The Hebrew FrameNet Project

Avi Hayoun, Michael Elhadad

Dept. of Computer Science
Ben-Gurion University
Beer Sheva, Israel
`{hayounav,elhadad}@cs.bgu.ac.il`

Abstract

We present the Hebrew FrameNet project, describe the development and annotation processes and enumerate the challenges we faced along the way. We have developed semi-automatic tools to help speed the annotation and data collection process. The resource currently covers 167 frames, 3,000 lexical units and about 500 fully annotated sentences. We have started training and testing automatic SRL tools on the seed data.

Keywords: FrameNet, Hebrew, frame semantics, semantic resources

1. Introduction

Recent years have seen growing interest in the task of **S**emantic **R**ole **L**abeling (SRL) of natural language text (sometimes called "shallow semantic parsing"). The task is usually described as the act of identifying the **semantic roles**, which are the set of semantic properties and relationships defined over constituents of a sentence, given a semantic context.

The creation of resources that document the realization of semantic roles in natural language texts, such as FrameNet (Fillmore and Baker, 2010; Ruppenhofer et al., 2010) and PropBank (Kingsbury and Palmer, 2002) have advanced the field of semantic analysis no end and have allowed the development of learning algorithms for automatically analyzing the semantic structure of text.

Shallow semantic analysis has been shown to contribute to the advancement of a wide spectrum of natural language processing tasks, ranging from information extraction (Surdeanu et al., 2003) and question answering (Shen and Lapata, 2007), to machine translation (Wu and Fung, 2009) and abstractive summarization (Melli et al., 2005).

1.1. FrameNet

FrameNet (Fillmore and Baker, 2010; Ruppenhofer et al., 2010) is a human-annotated linguistic resource with rich semantic content based on the linguistic theory of *Frame Semantics* proposed by Fillmore (1982). FrameNet defines a formal structure for semantic frames, and various relationships between and within them.

Each frame contains a list of frame-evoking words which also serve as the predicates of events described by the frames. These words are called **F**rame **E**voking **E**lements (**FEE**s) or **L**exical **U**nits (**LU**s). Additionally, each frame defines a list of event-participants and a list of constraints on and relationships between these participants. The participants are called **F**rame **E**lements (or **FE**s).

Finally[1], and perhaps most importantly, FrameNet contains human-annotated examples of realizations of frames and their structures in natural language.

1.2. FrameNet in Other Languages

The original FrameNet project has been adapted and ported to multiple languages. The most active international FrameNet teams include the Swedish FrameNet (SweFN) covering close to 1,200 frames with 34K LUs (Ahlberg et al., 2014); the Japanese FrameNet (JFN) with 565 frames, 8,500 LUs and 60K annotated example sentences (Ohara, 2013); and FrameNet Brazil (FN-Br) covering 179 frames, 196 LUs and 12K annotated sentences (Torrent and Ellsworth, 2013).

Inspired by the ideas developed by the Swedish FrameNet++ project (Friberg Heppin and Voionmaa, 2012) and by Petruck (Petruck, 2005; Petruck, 2009) and Boas (Boas, 2011), we have started the development of a Hebrew FrameNet, a semi-automatic translation of the English FrameNet.

In this paper, we present this new resource, the methods we used to develop it and the specific linguistic issues we faced while addressing frame annotations in a morphologically rich language like Hebrew.

In the rest of this paper, we first present the linguistic resources and supporting infrastructure used as a starting point for the project. We then discuss the process adopted to develop the Hebrew FrameNet resource, the tools developed, and how we addressed the linguistic issues we faced. Finally, we present the current state of the project and discuss future work.

The project includes a collaborative web-based annotation tool[2] which supports browsing, annotating and searching the Hebrew FrameNet. We are starting to train and test automatic Hebrew SRL systems on the annotated data that we are collecting.

2. Starting Points

The Hebrew FrameNet project is built on wide-coverage Hebrew linguistic resources that have been

---

[1]The FrameNet resource encodes additional information, such as inter-frame relationships and standard lexical patterns of the realization of frames in natural language, but these details are beyond the scope on this paper.

