Verb Alternations and Their Impact on Frame Induction

Esther Seyffarth
Dept. of Computational Linguistics
Heinrich Heine University
Düsseldorf, Germany
Esther.Seyffarth@hhu.de

Abstract

Frame induction is the automatic creation of frame-semantic resources similar to FrameNet or PropBank, which map lexical units of a language to frame representations of each lexical unit's semantics. For verbs, these representations usually include a specification of their argument slots and of the selectional restrictions that apply to each slot. Verbs that participate in diathesis alternations have different syntactic realizations whose semantics are closely related, but not identical. We discuss the influence that such alternations have on frame induction, compare several possible frame structures for verbs in the causative alternation, and propose a systematic analysis of alternating verbs that encodes their similarities as well as their differences.

1 Introduction

One of the aims of natural language processing is to access and process the meaning of texts automatically. For tasks like question answering, automatic summarization, or paraphrase detection, annotating the text with semantic representations is a useful first step. A good annotation represents the semantics of each element in the text as well as the relations that exist between the elements, such as the relations between a verb and its arguments and adjuncts.

Frame semantics, founded on work by Fillmore (1968), Minsky (1974), Barsalou (1992), and others, provides a powerful method for the creation of such representations. A frame-semantic representation of a concept expresses the attributes that contribute to its semantics as functional relations in a recursive attribute-value structure. Frame-semantic lexical resources, such as FrameNet (Ruppenhofer et al., 2006) or PropBank (Palmer et al., 2005), map lexical units of a language to frames, which are described in terms of the frame elements, or roles, that are central to the concept.

Since the manual compilation of such resources is time-consuming, costly, and error-prone, much can be gained from the use of semi-supervised or unsupervised methods. The process of creating a frame lexicon automatically is known as frame induction.

Unsupervised frame induction draws on observations about each lexical item's behavior in a corpus to create a frame lexicon. If two different lexical units can evoke the same frame, they share certain semantic properties, and the frame inducer has to determine the amount of semantic overlap between them based on observable and latent features of each of the lexical items. In this paper, we discuss some of the problems that occur in frame induction when lexical units have a relatively large semantic overlap, but are not close enough to each other to be treated as total synonyms.

While there are several types of frame-evoking predicates, we focus here on verbs and the frames they evoke. We assume that the meaning of a sentence can be expressed using the frame representation of the sentence's root verb. The semantic contribution of all other elements in the sentence is then specified with respect to their relation to the root verb. Therefore, it is crucial to assign the correct semantic frame to the root verb.

Diathesis alternations are a phenomenon that is observed when verbs can occur with more than one valency pattern. In some alternations, the different uses of the participating verbs introduce changes in the semantics as well (Levin, 1993). This is particularly relevant for alternations that change the *Aktionsart* of the verb, such as the causative-inchoative alternation (Levin, 1993, Chapter 1.1.2). Verbs in this alternation can be used either transitively or intransitively, where the transitive use adds a causative meaning to the sen-

Proceedings of NAACL-HLT 2018: Student Research Workshop, pages 17–24
New Orleans, Louisiana, June 2 - 4, 2018.

tence that is not present in the intransitive use, as in the sentences in (1).

(1) a. Mary opened the door.
b. The door opened.

In this paper, we argue that sentence pairs like this differ in their semantics to a degree that warrants their annotation with different frames, but in a way that still expresses their semantic relation to each other. We compare different possible frame representations for verbs in the causative alternation and propose some guidelines for the way a frame inducer should analyze alternating verbs.

2 The Causative Alternation

The causative alternation is characterized by a role switch (McCarthy, 2001) between transitive and intransitive uses of participating verbs, as shown in the sentences in (2). The role switch determines the position of the semantic THEME, the rent, as syntactic object (in (2-a)) or subject (in (2-b)). An AGENT is only present in the transitive sentence.

(2) a. They have increased my rent.
b. My rent has increased.

Both sentences describe a situation which results in the rent being higher than before.

Transitive uses of verbs in the causative alternation can be paraphrased as "cause to [verb]-intransitive" (see Levin, 1993, p. 27). For instance, sentence (2-a) can be paraphrased as "They have caused my rent to increase."

In some cases, this type of paraphrase is not completely synonymous with the causative use of the verb; for a discussion of such cases, see e.g. Dowty (1991, pp. 96-99) and Cruse (1972). These authors claim that a difference between scenarios of direct and indirect causation renders some paraphrases ungrammatical or different in meaning from the original sentences with causative verbs. However, for the purposes of this paper, we focus on the regular cases, where the causative use and the paraphrased form do express the same meaning.

