Not an Interlingua, But Close: Comparison of English AMRs to Chinese and Czech

Nianwen Xue,[*] Ondřej Bojar,[‡] Jan Hajič,[‡] Martha Palmer,[#] Zdeňka Urešová,[‡] Xiuhong Zhang[*]

[*]Computer Science Department, Brandeis University, USA
[‡]Charles University in Prague, Institute of Formal and Applied Linguistics
[#]Dept. of Linguistics, University of Colorado in Boulder, USA
{xuen,xhzhang}@brandeis.edu {bojar,hajic,uresova}@ufal.mff.cuni.cz martha.palmer@colorado.edu

Abstract

Abstract Meaning Representations (AMRs) are rooted, directional and labeled graphs that abstract away from morpho-syntactic idiosyncrasies such as word category (verbs and nouns), word order, and function words (determiners, some prepositions). Because these syntactic idiosyncrasies account for many of the cross-lingual differences, it would be interesting to see if this representation can serve, e.g., as a useful, minimally divergent transfer layer in machine translation. To answer this question, we have translated 100 English sentences that have existing AMRs into Chinese and Czech to create AMRs for them. A cross-linguistic comparison of English to Chinese and Czech AMRs reveals both cases where the AMRs for the language pairs align well structurally and cases of linguistic divergence. We found that the level of compatibility of AMR between English and Chinese is higher than between English and Czech. We believe this kind of comparison is beneficial to further refining the annotation standards for each of the three languages and will lead to more compatible annotation guidelines between the languages.

Keywords: treebank, semantic annotation, multilinguality

1. Introduction

Syntactic treebanks in several languages (Marcus et al., 1993; Hajič et al., 2003; Xue et al., 2005) and related annotated corpora such as Propbank (Palmer et al., 2005), Nombank (Meyers et al., 2004), TimeBank (Pustejovsky et al., 2003), FactBank (Saurí and Pustejovsky, 2009), and the Penn Discourse TreeBank (Prasad et al., 2008), coupled with machine learning techniques, have been used in many NLP tasks. These annotated resources enabled substantial amounts of research in different areas of semantic analysis. There had already been tremendous progress in syntactic parsing (Collins, 1999; Charniak, 2000; Petrov and Klein, 2007) and now in Semantic Role Labeling because of the existence of the PropBank (Gildea and Jurafsky, 2002; Pradhan et al., 2004; Xue and Palmer, 2004; Bohnet et al., 2013) and similar resources in other languages (Hajič et al., 2009), and TimeBank has fueled much research in the area of temporal analysis. There is a concern among NLP researchers, however, that the field of semantic parsing is getting too fragmented. Propbank annotation, for example, focuses on the predicate argument structure of verbs, just as NomBank focuses on the predicate argument structure of nouns. Each verb or noun instance is annotated independently of other predicates in the sentence, and there is not one single representation for the entire sentence. Moreover, there are semantic dependencies that are not covered by either PropBank or Nombank. Only a handful of resources for other languages, such as the the PDT (Hajič et al., 2003) provide full sentence semantic representations. This situation limits the utility of the resulting semantic analyzers. There have been recent on-going efforts to address this concern (Srikumar and Roth, 2013), and one such effort is the development of SemBank using Abstract Meaning Representation (AMR).

2. Abstract Meaning Representation (AMR)

An Abstract Meaning Representation is a rooted, directional and labeled graph that represents the meaning of a sentence and it abstracts away from such syntactic notions as word category (verbs and nouns), word order, morphological variation etc.. Instead, it focuses on semantic *relations* between *concepts* and makes heavy use of predicate-argument structures as defined in PropBank (for English). In addition, many function words (determiners, prepositions) are considered to be syntactic "sugar" and are not explicitly represented in AMR, except for the semantic relations they signal. Readers are referred to (Banarescu et al., 2013) for a complete description of AMR.

In AMR notation, we distinguish two major types of nodes: entity nodes and concept nodes. *Entity nodes* are those labelled with variables, where the variables refer to real-world events and entities. From each entity node, there is usually a link `instance-of` to a relevant concept node. *Concept nodes* represent the "dictionary definition" of the respective entity node(s). For events, the concept nodes are entries in a relevant subcategorization/valency dictionary, e.g. PropBank for English and Chinese or PDT-Vallex for Czech. For named entities, the concepts link to an established ontology of entity types etc. The concept nodes for predicates such as verbs typically have Arg0 and/or Arg1 links to their arguments, as illustrated in the figures below. Entity nodes can also have "mod" links. Usually, concept nodes are leaves in the graph while entity nodes have at least one outgoing edge to a concept node.

