Contrast Sets for Stativity of English Verbs in Context

Daniel Chen & Alexis Palmer
Department of Linguistics
University of Colorado
{daniel.chen-1, alexis.palmer}@colorado.edu

Abstract

For the task of classifying verbs in context as **dynamic** or **stative**, current models approach human performance, but only for particular data sets. To better understand the performance of such models, and how well they are able to generalize beyond particular test sets, we apply the **contrast set** (Gardner et al., 2020) methodology to stativity classification. We create nearly 300 contrastive pairs by perturbing test set instances just enough to change their labels from one class to the other, while preserving coherence, meaning, and well-formedness. Contrastive evaluation shows that a model with near-human performance on an in-distribution test set degrades substantially when applied to transformed examples, showing that the stative vs. dynamic classification task is more complex than the model performance might otherwise suggest. Code and data are freely available.[1]

1 Introducing stativity

Aspectual properties of verbs, and the clauses they inhabit, have the potential to support a range of natural language processing tasks, such as event ordering (Modi and Titov, 2014) and temporal relation classification (Costa and Branco, 2012), as well as contributing to speaker choices around situational construal (Trott et al., 2020). At the same time, verb and situational aspect are a complex set of interacting properties, in which the meaning of the verb, the nature of its arguments, adverbial modifiers, and grammatical features such as verb tense and nominal definiteness can all play a role in determining the aspectual make-up of a clause.

This sensitivity of aspectual categorization to small shifts in linguistic form is one reason that automatic prediction of aspectual classes is an especially challenging computational problem. In this paper we explore the stability of automatic classification for one particular facet of aspect: stativity of English verbs. Stativity reflects the degree to which a verb represents a static situation versus a situation that reflects some degree of dynamicity. Dynamic verbs typically involve some change of state. (1-3) below show examples of the three classes relevant for our study: DYNAMIC verbs, STATIVE verbs, and verbs for which annotators CANNOT_DECIDE.

(1) Table 7 <u>shows</u> results from the latest experiments. STATIVE

(2) Dr. Smith <u>showed</u> her students how to work with the new GPUs. DYNAMIC

(3) The earlier paper <u>shows</u> the effectiveness of incorporating linguistic features. CANNOT_DECIDE

In (1), the table is static, and the results exist in the table; no change of state is indicated. (2) highlights the dynamic sense of *show*, in which the Agent is giving a demonstration. (3) allows two readings. In the STATIVE reading, the result about linguistic features is a static property of the paper; it simply exists in the paper. In the DYNAMIC reading, the paper demonstrates the effectiveness of linguistic features through an argument that develops and progresses over the course of the paper.

Most verbs in English have a strong predominant category (stative or dynamic), yet allow for variable interpretation, depending on context. A smaller number of verbs (e.g. *show*), are highly flexible, with no strong statistical tendency in either direction (Friedrich and Palmer, 2014a; Falk and Martin, 2016). Because of this variability, automatic classification of aspectual properties requires contextual input and instance-level classification.

To better understand the ability of systems to automatically determine stativity, we produce contrast sets (section 2) for English verb stativity with 298 transformed instances. The contrast set instances and their ground-truth labels are extracted from the SitEnt corpus (section 2.2), and the transformed instances are produced using a range of

[1] https://github.com/dchensta/se_contrast

Proceedings of the 29th International Conference on Computational Linguistics, pages 4028–4036
October 12–17, 2022.

linguistically-motivated transformation strategies, detailed in section 3.

Using a standard modeling configuration (section 4), we show that classification performance on the transformed instances is substantially lower than on the original instances. According to Friedrich (2017), observed annotator agreement for this task ranges from 79% to 82%. On the original instances, the model achieves micro-averaged accuracy of nearly 80%, approaching human agreement for this task.[2] On the transformed instances, micro-averaged accuracy is well below 60%.

2 Building contrast sets for stativity

For many NLP tasks, neural models, especially those built on large language models, have been shown to be sensitive to annotation artifacts in the data on which they are trained and evaluated. High performance of classifiers often hinges on preserving these properties at evaluation time, and testing on out-of-distribution data can result in such dramatic performance decreases that the models no longer reliably perform the task they have been trained to do (Gururangan et al., 2018; Poliak et al., 2018; Geva et al., 2019, among others). These findings have given rise to methodologies for more careful evaluation of classification capability. Several different methods for improved evaluation have been proposed (Ribeiro et al., 2020; Gardner et al., 2020, among others).

