semantic categories are structured as hierarchical scales, where the middle values can group together with either end, but the extremes of the scale are highly unlikely to be categorized together in any language. Illustrations are provided in figure 1 and figure 2, and in the supplementary materials.

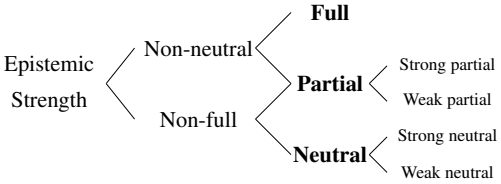


Figure 1: Annotation lattice for epistemic strength

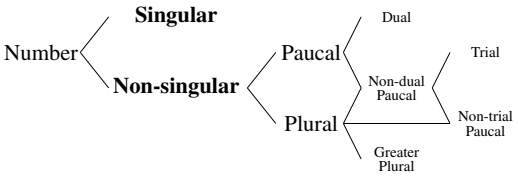


Figure 2: Annotation lattice for number

4.1 More Coarse-Grained Categories

If a language has more coarse-grained semantic categories in a certain domain than those provided in the base level of the lattice (in bold in figures 1-2), it might be difficult for annotators to judge which label to apply to a given use of such a category. For example, for any use of the Nambiquara Dubitative, one would have to judge whether it expresses NEUTRAL or PARTIAL support. This could lead to increased disagreements between annotators. On the one hand, one may still want to require annotators to adopt the base level categories. On the other hand, one might want to ease the annotation process for annotators of languages like Nambiquara.

The lattice architecture allows both goals to be met. As seen in figure 1, [FULL, PARTIAL] strength form an overlapping NON-NEUTRAL category; [PARTIAL, NEUTRAL] strength group together as NON-FULL. Following the aforementioned typological insight, no category groups together [FULL, NEUTRAL] to the exclusion of PARTIAL. Such a lattice avoids the drawback of a strict hierarchy in that it allows for flexibility in

the treatment of the in-between category, which can group with either FULL or NEUTRAL support.

For each use of the Nambiquara Dubitative, then, annotators would be encouraged to judge whether in context it expresses PARTIAL or NEUTRAL support. If such a judgement is too hard to make, annotators may use higher-level values in the lattice, in this case NON-FULL.

4.2 More Fine-Grained Categories

Even though annotators of languages with more fine-grained distinctions than the main level of the lattice should be able to accurately use this level, they may, with an eye on certain downstream applications, want to preserve more specific information encoded in the language. In the Universal Dependencies scheme, annotators are able to add lower-level language-specific categories where needed (e.g. Pyysalo et al. 2015 for Finnish). In order to eliminate the potential proliferation of incommensurable language-specific categories that could result from this, we would encourage annotators to use the base level values as much as possible. In addition, we would provide a set of typologically-based fine-grained categories on a lower level of the lattice. In figure 1, this corresponds to the [STRONG PARTIAL, WEAK PARTIAL, STRONG NEUTRAL, WEAK NEUTRAL] labels, in figure 2 to the [PAUCAL, PLURAL] labels and all labels subsumed underneath them.

In example (1a) from Limbu (van Driem, 1987, p. 244), annotators could follow the distinctions the language makes by labeling the epistemic marker *li·ya* as WEAK PARTIAL. In (1b), they can label *la?ba* as STRONG PARTIAL. Similarly, annotators for a language with fine-grained number categories, such as Yimas, could use the lower-level categories in figure 2. The Yimas Dual, used for reference to exactly two entities, can be marked as DUAL. The Yimas Paucal (typically used for reference to sets containing three to seven entities, Foley, 1991, p. 111) can be marked as NON-DUAL PAUCAL.

(1) a. *ya·?l li·ya.*
groan EPMOD
'He's perhaps groaning.'

b. *ya·?l la?ba.*
groan EPMOD
'He's probably groaning.'

In this way, the specific information expressed in these forms is preserved. At the same time, comparability to other languages is safeguarded: because of the structure of the lattice, lower-level annotations can be traced back, e.g. to the NON-SINGULAR base level category for the DUAL label, and to the PARTIAL category for the STRONG PARTIAL label, and compared to instances of this category in other languages.