Example 1 illustrates an AMR for the sentence "Where is Homer Simpson when you need him?" in English and in its Chinese and Czech translations.

AMR is not intended to be an Interlingua, but by abstracting away from word order, morphology and function words, AMR takes away several sources of cross-lingual differ-

(1) a. Where is Homer Simpson when you need him?

```
(b / be-located-at-91
    :ARG0 (p / person
                :name ( h / name
                            :op1 “Homer”
                            :op2 “Simpson”))
    :ARG1 (a / amr-unknown)
    :time (n / need-01
              :ARG0 (y / you)
              :ARG1 p))
```

b. 当你需要他时, 霍默　辛普森在哪里?

```
(b / be-located-at
    :ARG0 (p / person
                :name (n / name
                           :op1 ” 霍默　辛普森|Homer Simpson”))
    :ARG1 (a / amr-unknown)
    :time (n2 / 需要|need
              :ARG0 (y / 你|you)
              :ARG1 p))
```

c. Kde je Homer Simpson, když ho potřebujete?

```
(b / byt_umisten
    :ARG0 (p / person
                :name ( h / name
                            :op1 “Homer”
                            :op2 “Simpson”))
    :ARG1 (a / amr-unknown)
    :time (p2 / potřebovat-1|need-01
              :ARG0 (v / vy|you)
              :ARG1 p))
```

ences among languages. It would therefore be interesting to see to what extent a sentence and its translation in another language can result in structurally compatible AMRs.[1] The specific concepts or relation types are linked to language-specific dictionaries or ontologies and we don't *a priori* expect these to align, although it would be interesting to see if (and which part of) these dictionaries or ontologies can be unified. For example, it may be possible to use a single ontology for all date and time expressions as well as names for all languages, but the word sense information would have to be different for all languages. To atempt to find answers to these questions, we have created AMRs for 100 Chinese and Czech sentences translated from English to examine how compatible they are to the AMRs of their English originals. In the next two sections, we present a qualitative comparison of the AMRs between Chinese and English and between Czech and English.

[1]By structurally compatible AMRs we mean AMRs with all concepts and relations aligned.

2.1. Language resources behind AMRs

As outlined above, AMRs are an attempt to encompass several independent language resources in a unified framework for formal representation of the meaning of the sentence. Table 1 summarizes the resources that are being used, planned or considered for English, Chinese and Czech variants of AMR.

While some concept types are well covered by language-specific lexicons in several languages, others lack any resource whatsoever. For concepts distant enough from their syntactic realization (e.g. names but not verbs), we may even reuse existing ontologies across languages.

2.2. Related Work

A related formal representation of Functional Generative Description (FGD, (Sgall et al., 1986)) is one of the theories capturing the syntax-semantics interface. FGD has been extensively used as the basis of the Prague dependency treebanks (at the *tectogrammatical* annotation layer), including the parallel Prague Czech-English Dependency Treebank (Hajič et al., 2012). The theory is also well supported with

Concept Nodes for	English	Chinese	Czech
Events, i.e. mainly "verbal" nodes	PropBank	Chinese Propbank	PDT-Vallex
Named entities (NE)	Custom, based on OntoNotes NE guidelines		Czech Wikipedia (under consideration)
Basic concepts (e.g. "mother")	None, we simply use English nouns	None, just use Chinese words	CzechWordNet (under consideration)
Concepts involving numbers	Custom, e.g. special keywords (`distance-quantity`); times and dates similar to Timex from TimeBank		

Table 1: Summary of lexicons and ontologies that AMRs directly refer to.

tools for automatic analysis of sentences and for generation of Czech sentences from their respective deep analyses. Work on English generation using the tectogrammatical representation is underway in the context of another project.[2] The tectogrammatical representation seems to be very closely related to AMRs in that it covers the whole sentence, abstracts away from syntax and morphology, uses a predicate-argument type of lexicon for the annotation of verbs and their senses (the PDT-Vallex, (Hajič et al., 2003; Urešová, 2011)), but it does not abstract as much as AMR does. For example, no attempt is made in the tectogrammatical representation to map event nouns and adjectives to verbal frames, nor to use reification.