2.1 Contrast sets

In this work, we follow the contrast set methodology (Gardner et al., 2020). The core idea is to create **contrastive** evaluation data sets by having experts make small perturbations to instances in the original test sets. In our case, we vary the lexical aspect of the main verb so that the preferred label changes from DYNAMIC (4) to STATIVE (5):

(4) Mary ran the Buenos Aires Marathon.

(5) Mary was a participant in the Buenos Aires Marathon.

These perturbations need to strike a delicate balance. The changes should be large enough to change the gold label for the instance, yet small enough to retain meaning, coherence, and validity. We also aim to use a variety of strategies, so as not to introduce *new* unintended annotation artifacts.

[2]Note that cross-validation accuracy on the much larger training set ranges from 77-80%.

	DYN	STAT	CD
Training	18,357	15,507	8,445
Test	376	217	57
Contrast: Test_Orig	172	120	0
Contrast: Test_Trans	120	172	0

Table 1: Distribution of DYNAMIC (DYN), STATIVE (STAT), and CANNOT_DECIDE (CD) labels.

Once contrast sets have been built, we compare the performance of the model in question on the *transformed* test instances (with their new labels) to the performance of the same model on the *original* version of those same instances. Significant performance degradation on the transformed test data calls into question whether the model has learned to classify the phenomena modeled in the annotated training data. The **contrast set** consists of the paired original and transformed test instances.

2.2 Data

We use data from the SitEnt (situation entities) corpus (Friedrich and Palmer, 2014b; Friedrich, 2017). The corpus[3] combines data from MASC (Ide et al., 2008) with Wikipedia texts, creating a collection of documents from 13 different genres. Texts are segmented into clauses, and each clause is triply-annotated for stativity of the main verb, genericity of the main referent, habituality of the situation described, and finally, a clause-level situation type label (following Smith (2003), these labels distinguish between events, states, generics, generalizing sentences, facts, propositions, reports, questions, imperatives, and undecided). Gold labels come from a majority vote across the three annotators.

For model training, we use Friedrich et al. (2016)'s original training split, which consists of 324 documents, with 42,309 clauses. As a basis for building contrast sets, we select four documents from the original test set, each from a different genre: news, essay, journal, and Wikipedia. For each sentence, we create one contrast set by transforming the first clause in the sentence.[4] No contrast sets are created for clauses labeled CAN-

[3]https://github.com/annefried/sitent/tree/master/annotated_corpus

[4]To test the viability of using first clauses only, we created contrast sets for all clauses in the Wikipedia document and compared classifier performance for the full document vs. only initial clauses. There was no significant difference.

NOT_DECIDE. We produce 292 contrast sets.[5]

Table 1 shows the distribution of the three labels for our data set. **Test** refers to the four selected documents; **Contrast: Test_Orig** refers to the original versions for the 292 contrast sets; and **Contrast: Test_Trans** refers to the transformed counterparts of the original instances, with flipped labels. Note that DYNAMIC to STATIVE transformations outnumber STATIVE to DYNAMIC transformations.

3 Transformation strategies

After building the contrast sets, we perform an analysis of the linguistic properties of the various strategies used. Most transformations hinge on the main verb, either replacing the lexical item or changing the role of the verb so that it moves to a different structural and semantic configuration.

3.1 DYNAMIC –> STATIVE

Example sentences showing the DYNAMIC to STATIVE transformations can be found in Table 2.

1. THOUGHT VERB - Demote a dynamic verb from main to secondary verb by moving it into the subordinate THEME role for verbs of thinking, believing, or feeling.
2. COPULA - Replace main verb with a simple predication headed by a copular verb.
3. DESCRIPTIVE VERB - Replace the dynamic action with a descriptive verb, effectively reconfiguring the dynamic action as stative properties of the subject noun.
4. LIGHT VERB - Use the possessive [6] light verb construction with *have* to make *have* the new main verb.
5. SEMI-MODAL - Use a semi-modal verb (e.g. *need to*, *ought to*) marking deontic modality, which concerns the speaker's requirements and desires, as a "thought" or "emotion" from the speaker, who can be an unspecified authority with no referent.
6. DOWNGRADE TO PPL - Remove main verb from the clause by transforming it into a perfect passive participle that favors a descriptive, adjectival reading over a verbal reading.
7. ORDER - Switch the order of the clauses and insert a descriptive verb as the new main verb.

[5] This number is on par with the data sets described in Gardner et al. (2020), which range from 70 to 1000 contrast sets per task.