Dowty (1991) decomposes causative verbs as [x CAUSE [BECOME y]], where an inchoative event [BECOME y] is embedded in a causative one. Thus, for verbs in the causative alternation, the transitive sentence has a more complex semantic structure. This is not the case for verbs outside the alternation: Sentences (3-a) and (3-b) below have the same semantic complexity, and one does not describe something that causes the other.

(3) a. I'm eating an apple.
b. I'm eating.

If (3-a) is true, then (3-b) must necessarily also be true, but there is no causation relationship between the sentences. Instead, (3-b) is a less specific description of the same situation that is also described by (3-a).

Like Rappaport Hovav (2014), we assume that a causative sentence entails its inchoative counterpart. In both sentences in (2), the verb *increase* describes an event that affects the semantic THEME; if it is true that a CAUSE is responsible for the event that increases the rent, then a statement that does not contain a CAUSE, but otherwise describes the same event, is necessarily also true. In other words, it is impossible for (2-a) to be true without (2-b) also being true.

2.1 Possible Frame Representations

A frame representation of the sentences in (2) that focuses on their shared semantics is given in Figure 1. A similar structure could also be used to represent the sentences in (3).

$$\begin{bmatrix} \textit{increase} & \\ \text{CAUSE} & \text{they / } \emptyset \\ \text{THEME} & \text{my rent} \end{bmatrix}$$

Figure 1: Frame representation that is agnostic towards the question of causativity or inchoativity.

This analysis assigns an *increase* frame, independent of the syntactic realization and observed arguments. If a cause is given, it is included; otherwise, it is not, and the CAUSE slot is unfilled.

A disadvantage of this choice is that the structure does not differentiate between causative and inchoative uses of alternating verbs, except by the presence and absence of the CAUSE slot. As discussed in the previous section, this is acceptable for sentences like (3), but undesirable for sentences like (2).

The representation in Figure 1 is similar to the structure used in PropBank (see Palmer et al., 2005, p. 77). There, both uses of an alternating verb are associated with one shared frameset. The PropBank annotation makes no difference between the sentences in (2) and the sentences in (3).

The FrameNet frame hierarchy includes a mechanism to connect frames with an "is causative of" relation and its inverse, "is inchoative of" (Ruppenhofer et al., 2006, p. 85). Different frames for alternating verbs are stored in the database and connected by these relations. Figure 2 gives the FrameNet representations for the sentences in (2). For a discussion of these representations, see Ruppenhofer et al. (2006, p. 12 and pp. 15-16).

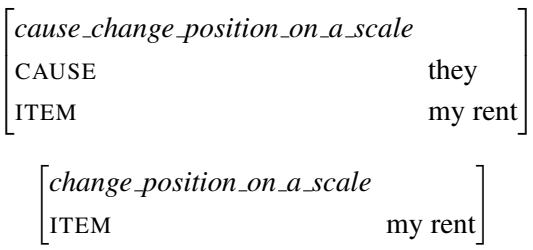

Figure 2: FrameNet representations for the sentences in (2).

These frames show a clear conceptual separation between the causative meaning of *increase* and its inchoative counterpart. The two frames are linked by the relations mentioned above, and their names hint at the semantic connection that exists between them.

However, it is not obvious from the separate frames that the two sentences are connected by an entailment relationship. If a sentence like (2-a) is observed, the first representation will be chosen; if a sentence like (2-b) is observed, the second representation will be chosen. Without referring to the FrameNet frame hierarchy, it will not be possible to recognize that sentence (2-a) necessarily entails sentence (2-b).

A frame representation that encodes this relationship more visibly is presented in Figure 3. That representation is analogous to the decomposition for causative verbs given by Dowty (1991), which is [x CAUSE [BECOME y]]. The specific sentence (2-a) is represented as [[they do something] CAUSE [BECOME [my rent is higher]]].

Figure 3 represents the event as a *causation* frame. The EFFECT slot of that frame is linked to the content of the [BECOME y] part of the decomposition, which is an *inchoation* frame. The RESULT of the inchoation is the formula's y.