Another well-known semantic representation is MRS (Copestake et al., 2005), but we are not aware of any treebank resources using it for such an inter-language comparison.

2.3. AMR resources used

For this study, we have used 100 annotated sentences from a blog on Virginia road construction, taken from the WB part of the Penn Treebank. These sentences have been annotated using AMRs, and also translated to Chinese and Czech and AMR-annotated in these two languages. The English text has 1676 word and punctuation tokens (using the Penn Treebank style tokenization), and its annotated AMR representation contains 1231 nodes.

3. English and Chinese AMRs

An analysis of the annotated English and Chinese AMRs shows that there are three scenarios. In the first scenario, translations of the same sentence are annotated with structurally compatible AMRs. Figure 1 illustrates such an example (please note that the annotation in the subsequent figures is color-keyed: blue is English, and red is Chinese). The AMRs of a Chinese sentence and its English translation show perfect alignment: all concepts and their relations are aligned, except for the tokens in Homer Simpson's name. This is a graphically expressed comparison of the Example 1(a) and (b).

In the second scenario, annotators of the different languages ended up with different AMRs, but the difference is a result of annotation choice and can potentially be reconciled. Such a difference is not different from inter-annotator inconsistency in an annotation task for the same language. An adjudication process or a refinement in annotation standards could potentially resolve this kind of difference. This is illustrated in Figure 2. In the Chinese AMR, the ARG1 of 报道 ("report") is a logical "and" instance connecting two concepts *m* and *h*. In the corresponding English AMR, there isn't such a logical "and" concept. This difference can potentially be reconciled because the difference results from different interpretations at the syntactic level. The Chinese sentence is interpreted as a coordination structure, which naturally maps to a logical "and" in the AMR. The English sentence, however, is interpreted as an unrestrictive relative clause, resulting in an "ARG4-of" relation. At the semantic level, such syntactic differences can potentially be glossed over, but the annotation standards or guidelines need to be refined so that annotators are instructed to abstract away from such differences.

In the third scenario, the differences in AMRs are due to different lexicalizations and such differences cannot be easily resolved without going to an even higher level of abstraction than the AMR representation currently provides.

Both scenarios can be found in Figure 3, which shows aligned AMRs for the Chinese sentence "这 是'一个 大 叫 '噢哦' 的 时刻" and its English translation "This is a major 'd'oh' moment.". In the Chinese AMR, the notion of "be-temporally-located-at" is reified and this results in an extra node that is not matched in the English AMR. However, this is a matter of annotation choice because reification is a legitimate alternative representation to having 这 ("this", node `t`) represented as the `domain` of 时刻 ("moment", node `t2`) instead, which would match the English AMR node. Since such differences in annotation choice can happen for the same language as well, it does not represent a cross-linguistic difference. In contrast, the fact that the English AMR has a node for "major" while the Chinese sentence has a node for 大叫 ("cry") is a difference in lexicalization. Such differences cannot be reconciled by making different annotation choices and can only resolved with a level of abstraction that AMR currently does not provide.

Figure 4 provides another example of difference of lexicalization between Chinese and English. The juxtaposed AMRs are for the Chinese sentence "《 里士满 时报 时事 通讯 》 报道 了 它 对 波瓦坦县 的 胡格诺 山路 的 影响 。" and its English translation "The Richmond Times - Dispatch tells the tale about the impact on the Huguenot Trail in Powhatan County." In one instance,

[2]Project QTLeap funded by the EU, `http://qtleap.eu`.