[6] verbs of possession have a STATIVE reading

3.2 STATIVE –> DYNAMIC

Example sentences showing each of the STATIVE to DYNAMIC transformations can be found in Table 3.

1. NEW PARTICIPANT - Choose a synonymous verb that introduces an agent who participates in a dynamic synonym of the original verb.
2. INSERT VERB - Replace a stative verb (typically the copula) with a dynamic verb or insert a dynamic verb as the new main verb, of which the original stative verb is a dependent.
3. BECOMING - Replace standard copula with an inflected form of the verbs *to become* or *to get*, and their synonyms. This preserves the copula's predicating structure while reformulating the event as dynamic.
4. UPGRADE - Upgrade a perfect passive participle or subordinate STATIVE verb to the main verb of the clause by adding a helping verb or deleting the main STATIVE verb.
5. HEAVY VERB - Replace a light verb construction like *have* with a heavy, dynamic verb.

4 Model, results, and discussion

Having built contrast sets, we now evaluate performance compared to the original instances.

4.1 Model

Our straightforward classification model first learns a contextualized representation for each clause using BERT (Devlin et al., 2019), followed by a regression layer to classify the clause representations as DYNAMIC, STATIVE, or CD. The logistic regression model is trained using the liblinear solver and L2 regularization. Running 5-fold cross-validation over the full training set, using the trained logistic regression model, yielded accuracy scores ranging from 77.7% to 80.13%, only slightly below observed human agreement, which Friedrich (2017) reports as ranging from 79% to 82%.

4.2 Results: Classifying contrast sets

Table 4 shows the model's performance on the original test set instances and the transformed instances. These figures include all clauses for the Wikipedia text and only initial clauses for the other three texts, and only clauses whose original labels are either STATIVE or DYNAMIC. On the original instances, the model achieves micro-averaged accuracy of nearly 80%, approaching human agreement for this task. On the transformed instances, micro-averaged accuracy is well below 60%.

Strategy	Original Instance (DYN)	Transformed Instance (ST)	#
THOUGHT VERB	During that time, the panel **said**,	During that time, the panel **believed** that	55
COPULA	Although it **affected** Youngstown and the surrounding area more than it affected other regions,	Although it **was** in Youngstown and the surrounding area more than in other regions,	48
DESCRIPTIVE VERB	"Your actions and failure to act **led to** violations of Senate rules	"Your actions and failure to act **constituted** direct violations of Senate rules	47
LIGHT VERB	**Scoring** higher than 21	**Having** a score higher than 21	14
SEMI-MODAL	The players' initial cards may be **dealt** face up	The players' initial cards **need to** consistently be dealt either face up	5
DOWNGRADE TO PPL	When **examining** the areas history, culture, and economic situation	**Based** on the area's history, culture, and economic situation	2
ORDER	the player or the dealer **wins** by having a score of 21 or by having the highest score	having a score of 21 **means** the player or the dealer wins	1

Table 2: Examples of DYNAMIC –> STATIVE transformations, along with the number of times each transformation strategy was used in the contrast sets.

Strategy	Original Instance (ST)	Transformed Instance (DYN)	#
NEW PARTICIPANT	11 plus the value of any other card will always **be** less than or equal to 21.	Players **add** 11 plus the value of any other card to get less than or equal to 21.	81
INSERT VERB	Since the 1960s, blackjack has **been** a high-profile target of advantage players, particularly card counters,	Since the 1960s, blackjack has **functioned** as a high-profile target of advantage players, particularly card counters,	19
BECOMING	One such bonus **was** a ten-to-one payout	One such bonus **became** a ten-to-one payout	13
UPGRADE PPL	Other casino games inspired by blackjack **include** Spanish 21 and pontoon.	Other casino games were **inspired** by blackjack, including Spanish 21 and pontoon.	5
HEAVY VERB	After receiving their initial two cards, players **have** the option of getting a "hit",	After receiving their initial two cards, players may **pursue** the option of getting a "hit",	2

Table 3: Examples of STATIVE –> DYNAMIC transformations, along with the number of times each transformation strategy was used in the contrast sets.

4.3 Results: Transformation strategies

Finally, we look at contrast set classification accuracy across different transformation strategies. In the STATIVE to DYNAMIC direction, the strategy most often classified correctly (56%) is NEW PARTICIPANT. In the reverse direction, the strategy of replacing the DYNAMIC main verb with a COPULA has the highest accuracy (52%). Notably, even the highly stative nature of the copula doesn't always result in the model recognizing the transformed clause as stative. (Detailed results in Appendix A.)