[2]`http://www.cs.bgu.ac.il/~nlpproj/newhebfn/`

developed in the past 15 years. We rely on the Hebrew lexicon described by Itai and Wintner (2008) and a Hebrew corpus, which contains about 1.75M sentences (an expansion of the Corpus of Contemporary Hebrew (Itai and Wintner, 2008)). We annotated all sentences using current state of the art automatic Hebrew annotators: a morphological analyzer and disambiguator (Adler and Elhadad, 2006; Tsarfaty and Goldberg, 2008), a POS tagger (Goldberg et al., 2008) and a syntactic parser (Goldberg and Elhadad, 2010). As part of this annotation effort, we used the Hebrew POS tagset (Netzer et al., 2007; Adler et al., 2008) together with this toolset (see Section 4.3.).

3. Development and Annotation Processes

Following the ideas of and initial analysis described in (Boas, 2011; Petruck, 2005; Petruck, 2009), we used the English FrameNet (version 1.5.1) as a basis to create the Hebrew project. We use the semantic frames as found in the original FrameNet, including definitions, participant roles, inter-role and inter-frame relationships as a language neutral pivot representation.

We describe the key design decisions we have adopted:

1. which frames to annotate;

2. how to select lexical units;

3. how to select exemplar sentences for each frame;

4. how to adapt the annotation to the specific Hebrew rich morphology configurations.

3.1. Initial Frame Selection

Two options have been considered in the past for selecting an initial subset of candidate frames for annotation: domain-based selection (Candito et al., 2014) and frequency-based selection. We opted for the frequency-based approach - aiming for large coverage.

We selected a subset of 200 of the most frequently used frames as candidates for annotation. We estimated frame frequency by inspecting the top 5K most commonly used words in contemporary American English[3]. We then searched FrameNet for frames in which the most popular words participate as lexical units (LUs). This is obviously a noisy process but provides a basis for ranking.

The decision to use English word-frequency, as opposed to Hebrew word-frequency, was a practical one, driven by three main factors:

1. The English data is readily available in the form of the FrameNet database itself.

2. Selecting the most frequent Hebrew words and mapping them to English LUs and frames is a much slower manual process; it is sometimes very difficult to disambiguate similar frames. For example, should **say.v** be an LU of the Frame

SPELLING_AND_PRONOUNCING, the frame COMMUNICATION or both?

3. Mapping words directly to Hebrew frames may have resulted in a sparse dataset, meaning we would have many frames with few LUs, as opposed to a smaller set of complete frames.

To verify that the frame frequency we estimated in the English FrameNet corresponds to similar coverage in contemporary Hebrew, we estimated the coverage of our corpus by the LUs of the selected English frames. We found a coverage of roughly 66%, sufficient to validate this approach.

3.2. Lexical Unit Selection

As neither the English FrameNet project nor the Hebrew lexicon we use contain word sense information[4], there was no direct way of automatically mapping English LUs to corresponding Hebrew terms.

To accelerate the manual process of adding Hebrew LUs to frames, we collected nearly 49K translation pairs for most lexical units occurring in the English FrameNet from online lexicographic resources. We then introduced lookup procedures as part of the online annotation tool we developed for the project (see Figure 1).

In order to populate a given frame with Hebrew LUs, annotators may select an English LU, for which our system suggests the available translations, and select the most appropriate translation. This translation pair selection is stored as a form of word sense disambiguation in the system. Additionally, annotators can add as LUs words which were not found in the translation suggestions as they see fit.

3.3. Exemplar Sentence Selection

The next step in the annotation process is the selection of exemplar sentences per LU for each frame.

To assist in this step, we prepared a corpus of 1.75M sentences collected from a variety of Modern Hebrew sources (newspapers, blogs, Wikipedia, medical resources) and pre-processed these sentences with automatic full morphological analysis and automatic syntactic parsing, as described in Section 2..

Annotators can access this annotated corpus from the Hebrew FrameNet annotation tool through a full text search interface, where annotators can search for lexical items irrespective of morphological inflection and refine the query by specifying part of speech, morphological features (number, gender, person etc.), and syntactic context (*e.g.*, word appearing as the subject of a specific verb).

We apply the syntactic diversification algorithm of Borin et al. (2012) to the search result set, so that the top N sentences presented to each annotator exhibit a wide range of syntactic constructs and lexical items, allowing annotators to quickly create a range of syntactic examples for a single semantic concept.