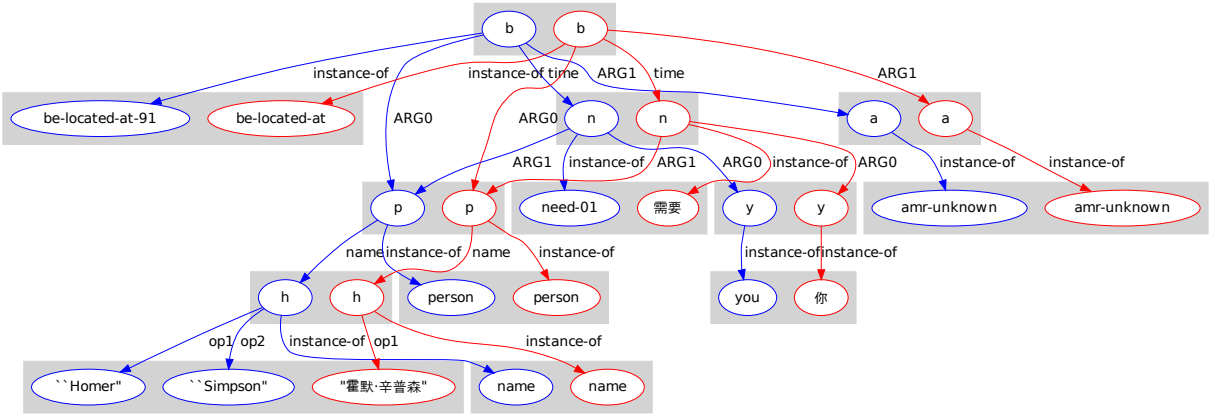

Figure 1: Structurally compatible AMRs (Chinese/English)

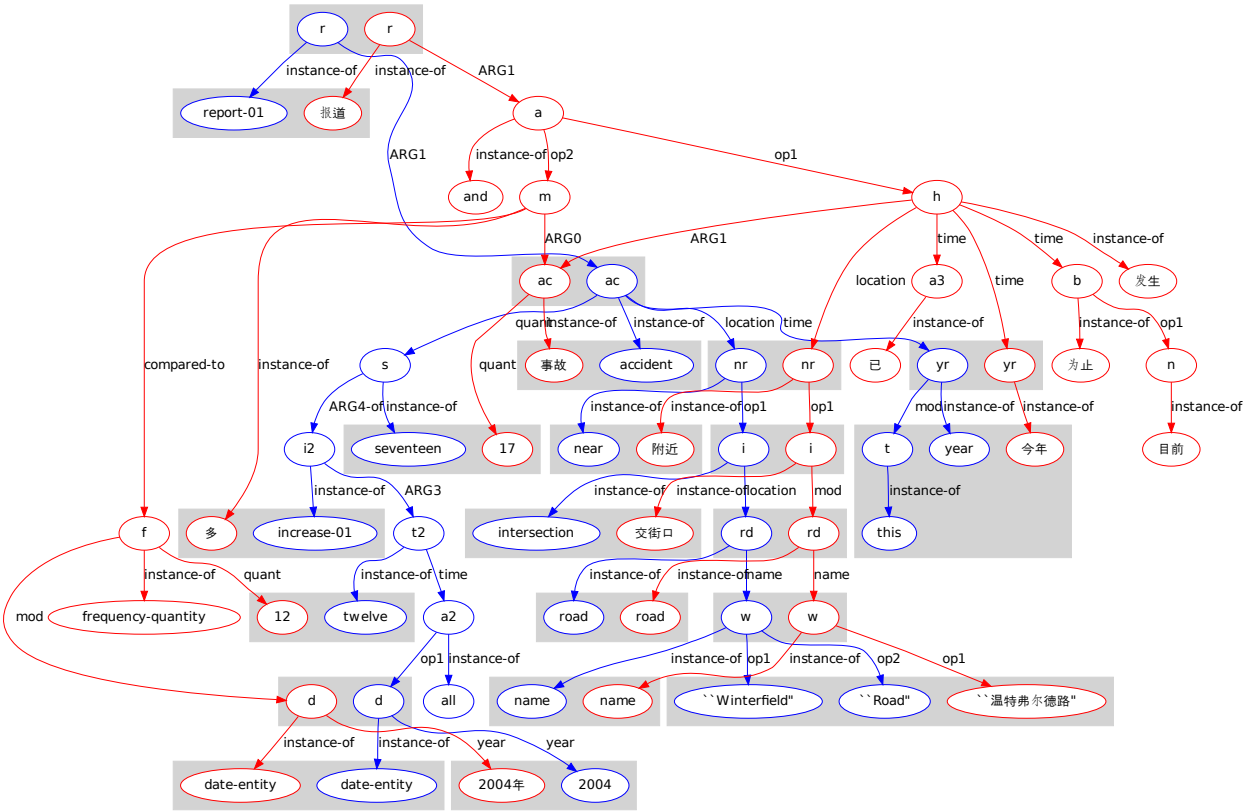

Figure 2: Different structure at the top level (Chinese/English)

the English multiword expression “tells the tale” is translated into a single word (“报道”) in Chinese. In contrast, the English AMR treats “tale” as an argument of “tell-01”, and this structure does not exist in the corresponding Chinese AMR. In another instance, the Chinese nominalized predicate “影响” has two arguments while its corresponding English predicate “impact-01” has only one predicate. This is because the English nominalized predicate “impact-01” has an implicit argument, the factors that bring about the impact that have never been made explicit. Unless one is to make explicit such implicit arguments, at the current level of abstraction, these lexical differences will remain.