Both THOUGHT VERB and DESCRIPTIVE VERB for converting DYNAMIC clauses to STATIVE perform poorly. This indicates that verbs of feeling, thinking, and wanting (THOUGHT VERB) and certain verbs like *signify*, *constitute*, and *include*, are not uniformly treated as STATIVE by the classifier. Other descriptive verbs, like *contain* and *resemble* do get accurately classified. Some verbs like *appear* sometimes get classified correctly as STATIVE, as in "All other cards appear as the numeric value", other times as DYNAMIC, as in "Cards appear either from one or two handheld decks, from a dealer's shoe, or from a shuffling machine."

The frequency of transformation strategy in the contrast sets does not correspond to high classification accuracy. In future, we will look at how frequent such constructions are in the training data, and whether their association with stativity labels matches linguistic expectations. We suspect that the different types of stativity may also play a role. For example, the stativity commonly associated with descriptions of mental processes is different from the attributional or predicational stativity often seen with copular constructions.

5 Related work on lexical aspect and verb stativity

Aspectual structure is complex and well-studied in the linguistics literature (Vendler, 1967; Comrie, 1976; Moens and Steedman, 1988; Smith, 1991, among many others). Classically, aspectual analysis involves the semantic properties of stativity, telicity, durativity, and iterativity. Croft et al. (2016) expand on this set of properties in their discussion of aspectual annotations within the Rich Event Description framework. Donatelli et al. (2018) propose methods for expanding the Abstract Meaning

Document	# Clauses Full Text	# Contrast Sets	Test Orig: Correct	Test Trans: Correct	Test Orig: Acc	Test Trans: Acc	Diff by # Clauses	Prop Diff
Wikipedia	114	102	70	60	67.39	58.82	-10	-9.8%
News	149	47	42	16	89.36	34.04	-26	-55.32%
Journal	67	19	11	11	57.89	57.89	-0	-0%
Essay	316	130	114	70	87.68	53.85	-44	-33.85%
Total or Microavg.	646	298	237	157	79.53	52.68		

Table 4: Accuracy on **Contrast: Test_Orig** and **Contrast: Test_Trans**, by document. "Diff by Clause" shows the # of clauses misclassified after transformation, and "Prop Diff" shows the percentage of misclassified contrast sets.

Representation framework with aspectual features, and such aspectual information is a key feature of the Uniform Meaning Representation framework (Van Gysel et al., 2021). In addition, aspectual properties are relevant at both the clause level and the level of individual verbs.

Our current focus is the verb-level property of **stativity**, sometimes referred to as *inherent lexical aspect*.[7] Stativity reflects the degree to which a verb represents a static situation versus a situation that reflects some degree of dynamicity. Dynamic verbs typically involve some change of state.

Computational approaches to stativity. Early approaches to computational analysis of verb stativity employ rule-based approaches based on known linguistic tests for stativity, such as the progressive test.[8] Klavans and Chodorow (1992) produce a type-level stativity rating for English verbs, based on the frequency with which verbs occur in various tenses in the Brown and Reader's Digest corpora. Dorr and Olsen (1997) treat stativity as one of several aspectual properties derivable from logical representations of verb meaning in the Lexical Conceptual Structure (LCS) framework (Jackendoff, 1983, 1990). Siegel and McKeown (Siegel, 1999; Siegel and McKeown, 2000) use a wide range of linguistic indicators to derive type-level stativity values for English verbs. Friedrich and Palmer (2014a) extend Siegel and McKeown's work to incorporate distributional features and perform classification in context. Kober et al. (2020) use distributional semantics to classify both stativity and telicity across genres. Falk and Martin (2016) take a more fine-grained approach to lexical aspect classification for French verbs in context, categorizing verbs across a set of 13 different verbal *readings*. Hermes et al. (2018) take a distributional approach to classifying German verbs for Aktionsart, and (Egg et al., 2019) provide a new annotated corpus and classification experiments for multiple components of aspect for German verbs.

Another important line of research (Govindarajan et al., 2019; Gantt et al., 2022) takes a broader view of event meaning, treating stativity as one of a number of aspectual features which together compose the meaning of an event. Similarly, work on clause-level semantic aspect classification (*aka* situation entity classification) (Friedrich et al., 2016; Becker et al., 2017; Dai and Huang, 2018) considers stativity as a key semantic property for determining clause-level aspect. Finally, Chen et al. (2021) use a sequence of rules to assign tense and aspect values to both verbal events and event nominals, making use of co-occurrence cues of part-of-speech tags, special lexical items, and semantic configurations that help the classifier select the right shade of aspect for a given situation.