These representations are inspired by Osswald and Van Valin Jr (2014). They visibly express the relationship between the frames and are consistent with the decompositional event structures used by Rappaport Hovav and Levin (1998) in their discussion of alternating verbs. The frame in Figure 3 provides a slot for the result state of the *inchoation* frame, here expressed as a frame of the type *higher*. Osswald and Van Valin Jr (2014) suggest that the semantics of verbs that express a gradual change of state should be represented in terms of the initial and final state and a specification of the relation between them. Thus, the frame should state that the rent had an initial and final amount, and the meaning of "They have increased my rent" is that the states are connected by a GREATER_THAN relation.

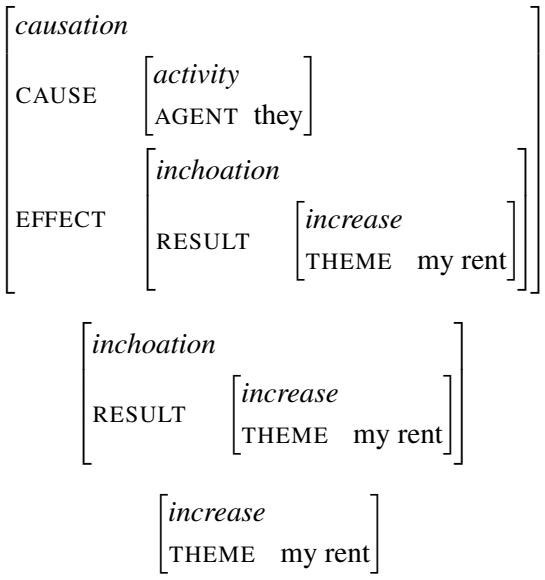

Figure 3: Decompositional frame representation of the sentences in (2), inspired by Osswald and Van Valin Jr (2014).

While the decompositional approach has advantages over the other representations shown above, it is difficult to use in a computational setting, where this level of detail in the expression of gradual changes is often neither attainable nor desirable. Additionally, only a subset of the alternating verbs specify a change of state that can be represented like this. Other verbs (*roll*, *fly*, ...) describe movement or induced action, and we aim at a representation that works universally for alternating verbs.

This is why we prefer the frame representations in Figure 4. They are related to the decompositional frames in Figure 3, but do not make use of the innermost embedding layer.

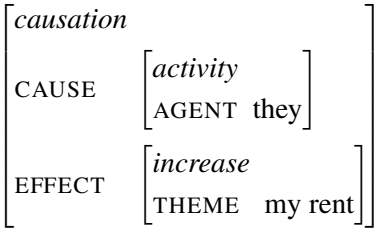

Figure 4: Frame representation for (2-a)

Compared to the frames in Figure 3, this structure is more compact. Like them, it acknowledges that the transitive use of *increase* adds not only an additional argument slot, but also a causative meaning that "wraps around" the *increase* frame.

This analysis is consistent with that of Rappaport Hovav (2014), who supports an analysis of alternating verbs that assumes they are lexically associated with their internal arguments only.

A question that is often discussed in research on diathesis alternations is that of underlying and derived forms; for an overview of perspectives on this question that are discussed in the literature, see e.g. Piñón (2001). The assumption is that knowledge of the derivation processes involved in diathesis alternations can inform the semantic analysis. Here, we do not claim that one use of alternating verbs is derived from the other. We do however note that inchoative sentences specify fewer participants of the event they describe, since they do not contain an argument slot for the entity that has causative control over the situation – unlike causative sentences, where these entities are part of the frame.

3 Frame Induction

Verbs that participate in diathesis alternations are likely to be observed in different syntactic configurations in a corpus. For a frame inducer, it is important to distinguish between these cases on the one hand, and homographs whose semantics are not (or less closely) related on the other hand. For instance, the system should recognize that the verb *run* in the sentences in (4) has two meanings that are less closely related than the transitive and intransitive meanings of *open* or *increase*.

(4) a. I ran ten miles this morning.
b. I started the program two days ago, but it's still running.

A frame induction system that uses the observed syntactic subjects and objects as indicators for the correct frame will initially treat both *run* and *increase* in the same way: It notices that the (syntactic) slot fillers in the observed sentences are not homogeneous enough to assign a single frame for both uses, and therefore assumes different frames. In the case of *run*, this is desirable, but in the case of verbs that participate in diathesis alternations, it results in an unnecessarily large number of frames in the lexicon, and the relationship between the semantics of transitive and intransitive uses of the verbs would not be represented.

The architecture we propose is one that combines existing probabilistic approaches to frame induction with an additional level of analysis that makes sure the learned frames are distinguished with regard to their participation in the alternation.