4. English and Czech AMRs

A similar picture to the one depicted above for English and Chinese arises when comparing English to Czech: the

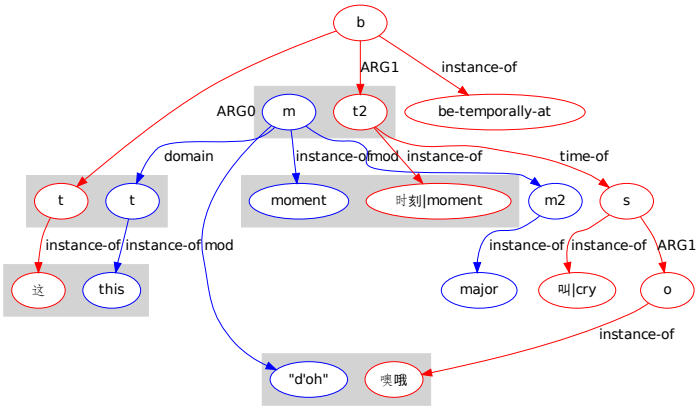

Figure 3: Difference in annotation choice and lexicalization (Chinese/English)

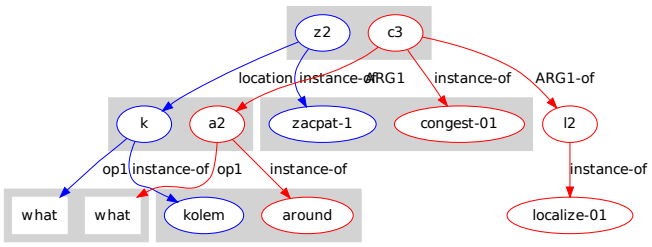

Figure 5: Annotation choice: reification (Czech/English)

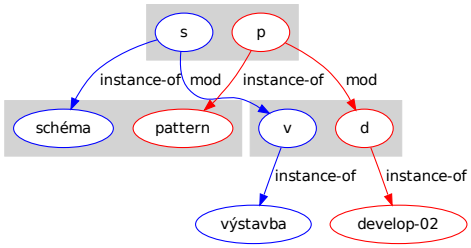

Figure 6: Event-object choice and ontology divergence (Czech/English)

AMRs for parallel sentences are either identical, differ in annotation choice, differ due to different ontologies (or lack of them), or are incompatible due to a substantially different sentence structure in the Czech translation. All of our 100 sentences (cf. Sect. 2.3.) have been doubly translated from English to Czech and reviewed. One set has then been revised, corrected and manually annotated with AMRs.

4.1. Annotation choice

In Fig. 5,[3] the reification of the notion "being located somewhere" (congestion taking place around something) in the English AMR leads to an extra structure (`l2 / localize-01`). In Czech, this information was captured by the arc label `location`. Since this is described as an alternative annotation by the AMR guidelines, such a case should not in fact be considered a "true" difference in the AMR structure. Fig. 6 can also be regarded as a difference in annotation choice only. The example shows that the phrase "*development patterns*", translated as "*schémata výstavby*", got represented using an event node `develop-02` in English and a object node `výstavba` in Czech. This is related to the underlying ontologies, in which some events might not be as strongly represented, leading to an "object" type of annotation instead.

[3]Please note that Czech is color-coded blue and English red in the comparison graphics in this Section.

4.2. Ontology difference

In Fig. 7, the difference lies in an annotation of a MWE in English as two nodes, whereas since the Czech translation is a single-word expression, the AMR for that naturally contains only one node ("*tell tale*" vs. "*popsat*").

Fig. 8 also shows an inverse case, where in English there is a single word while in Czech a three-word phrase ("*speeding*" vs. "*překračovat povolenou rychlost*" ("*to-surpass the-permitted speed*").

4.3. Incompatible structure

Fig. 7 shows also a different structural annotation caused by the insertion of the word "*vydání*" ("*issue*") which follows the practice of using this descriptor when an extra adjective is used ("*dnešní*"; "*today*"), especially with non-declining foreign names. Unless a specific rule is created that states that a newspaper entity name actually means an instance ("*issue*"), this difference is unresolvable the way the Czech sentence is formulated. Similar situations arise when the translator adds (in Czech) explicitly an event which is only implicit in the source (English), such as in "*resident engineer*" → "*odborník zaměstnaný v ... (specialist employed in/by ...)*". This case naturally results in an extra event in the Czech AMR, namely "*pracovat-01 (to-be-employed)*".