Annotators may, in addition, encounter typologically rare fine-grained categories that do not correspond to a pre-specified value in the lattice. They are encouraged in these cases to use base level categories from the lattice. If they feel very strongly that this is not sufficient for their purposes, they will be able to create a language-specific semantic label and specify its position in the lattice.

4.3 Cross-Cutting Categories

Languages with categories that cross-cut the distinctions in the lattice, such as the Jacaltec Strong Support vs. Weak Support system, are the hardest to deal with. The Few vs. Many verbal number system of Ainu, (typically called "Singular" and "Plural", Veselinova 2013), also shows this (2). *Ek* 'come' is used with a set of one to four participants, *arki* 'come' is used with more than four participants (Tamura 1988, p. 40) - cross-cutting the [SINGULAR, NON-SINGULAR] distinction.

(2) a. *tu okkaypo ek.*
two youth come.SG
'Two youths came.'

b. *tupesaniw ka arki ruwe ne.*
eight even come.PL NMLZ COP
'Eight people came.'

We present four options for the annotation of such cross-cutting categories, and argue that the fourth one strikes the best balance between ease of annotation and cross-lingual portability. Firstly, one could allow annotators to completely follow the distinctions their language makes. This would mean that Ainu annotators would establish a FEW category, subsuming the [SINGULAR, DUAL, TRIAL] categories in the lattice, and a MANY category, subsuming [NON-TRIAL PAUCAL, PLURAL]. Alternatively, these categories could be named SINGULAR and PLURAL, since they spread outwards from the cross-linguistic singular and plural prototypes. Along the same lines, Jacaltec annotators would establish a STRONG (or FULL) category for *chubil* and a WEAK (or NEUTRAL) category for *tato*.

This option gives maximal advantage to annotators, who can make use of the exact distinctions expressed in their language. They would not have to distinguish between the different uses of these forms.[2] It comes, however, with a great reduction in cross-linguistic comparability of the resulting annotations. Either the same semantic value will come to be annotated differently in different languages (partial epistemic support would be annotated as PARTIAL in most languages but as either FULL or NEUTRAL in Jacaltec), or the same annotation would mean different things in different languages (SINGULAR would mean "exactly one entity" in Yimas, but "one to three entities" in Ainu).

The second option is a weakened version of the first. Under this approach, the primary annotation of each form is the prototype of this category, but annotators are expected to add the accurate category of the more fine-grained level of the lattice as a secondary annotation.

The Ainu form *ek* would, then, be annotated as SINGULAR:SINGULAR when referring to the coming of one entity, and SINGULAR:NON-SINGULAR when referring to the coming of two or three entities. The first SINGULAR refers to the fact that the cross-linguistic singular category is the prototype of the semantic category expressed by Ainu *ek*. The second annotation expresses the actual semantic value of an utterance on the base level of the annotation lattice. As for modality, Jacaltec annotators would annotate strong partial and full support uses of *chubil* as FULL:STRONG PARTIAL and FULL:FULL respectively.

While this is probably fairly intuitive for annotators, the drawback is that labels such as STRONG PARTIAL no longer exclusively belong to one overarching category. In Jacaltec, it would belong under FULL, while in other languages it would fall under PARTIAL. As a result, annotators for languages with a canonical strong partial vs. weak partial distinction, as proper subcategories of the base level partial support category, would consistently have to employ a secondary annotation as well, specifying the overarching PARTIAL to make the value of this annotation clear. The necessity for two annotation labels to be selected for each form makes this solution fairly cumbersome.

The third option favours cross-linguistic comparison, but is perhaps less intuitive for annotators. It calls for consistent use of the categories specified in the lattice. In such a system, strong partial uses of Jacaltec *chubil* would always be PARTIAL:STRONG PARTIAL. In other words, annotation is done purely on semantic grounds, disregarding language-specific forms. This means that the various uses of the same (polysemous) Jacaltec form will receive different annotations. Even though we believe annotators for all languages should be able to distinguish the base level values of the lattice based on semantic criteria, interpreting such differences which lack overt expression in a language may still be challenging.