6 Conclusions

We apply the contrast set methodology to the task of classifying English verbs in context as stative or dynamic. We see a serious performance degradation on the transformed examples, suggesting the model has not learned a clean decision boundary for stativity. This first analysis suggests a need to more clearly define features that may bias clauses toward stative or dynamic readings.

The study would benefit from more data, across a wider range of text types. We would also like to investigate the effectiveness of contrastive evaluation for other semantic properties, using recently-developed methods for partially-automatic contrast set creation (Li et al., 2020; Bitton et al., 2021; Ross et al., 2021, among others).

[7] *Aktionsart* (Vendler, 1957) also models lexical aspect. Both *Aktionsart* and stativity are subject to coercion at the clause-level, as described in the introduction.

[8] Generally, static verbs in English cannot occur in progressive form: e.g. **I am knowing Thai*. This is one of the most robust of the linguistic tests, but it too is subject to exceptions: e.g. *I am liking this song!*

7 Acknowledgements

This project is supported, in part, by a grant from the Center for Humanities & the Arts at the University of Colorado Boulder.

References

Maria Becker, Michael Staniek, Vivi Nastase, Alexis Palmer, and Anette Frank. 2017. Classifying semantic clause types: Modeling context and genre characteristics with recurrent neural networks and attention. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 230–240, Vancouver, Canada. Association for Computational Linguistics.

Yonatan Bitton, Gabriel Stanovsky, Roy Schwartz, and Michael Elhadad. 2021. Automatic generation of contrast sets from scene graphs: Probing the compositional consistency of GQA. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 94–105, Online. Association for Computational Linguistics.

Daniel Chen, Martha Palmer, and Meagan Vigus. 2021. AutoAspect: Automatic annotation of tense and aspect for uniform meaning representations. In *Proceedings of The Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 36–45, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bernard Comrie. 1976. *Aspect: An introduction to the study of verbal aspect and related problems*, volume 2. Cambridge university press.

Francisco Costa and António Branco. 2012. Aspectual type and temporal relation classification. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 266–275, Avignon, France. Association for Computational Linguistics.

William Croft, Pavlina Pešková, and Michael Regan. 2016. Annotation of causal and aspectual structure of events in RED: a preliminary report. In *Proceedings of the Fourth Workshop on Events*, pages 8–17, San Diego, California. Association for Computational Linguistics.

Zeyu Dai and Ruihong Huang. 2018. Building context-aware clause representations for situation entity type classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3305–3315, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*. ArXiv: 1810.04805.

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

Bonnie J. Dorr and Mari Broman Olsen. 1997. Deriving verbal and compositonal lexical aspect for NLP applications. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 151–158, Madrid, Spain. Association for Computational Linguistics.

Markus Egg, Helena Prepens, and Will Roberts. 2019. Annotation and automatic classification of aspectual categories. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3335–3341, Florence, Italy. Association for Computational Linguistics.

Ingrid Falk and Fabienne Martin. 2016. Automatic identification of aspectual classes across verbal readings. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 12–22, Berlin, Germany. Association for Computational Linguistics.

Annemarie Friedrich. 2017. *States, events, and generics: computational modeling of situation entity types*. Ph.D. thesis, Universität des Saarlandes.

Annemarie Friedrich and Alexis Palmer. 2014a. Automatic prediction of aspectual class of verbs in context. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 517–523, Baltimore, Maryland. Association for Computational Linguistics.

Annemarie Friedrich and Alexis Palmer. 2014b. Situation entity annotation. In *Proceedings of LAW VIII - The 8th Linguistic Annotation Workshop*, pages 149–158, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Annemarie Friedrich, Alexis Palmer, and Manfred Pinkal. 2016. Situation entity types: automatic classification of clause-level aspect. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1757–1768, Berlin, Germany. Association for Computational Linguistics.

William Gantt, Lelia Glass, and Aaron Steven White. 2022. Decomposing and Recomposing Event Structure. *Transactions of the Association for Computational Linguistics*, 10:17–34.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer

Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1161–1166, Hong Kong, China. Association for Computational Linguistics.

Venkata Govindarajan, Benjamin Van Durme, and Aaron Steven White. 2019. Decomposing generalization: Models of generic, habitual, and episodic statements. *Transactions of the Association for Computational Linguistics*, 7:501–517.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Jürgen Hermes, M.P.R. Richter, and Claes Neuefeind. 2018. Supervised classification of aspectual verb classes in german - subcategorization-frame-based vs window-based approach: A comparison. In *ICAART*.