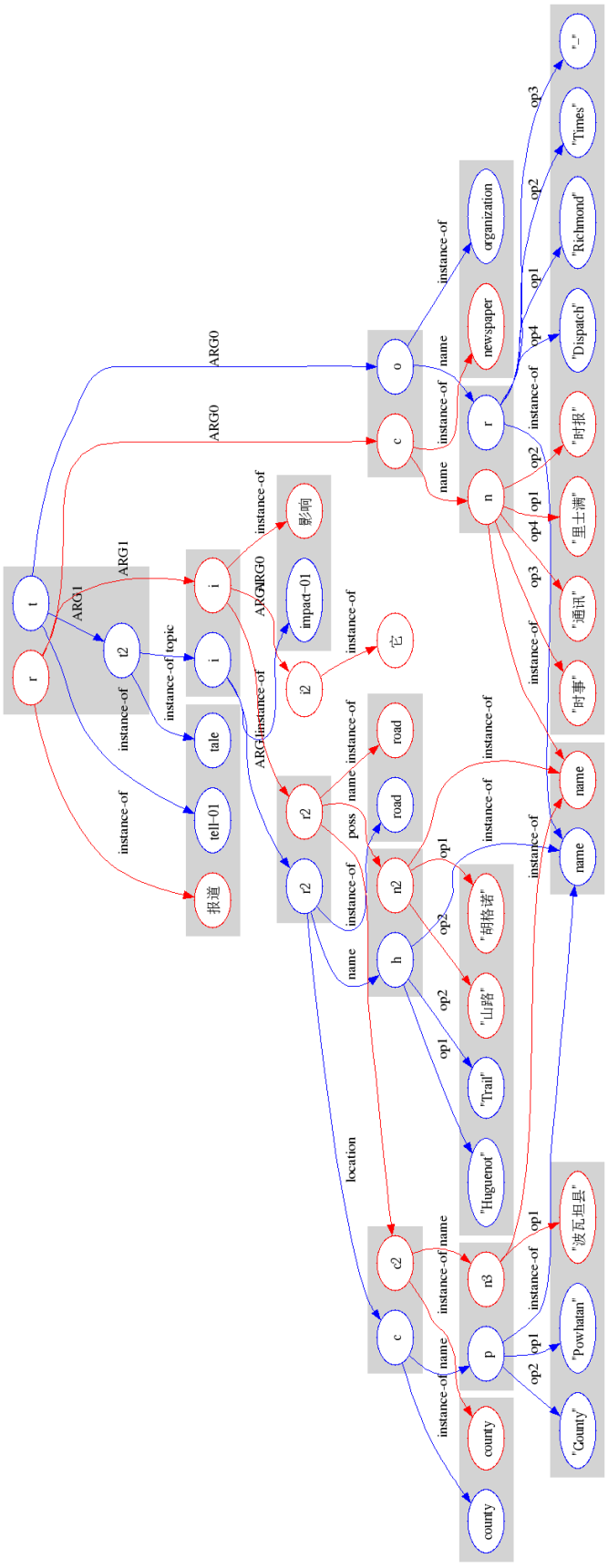

Figure 4: Difference in annotation choice and lexicalization (Chinese/English)

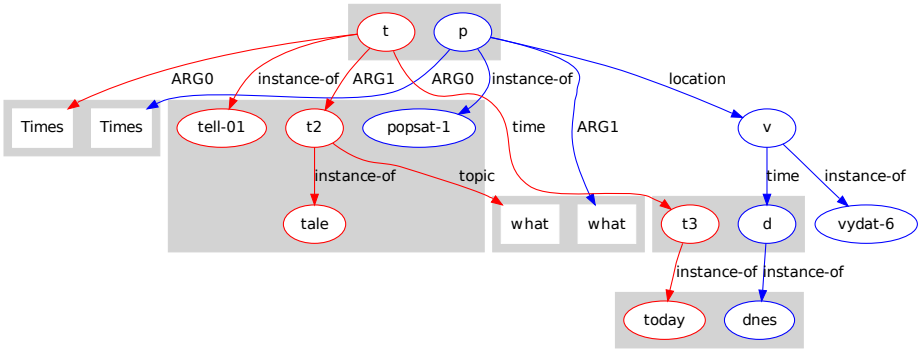

Figure 7: MWE and a structural divergence (Czech/English)