Therefore, we believe that our fourth option holds the most promise. This solution allows annotators to use a value in the lattice two levels higher than the markable meaning. For example, for any use of Jacaltec *chubil*, annotators would be allowed to use the label NON-NEUTRAL. This higher-level label allows for the inference that this particular use is either genuinely "in between" the two relevant base level categories (e.g. overlapping the prototypes of partial support and full support), or ambiguous between those two categories. In this way, two levels of the lattice that are problematic from a Jacaltec point of view (FULL vs. PARTIAL on the base level and FULL vs. STRONG PARTIAL at the lower level) are avoided. Of course, as was the case for the treatment of more coarse-grained categories, annotators are still encouraged to specify lower-level values when they can be clearly judged from the context. Thus, strong partial uses of Jacaltec *chubil* could be labeled either NON-NEUTRAL:STRONG PARTIAL, or simply NON-NEUTRAL.

Few cross-lingual annotation schemes have adopted explicit guidelines for languages whose categories cross-cut the pre-established values. Our use of a typologically motivated lattice to organize semantic categories provides various ways to deal with this issue, and at the same time captures insights into regularities in the division of semantic space. We believe that the fourth approach outlined in this section has the best chances of finding wide acceptance. It allows annotators for specific languages to do justice to the semantic structure of the language by recognizing the fine-grained uses of language-specific categories. In

[2] It must be kept in mind, however, that many formal grammatical categories in languages are polysemous. In semantic annotation, annotators must be wary of labeling expressions in a deterministic way based on the most prototypical use of a grammatical marker. Instead, each utterance must be judged based on its meaning in context.

addition, the use of a secondary annotation with a label not one, but two levels higher in the lattice avoids the problem of which superordinate category an in-between usage should be categorized as, and also guarantees cross-lingual portability.

5 Cross-Lingual Annotation Pilot

In order to explore the practicality of a semantic annotation scheme using a lattice structure and the guidelines for label selection outlined above, a small cross-lingual annotation experiment was performed, and is discussed in this section.

5.1 Annotation Procedure and Materials

Thirty-six English sentences expressing spatial figure-ground relations were taken from the STREUSLE corpus (Schneider et al., 2016), and provided thirty-six PPs as annotation targets. These sentences came originally from travel blogs, and were chosen to express spatial scenarios ranging from surface support, to attachment, to containment (figure 3, see also Bowerman and Choi 2001). This continuum was chosen because it is similar to the modality continuum discussed above. While it is exceedingly rare for languages to have one category for only support and containment, the attachment category frequently groups with either containment or support (Bowerman and Choi, 2001). In addition, the existence of spatial situations in between these three base level categories (such as adhesion, for a band-aid on a body part) allows us to confront difficult cross-cutting categories with our lattice architecture.

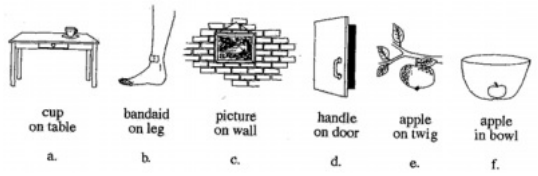


Figure 3: Support-Attachment-Containment continuum (Bowerman and Choi, 2001, p. 485)

Each sentence was translated into Dutch, Czech and Korean by a native speaker of each language (the first, third, and fourth authors of this paper, respectively), and annotated by the same native speaker. The English sentences were annotated by the second author, also a native speaker.

The lattice in figure 4 contains the annotation values, defined based on figure 3. The base level categories are [SUPPORT, ATTACHMENT, CONTAINMENT]. At the higher level, [NON-CONTAINMENT, NON-SUPPORT] group together [SUPPORT, ATTACHMENT] and [ATTACHMENT, CONTAINMENT], respectively. On the lowest level of the lattice, ADHESION cross-cuts the SUPPORT vs. ATTACHMENT distinction, while ATTACHED CONTAINMENT cross-cuts the ATTACHMENT vs. CONTAINMENT distinction.