Nancy Ide, Collin Baker, Christiane Fellbaum, Charles Fillmore, and Rebecca Passonneau. 2008. MASC: the manually annotated sub-corpus of American English. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Ray Jackendoff. 1983. *Semantics and Cognition*. The MIT Press.

Ray Jackendoff. 1990. *Semantic Structures*. The MIT Press.

Judith L. Klavans and Martin Chodorow. 1992. Degrees of stativity: The lexical representation of verb aspect. In *COLING 1992 Volume 4: The 14th International Conference on Computational Linguistics*.

Thomas Kober, Malihe Alikhani, Matthew Stone, and Mark Steedman. 2020. Aspectuality across genre: A distributional semantics approach. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4546–4562, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Chuanrong Li, Lin Shengshuo, Zeyu Liu, Xinyi Wu, Xuhui Zhou, and Shane Steinert-Threlkeld. 2020. Linguistically-informed transformations (LIT): A method for automatically generating contrast sets. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 126–135, Online. Association for Computational Linguistics.

Ashutosh Modi and Ivan Titov. 2014. Inducing neural models of script knowledge. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 49–57, Ann Arbor, Michigan. Association for Computational Linguistics.

Marc Moens and Mark Steedman. 1988. Temporal ontology and temporal reference. *Computational Linguistics*, 14(2):15–28.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Alexis Ross, Tongshuang Wu, Hao Peng, Matthew E. Peters, and Matt Gardner. 2021. Tailor: Generating and perturbing text with semantic controls. *CoRR*, abs/2107.07150.

Eric V. Siegel. 1999. Corpus-based linguistic indicators for aspectual classification. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 112–119, College Park, Maryland, USA. Association for Computational Linguistics.

Eric V. Siegel and Kathleen R. McKeown. 2000. Learning methods to combine linguistic indicators:improving aspectual classification and revealing linguistic insights. *Computational Linguistics*, 26(4):595–627.

Carlota Smith. 1991. The parameter of aspect. *Studies in Linguistics and Philosophy*, 43:27–64.

Carlota S. Smith. 2003. *Modes of Discourse: The local structure of texts*. Cambridge University Press.

Sean Trott, Tiago Timponi Torrent, Nancy Chang, and Nathan Schneider. 2020. (Re)construing meaning in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5170–5184, Online. Association for Computational Linguistics.

Jens E. L. Van Gysel, Meagan Vigus, Jayeol Chun, Kenneth Lai, Sarah Moeller, Jiarui Yao, Tim O'Gorman, Andrew Cowell, William Croft, Chu-Ren Huang, Jan Hajič, James H. Martin, Stephan Oepen, Martha Palmer, James Pustejovsky, Rosa Vallejos, and Nianwen Xue. 2021. Designing a Uniform Meaning Representation for Natural Language Processing. *KI - Künstliche Intelligenz*, 35(3):343–360.

Zeno Vendler. 1957. Verbs and times. *The Philosophical Review*, 66(2):143–160.

Zeno Vendler. 1967. Facts and events. *Linguistics in philosophy*, pages 122–146.

A Classification results by transformation strategy

Table 5 shows the distribution of correct labels assigned to transformed clauses using DYNAMIC to STATIVE strategies. Table 6 shows the same for STATIVE to DYNAMIC transformations.

Strategy	**Test Trans: Size**	**Test Trans: Correct**	**Acc**
THOUGHT VERB	55	10	18.18%
COPULA	48	25	52.08%
DESCRIPTIVE VERB	47	17	36.17%
LIGHT VERB	14	4	28.57%
SEMI-MODAL	5	0	0%
DOWNGRADE TO PPL	2	1	50%
ORDER	1	0	0%
Totals	172	57	33.14%

Table 5: Successful DYNAMIC > STATIVE transformation strategies, evaluated by accuracy of correctly identifying the contrast label.

Strategy	**Test Trans: Size**	**Test Trans: Correct**	**Acc**
NEW PARTICIPANT	81	45	55.55%
INSERT VERB	19	9	47.37%
BECOMING	13	9	69.23%
UPGRADE PPL	5	2	40%
HEAVY VERB	2	1	50%
Totals	120	66	55%

Table 6: Successful STATIVE > DYNAMIC transformation strategies, evaluated by accuracy of correctly identifying the contrast label.