---

[3] As reported in the Word Frequency Data corpus http://www.wordfrequency.info/free.asp (Davies, 2010)

[4] FrameNet LUs do have definitions, but they are not structured in a manner that enables automatic analysis.
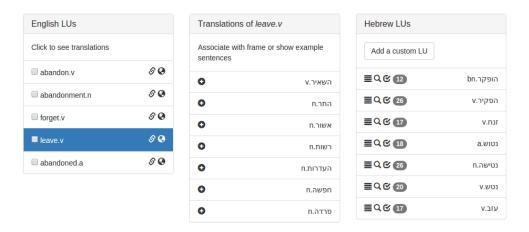
Figure 1: A list of automatic translations of the LU *leave.v* suggested as LUs for the Hebrew implementation of the **Abandonment** frame.

3.4. Exemplar Sentence Annotation

The final stage in the annotation process of a single Frame consists of annotating occurrences of frame elements with their roles within the selected exemplar sentences (see Figure 2).

While the exemplar sentences are annotated with syntactic information, we do not rely on this structure at the semantic annotation stage, as this information is added automatically, and is not reliable enough. As a result, our annotators must select spans of text to fill the various roles encoded in the semantic frame definition, as opposed to annotating syntactic structures, a method employed by some other projects (Erk et al., 2003; Candito et al., 2014). We are planning to compare the annotation decisions made during manual annotation with the spans predicted by the automatic parsing tools and present this information as a reviewing / quality control tool in the future.

4. Hebrew-Specific Issues

We identified the following issues specific to Hebrew in our initial annotation effort.

4.1. Multi-word Lexical Units

The English FrameNet project contains several multi-word LUs (MWLUs), such as *give up.v* and *turn in.v*, which are annotated as contiguous units. In Hebrew, we found many complex morphological and syntactic

variants of MWLUs. For example, while | מזל רע / *mazal ra* / bad luck |

is a contiguous unit, the LU | יצר קשר / *yatzar kesher* / contacted | in both

of the following sentences illustrates the wide range of discontinuous constructions we observed:

1. | דודו / *dudu* / David | יצר קשר / *yatzar kesher* / contacted | עם חייזרים / *im chayzarim* / aliens | .

2. | הוא / *hu* / He | יצר איתם קשר / *yatzar itam kesher* / contacted them | בשנות ה-90 / *be'shnot ha-90* / in the '90s | .

We decided to enable annotation of discontinuous units - but distinguish between LUs that must appear as contiguous (marked by a binary flag in the multi-word LU) and those that can be broken apart.

4.2. Role-Bearing Phrases Embedded in Morphology

Due to the rich morphology of Hebrew, a role-bearing phrase can be embedded as a morpheme in another word in the text. For example, consider this sentence from the **Ingestion** frame:

| אכלתי / *akhalti* / I ate | תפוח / *tapuach* / an apple | .

If the first-person singular pronoun suffix תי in the word אכלתי were a separate token, the correct annotation would be:

| אכל / *akhal* / Ate | [תי / *ti* / I INGESTOR] | [תפוח / *tapuach* / an apple IN- | GESTIBLES].

The AGENT is instead embedded in another token, since the subject is actually realized as a morpheme of the verb. This issue is somewhat similar to that of compound words in Swedish (Heppin and Gronostaj, 2012). In Hebrew, however, the compounding of role-bearing elements is limited to pronouns, and more closely resembles the same issue in Spanish, Italian and some other Latin-derived languages.

As a result, instead of adding an extra annotation layer[5], like in the SweFN project, we decided to followed Petruck's recommendation; borrowing from the Spanish FrameNet project, we annotate such roles as "externally constructionally null instantiated" (ECNI) (Subirats, 2009).

---

[5]Since the pronoun compounding in Hebrew is not lim-

Figure 2: An exemplar sentence for the LU אכל.v / *achal* / ate.v in the frame **Ingestion**.