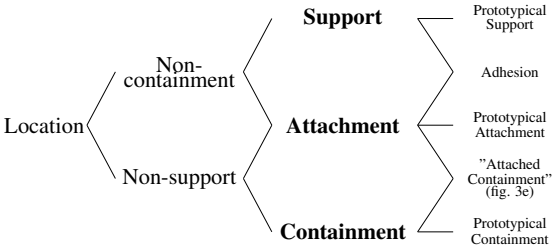


Figure 4: Annotation lattice for spatial relations

Annotators were given the following guidelines:

1. Choose a label from the base level of the lattice based on the meaning of the sentence.
2. If the sentence is ambiguous between two base level values, choose the relevant overarching category.
3. If the sentence expresses a category that is in between two base level values, choose the relevant lower-level category when confident. Otherwise, choose the applicable coarse-grained category above the base level.
4. If the sentence expresses a more fine-grained distinction within one of the base level categories which is not given in the lattice, simply use the applicable base level value.

5.2 Evaluation Procedure

We are aware of few previous experiments annotating multilingual parallel corpora with one set of semantic categories. Closest to our pilot study is probably Zufferey and Degand (2017), who calculate agreement between annotations of a parallel corpus in English, French, German, Dutch, and Italian. Pairwise agreement between English and every other language is reported for each level of the hierarchy in which their categories are structured. The agreement values are given only in raw percentages.

We report pair-wise agreement between all pairs of languages in our pilot. We report both the ex-

act correspondence of annotations between languages, and the compatibility of these annotations. The first set of values is conceptualized as a measure of the discrepancies between the semantic categories of individual languages. For example, an attachment scenario might be annotated as ATTACHMENT in Dutch (which has a preposition *aan* specialized for attachment), but as NON-CONTAINMENT in English, because of its more coarse-grained semantic structure. Under this first measure, these cases are counted as disagreements.

Under the second measure, they are seen as compatible. Since ATTACHMENT is a subcategory of NON-CONTAINMENT, the Dutch annotation can be traced back in the lattice to NON-CONTAINMENT, and the two languages have equivalent annotations on this level. The difference between the exact correspondence score for a language pair and its compatibility score measures the portability of the lattice architecture, and its ability to abstract away from language-specific subdivisions of semantic space.

Both the exact correspondence measure and the compatibility measure are reported as agreement proportions, and as Cohen's Kappa scores (Cohen, 1960). We believe that, even though we are calculating cross-lingual interannotator agreement rather than monolingual agreement between two annotators, the tasks performed by the annotators are still comparable. Since we use a parallel corpus and the same set of annotation values, Cohen's Kappa provides a meaningful measure of how much the proposed annotation system improves labeling over a chance distribution.

5.3 Annotation Results

Table 1 reports cross-lingual interannotator agreement for identity between the chosen labels. The raw proportions of agreement are high, ranging from 82% (Czech-English and Korean-English) to 93% (Czech-Dutch). The Cohen's Kappa scores are also acceptable (between 0.64 and 0.86).

As shown in table 2, pairwise compatibility proportions are on average 7% higher than the corresponding identity scores, and compatibility Kappa scores are on average 0.15 higher than the corresponding identity scores. All language pairs show agreement greater than 90%, and all but one show a Kappa value greater than 0.80.

The organization of annotation categories in a lattice paired with clear guidelines as to which levels of the lattice to use in different situations therefore seems to be a promising way of guaranteeing both ease of annotation and cross-linguistic comparability. It seems fairly successful at abstracting away from language-specific differences in category boundaries, as evidenced by the improvement in the scores for compatibility of annotations as compared to those for exact identity.

	Czech	Dutch	English
Dutch	93% ($\kappa = 0.86$)		
English	82% ($\kappa = 0.64$)	86% ($\kappa = 0.74$)	
Korean	85% ($\kappa = 0.67$)	89% ($\kappa = 0.78$)	85% ($\kappa = 0.66$)

Table 1: Identity between cross-lingual annotations

	Czech	Dutch	English
Dutch	96% ($\kappa = 0.91$)		
English	93% ($\kappa = 0.86$)	94% ($\kappa = 0.90$)	
Korean	90% ($\kappa = 0.79$)	97% ($\kappa = 0.94$)	92% ($\kappa = 0.84$)

Table 2: Compatibility between cross-lingual annotations

A reviewer points out that it is hard to assess the improvement our annotation lattice offers over a flat annotation scheme where annotators are required to choose between [SUPPORT, ATTACHMENT, CONTAINMENT]. We agree that a comparison with such a control condition would be interesting. However, re-annotating this small corpus with such a flat annotation scheme would lead to skewed results, because the present annotators have built up familiarity with the sentences. Since time constraints prevent us from conducting a new annotation experiment in accordance with this suggestion, or from finding new annotators to provide the baseline annotation, we will simply keep it in mind for further work.