Thus, the system will be able to make an informed decision whether different uses of a verb are due to non-obligatory arguments, as in the sentences in (3); due to polysemy, as in the sentences in (4); or due to a diathesis alternation, as in the sentences in (2). In the first case, a single frame must be assigned to all uses of the verb, with optional arguments as observed in the data. In the second case, different semantic frames need to be assigned, because the meanings of the different uses of the verb do not have a large enough overlap to be subsumed into the same frame. Finally, if the frame inducer has correctly identified the verb as one which does participate in a diathesis alternation, the alternating frame structure of that alternation can be used to create frames for the particular verb in a way that relates the meanings of the different uses to each other.

In order to distinguish between alternating verbs and non-alternating verbs, the frame inducer can employ one of the methods for alternation identification that have been proposed in the literature. Early approaches to this relied on WordNet or similar resources to identify the slot overlap of different subcategorization frames (McCarthy and Korhonen, 1998; Schulte im Walde, 2000; McCarthy, 2001) or on the evaluation of latent semantic properties of the slot fillers, approximated using manually-defined rules (Stevenson and Merlo, 1999). More recently, distributional representations of slot fillers have been used to create clusters whose overlap can be used for the distinction (Baroni and Lenci, 2009; Sun and Korhonen, 2009; Sun et al., 2013). We propose that distributional methods be used by the frame inducer, in order to

minimize the dependence on manually-created resources.

The frame inducer can store the different frames with cross-references between the alternating variants. This ensures that the core semantics that describes the event and its result state is stored in one place only – the inchoative frame –, while the lexicon entry for the causative frame can access the inchoative frame to build the causative semantics around it.

Frame induction is notoriously difficult to evaluate quantitatively. Since the structure of semantic frames depends on a number of factors, including the perspective on the event, the desired level of granularity of the description, and the application context in which the frame representation is to be used, there is no single, objectively correct frame structure for a given event. However, to get a general indicator for the performance of the system, one can evaluate the induced frames against resources like SemLink (Bonial et al., 2013) to calculate the amount of overlap between the induced frame hierarchy and manually-created hierarchies. Note that among other data sources, SemLink contains annotations from FrameNet, where causative and inchoative uses of alternating verbs are associated with different frames, but also annotations from PropBank, where no distinction is being made. We leave a detailed specification of an optimal evaluation setup to future work.

4 Discussion

The frame representation we propose has a number of advantages in the context of the semantic tasks mentioned in the introduction.

First, it expresses overtly the fact that the main difference between the semantics of the causative and inchoative use of each verb is the added causative meaning in the transitive form. This helps because inferences about the result state of the event should be consistent across both transitive and intransitive uses, and the structure we propose that embeds the inchoative frame into the causative one means that all relevant information about the result state can be derived from the embedded frame as needed.

A similar point applies to the question of entailment. As mentioned in Section 2, a transitive sentence with a verb in the causative alternation as the root entails the intransitive version of that sentence. With our frame structure, this entailment is expressed in the frame already, since the truth of the embedded frame contributes directly to the truth of the whole frame.

We also look to frame-semantic parsing in choosing this frame structure. The paraphrase of sentence (2-a) as "They have caused my rent to increase" is an illustration of the benefit of our analysis: A frame-semantic annotation of the paraphrased sentence should include a structure that is like the one we propose for the causative use of *increase*.

Contrast this with the way a parser that relies on FrameNet frames (see Figure 2 above) would analyze the paraphrase: FrameNet will assign a *causation* frame to the verb *cause*, with the increasing of the rent being specified in the EFFECT slot of the *causation* frame. The resulting mismatch does not encode the fact that the paraphrase carries the same meaning as the causative use of the verb.

4.1 Productivity of the Causative Alternation

We view the causative alternation as an open class. Levin (1993) lists verbs that regularly participate in the alternation and verbs that cannot alternate, but we assume that there are also verbs that can be used in an alternating way to produce a novel causative construction. An example is given in (5).

(5) The bad weather decided him to take the car.[1]

Sentences like this are indicative of a certain productivity of the causative alternation. When a verb that usually only occurs in intransitive forms is used transitively, we can assume that this change adds a causative dimension. Our frame structure allows us to embed any verb frame into a causative frame to create the frame in Figure 5 for sentence (5). Note that the role of AGENT is being filled by different entities in the subframes CAUSE and EFFECT. This is different from the frame given in Figure 4 because *decide* does not usually select a THEME.

[1]This sentence is a heavily adapted version of a passage from Chapter 4 of David Lodge's novel *The British Museum is Falling Down* (1965).