4.3. What Units can Evoke Frames: POS Tagset

Traditionally, verbs, nouns, adjectives and prepositions are considered as candidate LUs (Fillmore et al., 2002). However, as discussed by Adler et al. (2008), part-of-speech tagsets must sometimes be modified for different languages. In the case of Hebrew, we use the *Beinoni* tag, which occupies a middle place between noun and verb and is most closely related to participial forms. From a semantic point of view, according to traditional descriptions, Hebrew *Beinoni* forms do not denote a fixed state, but activities, in contrast to nouns and adjectives. As an example, consider the difference between the semantic information carried by the word מטפס in the following sentences:

1. שלמה | הוא | מטפס הרים .
   Shlomo | *hu* | *metappes harim*
   Solomon | is | (a) mountain climber

2. הילד | מטפס | במעלה
   ha'yeled | *metappes* | *be'maaleh*
   The child | (is) climbing | up
   סולם החבלים .
   sulam ha'chavalim
   the rope ladder

We distinguish between these two usages of the word, and annotate the predicate-denoting form as *Beinoni* - and thus as a frame-evoking lexical unit.

4.4. Word Segmentation

Due to word formation rules in Hebrew, word segmentation is a difficult task (Goldberg and Elhadad, 2013). For example, the word בצל is ambiguous and can be read either as a single word (onion) or two (ב-צל in the shade). As part of our pre-processing pipeline, we use Goldberg and Elhadad's text segmentation method, which achieves about 94% accuracy on the data we evaluated. To address the potential errors in annotation, we enable manual annotation of text spans that fall within a single sequence of non-space letters.

5. Semi-Supervised Dataset Expansion

Fürstenau and Lapata (2012) presented a semi-supervised method for automatically expanding the FrameNet corpus, using structural alignment of dependency-parse trees. They reported a success rate of approximately 33% of the projected annotations as completely correct. We implemented this algorithm in the Hebrew FrameNet tool to assist annotators; manual annotations are used to seed automatic annotations of candidate sentences from the Hebrew Wikipedia. We computed Word2Vec word embeddings on the Hebrew corpus, taking into account syntactic relations in the dependency trees as opposed to n-grams following (Levy and Goldberg, 2014). Candidate sentences are extracted using a lexical similarity measure, based on the computed word embeddings.

The rate of fully accurate annotation projections in our pipeline is ∼10%, much lower than the reported statistics in the English setting. Due to the relatively low accuracy of the automatic projection method, the projected annotations are manually reviewed before being accepted into the Hebrew FrameNet corpus.

5.1. Error Analysis

Analysis of the results of the method revealed that parser errors and lexical scoring errors introduce noise, which significantly degrades the accuracy of the projections. These two problems can be alleviated to some extent by retraining both the Word2Vec embeddings and the dependency parser on more syntactically- and lexically-diverse corpora.

---

ited to verbs, we are likely to add additional annotation layers in the future and adapt annotation practices more similar to those of SweFN.

An additional issue uncovered is that of word ambiguity. Some LUs are highly ambiguous in our Hebrew dataset. Since the alignment method considers lexical similarity to be twice as important as syntactic similarity, such ambiguity often leads to a high rate of false positives in the resulting projected annotations. A good example of this is the LU

עלה
Ala
boarded

in the BOARD_VEHICLE frame. In addition to "boarded", this word can also mean either "climbed" or "immigrated". This ambiguity results in projections of BOARD_VEHICLE annotations onto sentences which should be tagged as examples of another frames, such as CHANGE_POSITION_ON_A_SCALE (of which **climb.v** is an LU).

We are in the process of testing various ideas to improve the results of this method by addressing the issues found in the error analysis. Bases on the success in the English setting, we expect it to significantly accelerate the annotation effort in the future.

6. Software Tools and Infrastructure

As in other FrameNet projects, openness is a driving principle which influences architectural decisions (Forsberg, 2011). Accordingly we used freely available software tools and linguistic resources and published a well-documented API to access the Hebrew FrameNet. Our semantic annotation data is stored in a MongoDB[6] database, while our corpus, along with all morphological and syntactic data are stored in an ElasticSearch[7] cluster. The annotation tool is built on modern web technologies and is accessible through a REST API endpoint.

Additionally, the full annotation process, from the assignment of LUs to frames to the final annotation of exemplar sentences, is recorded in our database, while the annotation decisions are all subject to a review stage, before being finalized in a collaborative process.

7. Project Status and Future work

7.1. Project Status

As of October 2015, Hebrew FrameNet contains 3K LUs across 167 frames, with an average of 18 LUs per frame.