5.4 Error Analysis

The differences between the values in table 1 and table 2 stem from annotations which are compatible, but not identical between languages. These annotations reflect both the presence of more coarse-grained categories and cross-cutting categories. As for the former case, examples such as (3a) were annotated as SUPPORT in Czech and Dutch, but as NON-CONTAINMENT in English and (sometimes) Korean. The lattice thus allows anno-

tators in languages with coarse-grained categories to suspend judgement on the base level annotation categories where necessary, while maintaining cross-linguistic comparability.

(3) a. *...right* ***on the back of my car***.
b. *...had nail polish* ***on a couple of toes***.

The same can largely be said for cross-cutting categories. For the single example of surface adhesion in our corpus (3b), the English and Dutch annotators followed guideline 3, choosing the lower-level ADHESION category. The Czech and Korean annotators chose ATTACHMENT and SUPPORT, respectively, both of which are compatible with the Dutch and English choices. This yields compatible annotations in five of the six language pairs, indicating that a category lattice does fairly well in treating cross-cutting categories.

This sentence also illustrates again the problematic character of continuous semantic categories with values in between the base level annotation categories. The ADHESION category cross-cuts the SUPPORT vs. ATTACHMENT distinction, and annotators for different languages (and, conceivably, within one language) will sometimes make different judgements as to which of these two base level categories is appropriate. Choosing a category two levels higher in the lattice instead of just one, as proposed in this paper, would ideally prevent disagreements.

Disagreements also arose with the examples in 4, for which we offer two tentative explanations. Examples (4a-4b), on the one hand, seem likely to give rise to different conceptualizations on the part of annotators. One can interpret *the product* in (4a) to be strictly on top of the hair (leading to the SUPPORT annotations in Dutch and Korean), as clinging to every single hair (resulting in the English ATTACHMENT annotation), or as being contained within the space delimited by the totality of the hair (explaining the Czech CONTAINMENT annotation). Similar conceptualizations can be proposed for *on burger* in (4b): the meat can be seen as contained within the space delimited by the two halves of the bun, or as supported by the bottom half of the bun. Such alternative construals are likely to lead to a certain proportion of disagreements.

(4) a. *...put product* ***on my hair****...*
b. *No meat* ***on burger****...*
c. *...when I am* ***in the chair****...*

The disagreement in (4c) - CONTAINMENT in English vs. SUPPORT in Czech, Dutch, and Korean - is likely to stem from different language-specific conventionalized construals for specific figure-ground configurations. In Dutch, for example, the most natural translation of *in the chair* would be *op de stoel*, using the prototypical support preposition *op*. Using *in*, the containment preposition, is hardly possible. In other words, the relation between a sitter and a chair is always construed as a support relation rather than a containment relation. There does not seem to be a straightforward solution for such cases either. It remains to be seen, however, whether this source of disagreements is recurrent across semantic domains - it might well be more common in the domain of figure-ground relations than in other regions of conceptual space.

6 Conclusions

This paper proposes a lattice-like architecture of cross-lingual semantic annotation systems, with category labels organized in different levels and forming overlapping groupings. This allows us to be faithful to both individual languages and typological generalizations. An approach where cross-cutting categories either receive a low-level, highly specific label (when annotators are confident), or a high-level and uncontroversial label, presents a middle ground between maximizing ease of annotation and maximizing typological rigor. An exploratory cross-lingual annotation task on a small parallel corpus in four languages shows that such an approach has the potential to tackle the issues discussed.

Acknowledgements

This research was supported in part by grant 1764091 by the National Science Foundation to the last author.

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.

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, Sofia, Bulgaria. Association for Computational Linguistics.

Robert Botne. 2012. Remoteness distinctions. In Robert I. Binnick, editor, *The Oxford handbook of tense and aspect*, pages 536–562. Oxford University Press.

Melissa Bowerman and Soonja Choi. 2001. Shaping meanings for language: universal and language-specific in the acquisition of semantic categories. In *Language acquisition and conceptual development*, pages 475–511. Cambridge University Press.