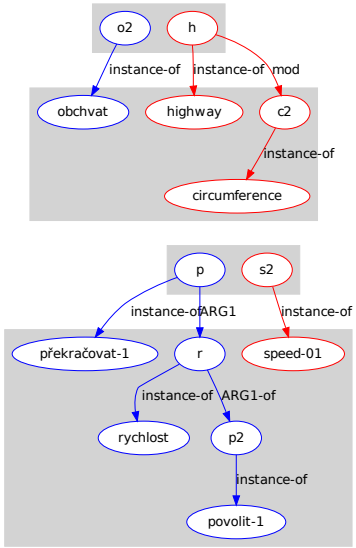

Figure 8: Multiword expression vs. single word (Czech/English)

4.4. Overall observations for Czech vs. English

We have found by manual inspection that only 29 Czech sentences have structurally identical annotation (and possibly differ only in argument labeling) and 18 additional sentences contained differences which can be considered "local". This suggests that over a half of the annotated sentences differ structurally in some more profound way.

5. Conclusions and Future Work

We have described an ongoing effort to analyze and refine the Abstract Meaning Representation (AMR) annotation framework aimed at semantic representations of sentences. By comparing English to Chinese and Czech annotations on parallel texts, we reveal some natural divergences between the language pairs but also some points of AMR that could still use refinement. We have found that there is a relatively large number of Czech sentences that differ structurally from the annotated English counterparts.

The illustrated divergences indicate that using AMRs, e.g., as a sort of a transfer layer in machine translation may require quite large and complex entries (elementary (sub)graphs) in the "translation dictionary". Using AMRs for just the source or just the target, as proposed by Kevin Knight (Jones et al., 2012) can be thus more appropriate. It has yet to be seen how exactly AMRs can be deployed in information extraction, entailment and other semantic tasks. In any case, contrastively comparing AMRs across languages is definitely beneficial for further refinement of the annotation specification and guidelines and the associated annotation practice.

In any case, we believe that a further study of the differences is necessary, to analyze not only the annotation differences in detail and relate them to the AMR giudelines, but also to investigate the influence of translation choice and creativity, which in fact might account for a non-trivial number of differences, too.

6. Acknowledgements

We gratefully acknowledge the support of the National Science Foundation Grant NSF: 0910992 IIS:RI: Large: Collaborative Research: Richer Representations for Machine Translation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. The project has been partially supported by the grant No. GPP406/13/03351P of the Grant Agency of the Czech Republic and by the grant LH12093 of the Ministry of Education, Youth and Sports of the Czech Republic. This work has been using language resources and tools developed and stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2010013).

7. References

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., and Schneider, N. (2013). Abstract Meaning Representation for Sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop*, Sophia, Bulgaria.

Bohnet, Bernd, Nivre, Joakim, Boguslavsky, Igor, Farkas, Richard, Ginter, Filip, and Hajič, Jan. (2013). Joint morphological and syntactic analysis for richly inflected lan-

guages. *Transactions of the Association for Computational Linguistics*, 1:415–428.

Charniak, Eugene. (2000). A maximum-entropy-inspired parser. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pages 132–139. Association for Computational Linguistics.

Collins, Michael. (1999). *Head-driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania.

Copestake, Ann, Flickinger, Dan, Pollard, Carl, and Sag, IvanA. (2005). Minimal recursion semantics: An introduction. *Research on Language and Computation*, 3(2-3):281–332.

Gildea, Daniel and Jurafsky, Daniel. (2002). Automatic labeling of semantic roles. *Computational linguistics*, 28(3):245–288.

Hajič, Jan, Panevová, Jarmila, Urešová, Zdeňka, Bémová, Alevtina, Kolářová, Veronika, and Pajas, Petr. (2003). PDT-VALLEX: Creating a large-coverage valency lexicon for treebank annotation. In Nivre, Joakim and Hinrichs, Erhard, editors, *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*, volume 9 of *Mathematical Modeling in Physics, Engineering and Cognitive Sciences*, pages 57–68, Vaxjo, Sweden. Vaxjo University Press.