Of the 3K LUs, 43% are verbs, 46% are nouns, 7% are adjectives, 2% are adverbs. The final 2% are filled by other POS types (modal, copula, etc.).

There are 423 annotated exemplar sentences across 66 LUs, with an average of 6.41 sentences per LU. Before starting a more intense annotation campaign, we are now reviewing the linguistic issues faced during the initial annotation trial and assessing the potential to speed up annotation with semi-supervised expansion.

[6]https://www.mongodb.org/
[7]https://www.elastic.co/products/elasticsearch

7.2. Future Work

This work provides a basis on which an automatic SRL system for the Hebrew language can be constructed. We are investigating the current state-of-the-art English SRL system (Roth and Lapata, 2015) as a starting point.

We are planning to apply the ideas put forth by Boas (2005) as a form of evaluation for automatic SRL systems in multiple languages. We use an aligned Hebrew-English corpus and estimate the accuracy of an automatic annotator by comparing the annotations of an English SRL system with those produced by the Hebrew system. The Hebrew annotation is deemed correct when we verify that the same frames are identified across the English and Hebrew sides of an aligned sentence pair, and that the assigned roles can be aligned across languages using lexical alignment methods. We have collected an additional 600K sentence pairs over the corpus of Tsvetkov and Wintner (2012) and prepared a syntactically annotated version of this corpus. We are also investigating whether presenting such aligned sentences as part of the manual annotation process helps in the annotation process.

8. Conclusion

We presented the development process and the current status of the Hebrew FrameNet project. Using the Berkeley FrameNet as a basis and adopting a frequency-based approach to annotation, we are on our way to constructing a frame-semantic lexicon for the Hebrew language and assessing how this resource can be used to develop automatic SRL tools for Hebrew.

References

Adler, M. and Elhadad, M. (2006). An unsupervised morpheme-based hmm for hebrew morphological disambiguation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 665–672. Association for Computational Linguistics.

Adler, M., Netzer, Y. D., Goldberg, Y., Gabay, D., and Elhadad, M. (2008). Tagging a hebrew corpus: the case of participles. In *LREC*. Citeseer.

Ahlberg, M., Borin, L., Dannélls, D., Forsberg, M., Toporowska Gronostaj, M., Friberg Heppin, K., Johansson, R., Kokkinakis, D., Olsson, L.-J., and Uppström, J. (2014). Swedish framenet++ the beginning of the end and the end of the beginning.

Boas, H. C. (2005). Semantic frames as interlingual representations for multilingual lexical databases. *International Journal of Lexicography*, 18(4):445–478.

Boas, H. C. (2011). Constructing parallel lexicon fragments based on english framenet entries: Semantic and syntactic issues. *Multilingual Resources and Multilingual Applications. Proceedings of the German Society for Computational Linguistics and Language Technology (GSCL)*, pages 9–18.

Borin, L., Forsberg, M., Heppin, K. F., Johansson, R., and Kjellandsson, A. (2012). Search result diversification methods to assist lexicographers. In *Proceedings of the Sixth Linguistic Annotation Workshop*, pages 113–117. Association for Computational Linguistics.

Candito, M., Amsili, P., Barque, L., Benamara Zitoune, F., De Chalendar, G., Djemaa, M., Haas, P., Huyghe, R., Yannick Mathieu, Y., Muller, P., et al. (2014). Developing a french framenet: Methodology and first results.

Davies, M. (2010). The corpus of contemporary american english as the first reliable monitor corpus of english. *Literary and Linguistic Computing*, 25(4):447–464.

Erk, K., Kowalski, A., and Pado, S. (2003). The salsa annotation tool. In *Proceedings of the Workshop on Prospects and Advances in the Syntax/Semantics Interface*, pages 1–4. Citeseer.

Fillmore, C. J. and Baker, C. (2010). A frames approach to semantic analysis. *The Oxford handbook of linguistic analysis*, pages 313–339.

Fillmore, C. J., Baker, C. F., and Sato, H. (2002). The framenet database and software tools. In *LREC*.

Fillmore, C. J., (1982). *Frame semantics*, pages 111–137. Hanshin Publishing Co., Seoul, South Korea.