Kasper Boye. 2012. *Epistemic meaning: A crosslinguistic and functional-cognitive study*, volume 43 of *Empirical Approaches to Language Typology*. De Gruyter Mouton, Berlin.

Joan L. Bybee, Revere Dale Perkins, and William Pagliuca. 1994. *The Evolution of Grammar: Tense, Aspect, and Modality in the Languages of the World*. University of Chicago Press Chicago.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Bernard Comrie. 1989. *Language universals and linguistic typology: Syntax and morphology*. University of Chicago press.

Greville G. Corbett. 2000. *Number*. Cambridge University Press.

Colette Grinevald Craig. 1977. *The structure of Jacaltec*. University of Texas Press.

William Croft. 2002. *Typology and universals*. Cambridge University Press.

William Croft. 2012. *Verbs: Aspect and Causal Structure*. Oxford University Press.

William Croft, Pavlína Pešková, and Michael Regan. 2017. Integrating decompositional event structures into storylines. In *Proceedings of the Events and Stories in the News Workshop*, pages 98–109, Vancouver, Canada. Association for Computational Linguistics.

Östen Dahl. 1981. On the definition of the telic-atelic (bounded–nonbounded) distinction in tense and aspect. *Syntax and Semantics*, 14:79–90.

Östen Dahl. 1983. Temporal distance: Remoteness distinctions in tense-aspect systems. *Linguistics*, 21(1):105–122.

Georg van Driem. 1987. *A Grammar of Limbu*. Mouton de Gruyter.

William A Foley. 1991. *The Yimas language of New Guinea*. Stanford University Press.

Kira Griffitt, Jennifer Tracey, Ann Bies, and Stephanie Strassel. 2018. Simple semantic annotation and situation frames: Two approaches to basic text understanding in LORELEI. In *Proceedings of the 11th Language Resources and Evaluation Conference*, pages 1672–1676, Miyazaki, Japan. European Language Resource Association.

John Haiman. 1980. *Hua, a Papuan language of the Eastern highlands of New Guinea*. John Benjamins Publishing.

Martin Haspelmath. 2003. The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison. In *The new psychology of language*, pages 217–248. Psychology Press.

Graham Katz and Fabrizio Arosio. 2001. The annotation of temporal information in natural language sentences. In *Proceedings of the ACL 2001 Workshop on Temporal and Spatial Information Processing*. Association for Computational Linguistics.

Julia Lavid, Marta Carretero, and Juan Rafael Zamorano-Mansilla. 2016. Contrastive annotation of epistemicity in the multinot project: preliminary steps. In *Proceedings of the ISA-12, Twelfth Joint ACL-ISO Workshop on Interoperable Semantic Annotation, held in conjunction with Language Resources and Evaluation Conference*, pages 81–88.

Malvina Nissim, Paola Pietrandrea, Andrea Sanso, and Caterina Mauri. 2013. Cross-linguistic annotation of modality: a data-driven hierarchical model. In *Proceedings of the 9th Joint ISO - ACL SIGSEM Workshop on Interoperable Semantic Annotation*, pages 7–14, Potsdam, Germany. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Frank Robert Palmer. 2001. *Mood and modality*. Cambridge University Press.

Edoardo Maria Ponti, Helen O'Horan, Yevgeni Berzak, Ivan Vulic, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2018. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *CoRR*, abs/1807.00914.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The penn discourse treebank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*,

pages 2961–2968, Marrakech, Morocco. European Language Resources Association (ELRA).

James Pustejovsky, José M Castano, Robert Ingria, Roser Sauri, Robert J Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R Radev. 2003. Timeml: Robust specification of event and temporal expressions in text. *New directions in question answering*, 3:28–34.

Sampo Pyysalo, Jenna Kanerva, Anna Missilä, Veronika Laippala, and Filip Ginter. 2015. Universal dependencies for Finnish. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pages 163–172, Vilnius, Lithuania. Linköping University Electronic Press, Sweden.

Patrick Saint-Dizier. 2006. PrepNet: a multilingual lexical description of prepositions. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, pages 1021–1026, Genoa, Italy. European Language Resources Association (ELRA).

Nathan Schneider, Jena D. Hwang, Vivek Srikumar, Meredith Green, Abhijit Suresh, Kathryn Conger, Tim O'Gorman, and Martha Palmer. 2016. A corpus of preposition supersenses. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 99–109, Berlin, Germany. Association for Computational Linguistics.