Hajič, Jan, Ciaramita, Massimiliano, Johansson, Richard, Kawahara, Daisuke, Martí, Maria, Màrquez, Lluís, Meyers, Adam, Nivre, Joakim, Padó, Sebastian, Štěpánek, Jan, Straňák, Pavel, Surdeanu, Mihai, Xue, Nianwen, and Zhang, Yi. (2009). The CoNLL-2009 Shared Task: Syntactic and Semantic Dependencies in Multiple Languages. In Hajič, Jan, editor, *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL): Shared Task*, pages 1–18, Boulder, CO, USA. Association for Computational Linguistics.

Hajič, Jan, Hajičová, Eva, Panevová, Jarmila, Sgall, Petr, Bojar, Ondřej, Cinková, Silvie, Fučíková, Eva, Mikulová, Marie, Pajas, Petr, Popelka, Jan, Semecký, Jiří, Šindlerová, Jana, Štěpánek, Jan, Toman, Josef, Urešová, Zdeňka, and Žabokrtský, Zdeněk. (2012). Announcing Prague Czech-English Dependency Treebank 2.0. In *Proceedings of the Eighth International Language Resources and Evaluation Conference (LREC'12)*, pages 3153–3160, Istanbul, Turkey, May. ELRA, European Language Resources Association.

Hajič, Jan, Böhmová, Alena, Hajicová, Eva, and Hladká, Barbora. (2003). The Prague Dependency Treebank: A Three Level Annotation Scenario. In Abeillé, Anne, editor, *Treebanks: Building and Using Annotated Corpora*. Kluwer Academic Publishers.

Jones, Bevan, Andreas, Jacob, Bauer, Daniel, Hermann, Karl Moritz, and Knight, Kevin. (2012). Semantics-based machine translation with hyperedge replacement grammars. In *COLING*, pages 1359–1376.

Marcus, Mitchell P., Santorini, Beatrice, and Marcinkiewicz, Mary Ann. (1993). Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330.

Meyers, A., Reeves, R., Macleod, C., Szekely, R., Zielinska, V., Young, B., and Grishman, R. (2004). The NomBank Project: An Interim Report. In *Proceedings of the NAACL/HLT Workshop on Frontiers in Corpus Annotation*, pages 24–31, Boston, Massachusetts.

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

Petrov, Slav and Klein, Dan. (2007). Improved inference for unlexicalized parsing. In *HLT-NAACL*, pages 404–411.

Pradhan, Sameer S, Ward, Wayne, Hacioglu, Kadri, Martin, James H, and Jurafsky, Daniel. (2004). Shallow semantic parsing using support vector machines. In *HLT-NAACL*, pages 233–240.

Prasad, Rashmi, Dinesh, Nikhil, Lee, Alan, Miltsakaki, Eleni, Robaldo, Livio, Joshi, Aravind, and Webber, Bonnie. (2008). The Penn Discourse Treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco.

Pustejovsky, James, Hanks, Patrick, Saurí, Roser, See, Andrew, Day, David, Ferro, Lisa, Gaizauskas, Robert, Lazo, Marcia, Setzer, Andrea, and Sundheim, Beth. (2003). The TimeBank Corpus. *Corpus Linguistics*, pages 647–656.

Saurí, Roser and Pustejovsky, James. (2009). Factbank: a corpus annotated with event factuality. *Language resources and evaluation*, 43(3):227–268.

Sgall, Petr, Hajičová, Eva, and Panevová, Jarmila. (1986). *The Meaning of the Sentence and Its Semantic and Pragmatic Aspects*. Academia/Reidel Publishing Company, Prague, Czech Republic/Dordrecht, Netherlands.

Srikumar, Vivek and Roth, Dan. (2013). Modeling semantic relations expressed by prepositions. *Transactions of the Association for Computational Linguistics*, 1:231–242.

Urešová, Zdeňka. (2011). *Valenční slovník Pražského závislostního korpusu (PDT-Vallex)*. Studies in Computational and Theoretical Linguistics. Ústav formální a aplikované lingvistiky, Praha, Czechia.

Xue, Nianwen and Palmer, Martha. (2004). Calibrating features for semantic role labeling. In *EMNLP*, pages 88–94.

Xue, Nianwen, Xia, Fei, dong Chiou, Fu, and Palmer, Martha. (2005). The Penn Chinese TreeBank: Phrase Structure Annotation of a Large Corpus. *Natural Language Engineering*, 11(2):207–238.