Forsberg, M. (2011). Green resources in plain sight: opening up the swefn++ project. *NEALT PROCEEDINGS SERIES VOL. 13*, page 7.

Friberg Heppin, K. and Voionmaa, K. (2012). Practical aspects of transferring the english berkeley framenet to other languages. In *SLTC 2012 The Fourth Swedish Language Technology Conference Lund, October 24-26, 2012 Proceedings of the Conference*, pages 28–29.

Fürstenau, H. and Lapata, M. (2012). Semi-supervised semantic role labeling via structural alignment. *Comput. Linguist.*, 38(1):135–171, March.

Goldberg, Y. and Elhadad, M. (2010). An efficient algorithm for easy-first non-directional dependency parsing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 742–750. Association for Computational Linguistics.

Goldberg, Y. and Elhadad, M. (2013). Word segmentation, unknown-word resolution, and morphological agreement in a hebrew parsing system. *Comput. Linguist.*, 39(1):121–160, March.

Goldberg, Y., Adler, M., and Elhadad, M. (2008). Em can find pretty good hmm pos-taggers (when given a good start). In *ACL*, pages 746–754.

Heppin, K. F. and Gronostaj, M. T. (2012). The rocky road towards a swedish framenet-creating swefn. In *LREC*, pages 256–261.

Itai, A. and Wintner, S. (2008). Language resources for hebrew. *Language Resources and Evaluation*, 42(1):75–98.

Kingsbury, P. and Palmer, M. (2002). From treebank to propbank. In *In Language Resources and Evaluation*.

Levy, O. and Goldberg, Y. (2014). Dependencybased word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 302–308.

Melli, G., Wang, Y., Liu, Y., Kashani, M. M., Shi, Z., Gu, B., Sarkar, A., and Popowich, F. (2005). Description of squash, the sfu question answering summary handler for the duc-2005 summarization task. *safety*, 1:14345754.

Netzer, Y., Adler, M., Gabay, D., and Elhadad, M. (2007). Can you tag the modal? you should. In *Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources*, pages 57–64. Association for Computational Linguistics.

Ohara, K. (2013). Toward constructicon building for japanese in japanese framenet. *Veredas: Frame Semantics and Its Technological Applications*, 17(1):11–28.

Petruck, M. R. (2005). Towards hebrew framenet. *Kernerman Dictionary News*, 13:12.

Petruck, M., (2009). *Typological Considerations in constructing a Hebrew FrameNet*, pages 183–208. Mouton de Gruyter, Berlin.

Roth, M. and Lapata, M. (2015). Context-aware frame-semantic role labeling. *Transactions of the Association for Computational Linguistics*, 3:449–460.

Ruppenhofer, J., Ellsworth, M., Petruck, M. R. L., Johnson, C. R., and Scheffczyk, J. (2010). FrameNet II: Extended Theory and Practice. Technical report, International Computer Science Institute in Berkeley, 09.

Shen, D. and Lapata, M. (2007). Using semantic roles to improve question answering. In *EMNLP-CoNLL*, pages 12–21.

Subirats, C. (2009). Spanish framenet: A frame-semantic analysis of the spanish lexicon. *Multilingual FrameNets in Computational Lexicography. Methods and Applications, Mouton de Gruyter, Berlin/New York*, pages 135–162.

Surdeanu, M., Harabagiu, S., Williams, J., and Aarseth, P. (2003). Using predicate-argument structures for information extraction. In *Proceedings of*

the 41st Annual Meeting on Association for Computational Linguistics-Volume 1, pages 8–15. Association for Computational Linguistics.

Torrent, T. T. and Ellsworth, M. (2013). Behind the labels: criteria for defining analytical categories in framenet brasil. *Veredas: Frame Semantics and Its Technological Applications*, 17(1):44–65.

Tsarfaty, R. and Goldberg, Y. (2008). Word-based or morpheme-based? annotation strategies for modern hebrew clitics. In *LREC*.

Tsvetkov, Y. and Wintner, S. (2012). Extraction of multi-word expressions from small parallel corpora. *Natural Language Engineering*, 18(04):549–573.

Wu, D. and Fung, P. (2009). Semantic roles for smt: a hybrid two-pass model. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 13–16. Association for Computational Linguistics.