Suzuko Tamura. 1988. *The Ainu Language*. Sanseido, Tokyo. Reprinted in 2000, translated from Japanese into English by Sanseido Co. Ltd.

Ljuba N. Veselinova. 2013. Verbal number and suppletion. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

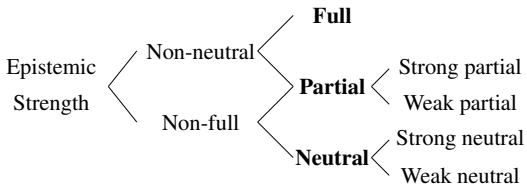
Nianwen Xue, Ondrej Bojar, Jan Hajic, Martha Palmer, Zdenka Uresova, and Xiuhong Zhang. 2014. Not an interlingua, but close: Comparison of English AMRs to Chinese and Czech. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1765–1772, Reykjavik, Iceland. European Language Resources Association (ELRA).

Sandrine Zufferey and Liesbeth Degand. 2017. Annotating the meaning of discourse connectives in multilingual corpora. *Corpus Linguistics and Linguistic Theory*, 13(2):399–422.

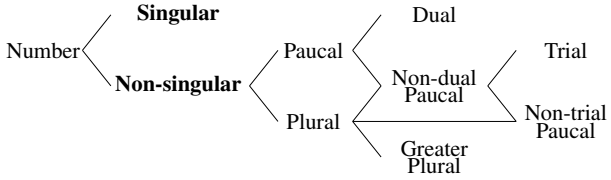
Sandrine Zufferey, Liesbeth Degand, Andrei Popescu-Bellis, and Ted Sanders. 2012. Empirical validations of multilingual annotation schemes for discourse relations. In *Eighth Joint ISO-ACL/SIGSEM Workshop on Interoperable Semantic Annotation*, pages 77–84, Pisa, Italy. Association for Computational Linguistics.

Supplementary Materials

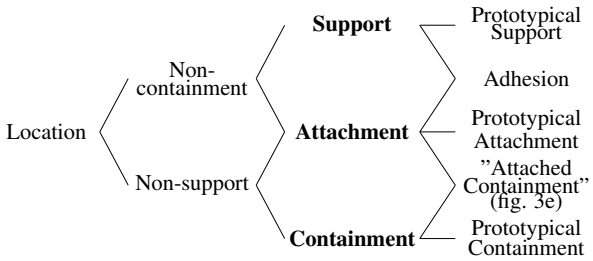
Proposed annotation lattice for epistemic strength