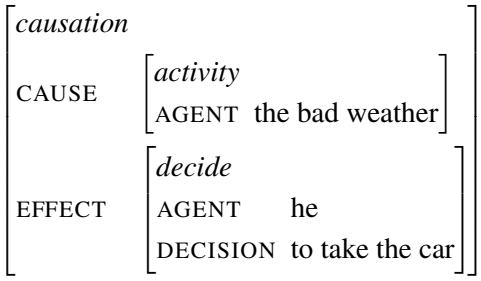

Figure 5: Frame representation for (5)

Separating the core semantics of each verb from the alternating mechanism results in a frame lexicon that is more flexible and therefore better equipped to deal with unseen constructions than the alternative analyses we have discussed.

4.2 Extending the Analysis to Other Alternations

So far, we have focussed on the causative alternation, but similar analyses are also conceivable for other alternations. In the conative alternation (Levin, 1993, p. 41), one of the syntactic environments in which the alternating verb can occur describes an attempted action and lacks entailment of success or completion of the action described. A representation that uses embedded frames, analogous to the ones described above, may look like the frame in Figure 6. The frame represents the semantics of the sentence in (6).

(6) Mary cut at the rope.

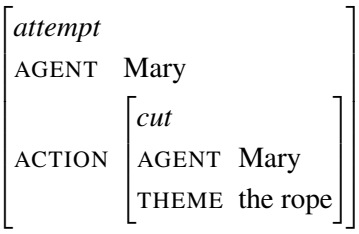

Figure 6: Frame representation for (6)

As in the causative alternation, a precondition for such an analysis is knowledge of the alternatability of a sentence's root verb and access to rules that control the creation of the embedding frames. Having access to rules that govern the creation of such complex frames allows the frame inducer to represent the conative meaning by combining the *attempt* part of the meaning with a reference to the *cut* frame stored in the lexicon.

5 Future Work

We will employ the strategies outlined here to develop a frame inducer that is sensitive to slight differences in semantics, such as the ones observed in diathesis alternations, and that is equipped to handle these differences in a systematic way.

The resulting system will be semi-supervised, since the productive rules for complex frames like the ones presented above can be created manually. The frame inducer then has the task of identifying verbs that participate in the alternations to which the pre-defined rules apply, and of storing the different uses of the alternating verbs in an appropriate way. Using cross-references to link a causative frame to the inchoative frame embedded in it will ensure that the lexicon can be kept to a small size while providing as much expressive power as necessary for the semantic distinctions at hand.

While we are optimistic about the system suggested here for the treatment of alternating verbs, we are aware that frame induction is not a trivial task. Particularly, we note that it is difficult, if not impossible, to argue that one specific frame representation of a concept is correct while another is incorrect. The way frames are being formed to represent semantics is highly subjective, and the decisions one makes always have to depend on the purpose for which the frame lexicon is being created.

However, we find it important to identify ways in which the induction of frames may be systematized. We are convinced that the idea of storing complex frames in the lexicon that embed semantically related frames is useful for the analysis of diathesis alternations as well as similar phenomena.

An important part of working on frame induction is the exploration of different ways to evaluate the induced frame hierarchy. In addition to the approach mentioned in Section 3, where the overlap of the new hierarchy and some manually-built resource is being determined, we are also interested in the possibility of extrinsic evaluation. For instance, a question-answering task may be set up and tested using the output of versions of the frame inducer with and without alternation-specific functions, in order to enable a comparison of each system's success in this type of application context.

6 Conclusion

Diathesis alternations pose a challenge to the creation of frame-semantic resources, because they license verb uses that are closely related, but not similar enough to be treated as synonyms. In this paper, we argued that alternating verbs should be represented with frames that highlight this relationship while also specifying the differences between the alternating verb uses.

Unlike the frames defined in PropBank and FrameNet for alternating verbs, our proposed analysis involves the embedding of one frame (the "core meaning" of the verb) into another (the causation frame that "wraps around" the core meaning in transitive uses). We find that this analysis is consistent with the appropriate analysis when parsing a sentence like (5), where a verb that may not be stored in the lexicon as having a causative property (here, the verb *decide*) is used exactly like verbs that participate in the alternation. We wish to minimize the difference between such analyses that are conducted at parsing time and the entries in the frame lexicon.