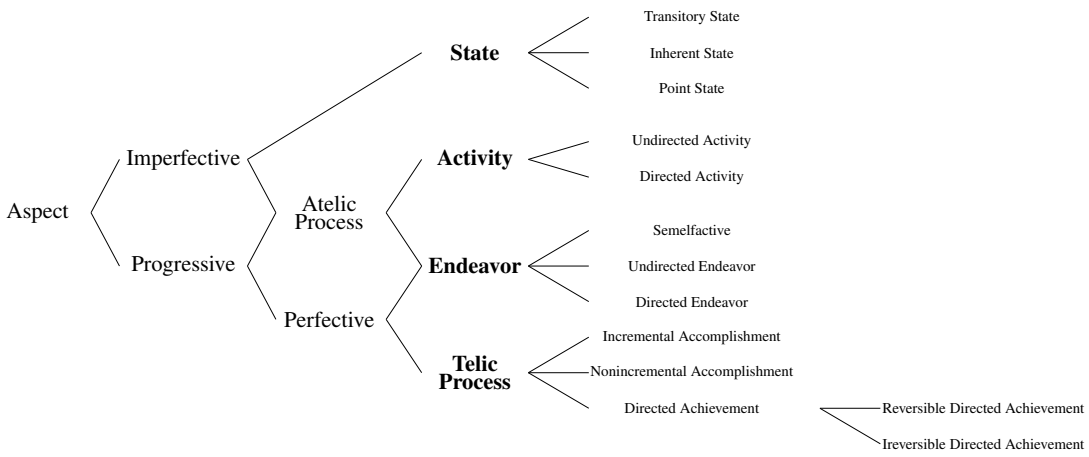
Proposed annotation lattice for number



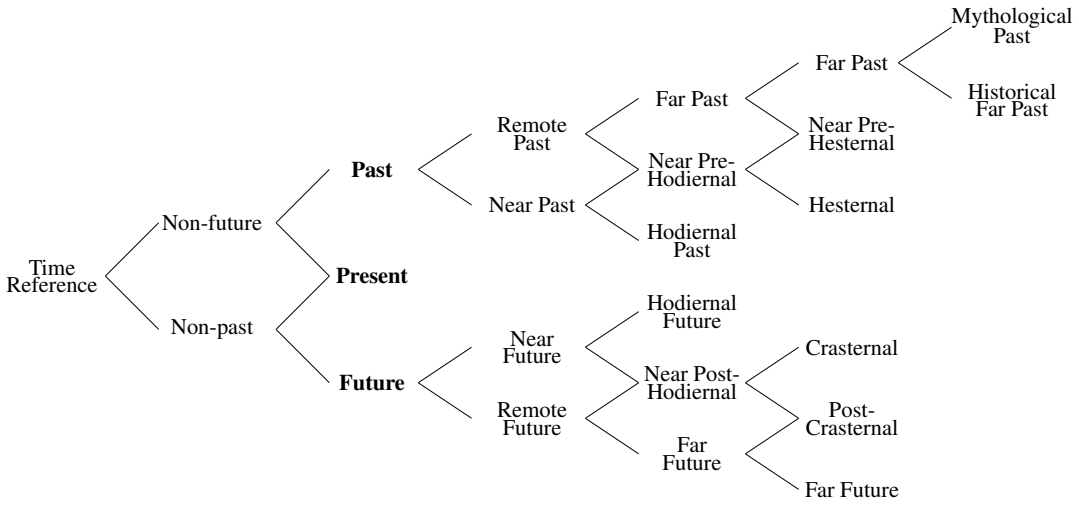
Proposed annotation lattice for spatial relations



Proposed annotation lattice for aspect



Proposed annotation lattice for time reference



These lattices are based on Dahl (1983), Bybee et al. (1994) and Botne (2012) for time reference, Boye (2012) for epistemic strength, Corbett (2000) for number, and Bowerman and Choi (2001) for spatial relations. The aspect lattice is based on the fine-grained aspectual types defined in Croft (2012), with the addition of the category of endeavors (processes that terminate without reaching a natural endpoint or telos), described in Croft et al. (2017). Endeavors are sometimes grouped with telic processes, sometimes not (Dahl, 1981). Imperfectives group together unbounded processes and states, while progressives group together processes, unbounded or bounded (although they describe the state of being in the middle of the process).

References

Robert Botne. 2012. Remoteness distinctions. In Robert I. Binnick, editor, *The Oxford handbook of tense and aspect*, pages 536–562. Oxford University Press.

Melissa Bowerman and Soonja Choi. 2001. Shaping meanings for language: universal and language-specific in the acquisition of semantic categories. In *Language acquisition and conceptual development*, pages 475–511. Cambridge University Press.

Kasper Boye. 2012. *Epistemic meaning: A crosslinguistic and functional-cognitive study*, volume 43 of *Empirical Approaches to Language Typology*. De Gruyter Mouton, Berlin.

Joan L. Bybee, Revere Dale Perkins, and William Pagliuca. 1994. *The Evolution of Grammar: Tense, Aspect, and Modality in the Languages of the World*. University of Chicago Press Chicago.

Greville G. Corbett. 2000. *Number*. Cambridge University Press.

William Croft. 2012. *Verbs: Aspect and Causal Structure*. Oxford University Press.

William Croft, Pavlína Pešková, and Michael Regan. 2017. Integrating decompositional event structures into storylines. In *Proceedings of the Events and Stories in the News Workshop*, pages 98–109, Vancouver, Canada. Association for Computational Linguistics.

Östen Dahl. 1981. On the definition of the telic-atelic (bounded–nonbounded) distinction in tense and aspect. *Syntax and Semantics*, 14:79–90.

Östen Dahl. 1983. Temporal distance: Remoteness distinctions in tense-aspect systems. *Linguistics*, 21(1):105–122.