The successful induction of frames of the type described here depends on the successful identification of the alternation. If the frame inducer mistakes a verb that has several unrelated meanings for a verb that participates in the alternation, the system will create frames that are inappropriate for that verb. For instance, the sentences in (4) should not be analyzed with frames that embed one another, since the meanings of their root verbs are too dissimilar and there is no entailment relation between the different uses of *run*.

A frame induction system that follows the suggestions outlined in this paper will be able to represent the semantics of alternating verbs (and phenomena that exhibit similar behaviors) in a way that not only clarifies the semantic relations that exist between the different uses of the verbs, but is also consistent with annotations that are created in the context of frame-semantic parsing.

Acknowledgments

The work presented in this paper was financed by the Deutsche Forschungsgemeinschaft (DFG) within the CRC 991 "The Structure of Representations in Language, Cognition, and Science". I would like to thank Laura Kallmeyer and all reviewers of this work for their valuable comments and suggestions.

References

Marco Baroni and Alessandro Lenci. 2009. One Distributional Memory, Many Semantic Spaces. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 1–8, Athens, Greece. Association for Computational Linguistics.

Lawrence W. Barsalou. 1992. Frames, concepts, and conceptual fields. In A. Lehrer and E. Kittay, editors, *Frames, Fields, and Contrasts: New Essays in Semantic and Lexical Organization*, pages 21–74. Lawrence Erlbaum Associates, Hillsdale, N.J.

Claire Bonial, Kevin Stowe, and Martha Palmer. 2013. Renewing and Revising SemLink. In *Proceedings of the 2nd Workshop on Linked Data in Linguistics (LDL-2013): Representing and linking lexicons, terminologies and other language data*, pages 9–17. Association for Computational Linguistics.

Donald A. Cruse. 1972. A Note on English Causatives. *Linguistic Inquiry*, 3(4):520–528.

David R. Dowty. 1991. The semantics of aspectual classes of verbs in English. In *Word meaning and Montague grammar*, pages 37–132. Springer.

Charles J. Fillmore. 1968. The Case for Case. In E. Bach and R. T. Harms, editors, *Universals in Linguistic Theory*, pages 1–88. Holt, Rinehart and Winston.

Beth Levin. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago press.

Diana McCarthy. 2001. *Lexical Acquisition at the Syntax-Semantics Interface: Diathesis Alternations, Subcategorization Frames and Selectional Preferences*. Ph.D. thesis, University of Sussex, UK.

Diana McCarthy and Anna Korhonen. 1998. Detecting Verbal Participation in Diathesis Alternations. In *COLING 1998 Volume 2: The 17th International Conference on Computational Linguistics*.

Marvin Minsky. 1974. A Framework for Representing Knowledge. Technical report, Massachusetts Institute of Technology, Cambridge, MA, USA.

Rainer Osswald and Robert D. Van Valin Jr. 2014. FrameNet, Frame Structure, and the Syntax-Semantics Interface. In Thomas Gamerschlag, Doris Gerland, Rainer Osswald, and Wiebke Petersen, editors, *Frames and Concept Types*, pages 125–156. Springer.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–106.

Christopher Piñón. 2001. A Finer Look at the Causative-Inchoative Alternation. *Semantics and Linguistic Theory*, 11:346–364.

Malka Rappaport Hovav. 2014. Lexical content and context: The causative alternation in English revisited. *Lingua*, 141:8–29.

Malka Rappaport Hovav and Beth Levin. 1998. Building Verb Meanings. In *The projection of arguments: Lexical and compositional factors*, pages 97–134. University of Chicago Press.

Josef Ruppenhofer, Michael Ellsworth, Miriam R.L. Petruck, Christopher R. Johnson, and Jan Scheffczyk. 2006. *FrameNet II: Extended Theory and Practice*. International Computer Science Institute, Berkeley, California. Distributed with the FrameNet data.

Suzanne Stevenson and Paola Merlo. 1999. Automatic Verb Classification Using Distributions of Grammatical Features. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 45–52. Association for Computational Linguistics.

Lin Sun and Anna Korhonen. 2009. Improving Verb Clustering with Automatically Acquired Selectional Preferences. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 638–647. Association for Computational Linguistics.

Lin Sun, Diana McCarthy, and Anna Korhonen. 2013. Diathesis alternation approximation for verb clustering. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 736–741. Association for Computational Linguistics.

Sabine Schulte im Walde. 2000. Clustering Verbs Semantically According to their Alternation Behaviour. In *COLING 2000 Volume 2: The 18th International Conference on Computational Linguistics*, pages 747–753. Association for Computational Linguistics.