The Aligned Multimodal Movie Treebank: An audio, video, dependency-parse treebank

Adam Yaari
MIT CSAIL–CBMM

Jan DeWitt
MIT CSAIL-CBMM

Henry Hu
MIT CSAIL-CBMM

Bennett Stankovits
MIT CSAIL-CBMM

Sue Felshin
MIT CSAIL-CBMM

Yevgeni Berzak
Technion

Helena Aparicio
Cornell University

Boris Katz
MIT CSAIL-CBMM

Ignacio Cases*
MIT CSAIL-CBMM

Andrei Barbu*
MIT CSAIL-CBMM

Abstract

Treebanks have traditionally included only text and were derived from written sources such as newspapers or the web. We introduce the Aligned Multimodal Movie Treebank (AMMT)†, an English language treebank derived from dialog in Hollywood movies which includes transcriptions of the audio-visual streams with word-level alignment, as well as part of speech tags and dependency parses in the Universal Dependencies (UD) formalism. AMMT consists of 31,264 sentences and 218,090 words, that will amount to the 3rd largest UD English treebank and the only multimodal treebank in UD. We find that parsers on this dataset often have difficulty with conversational speech and that they often rely on punctuation which is often not available from speech recognizers. To help with the web-based annotation effort, we also introduce the Efficient Audio Alignment Annotator (EAAA)‡, a companion tool that enables annotators to significantly speed-up their annotation processes.

Keywords: multimodal, video, audio, treebank, Universal Dependency parsing

Correspondence to {yaari,cases,abarbu}@mit.edu
* Equal senior contribution.

†https://github.com/abarbu/ammt
‡https://github.com/abarbu/audio-annotation

1 Introduction

Treebanks are fundamental resources in Natural Language Processing (Nivre et al., 2016). Despite their central role, most existing treebanks are derived from single-modality texts such as newspapers, blogs, and other online communities. The vocabulary, syntax, and statistics of spoken and written language can be quite different from one another (Caines et al., 2017). To complement these datasets and aid the advent of multimodal conversational agents, we have created a new dataset, the Aligned Multimodal Movie Treebank, AMMT, the content of which is derived from language spoken in Hollywood movies. AMMT is released publicly under an open source license and will be contributed to the Universal Dependencies (UD) (Nivre et al., 2020) treebanks.

Speech based treebanks have proven to be a resource of enormous importance to the NLP research community (Ahrenberg, 2007; Nivre et al., 2006). We find Treebank-3 of the Penn Treebank (Marcus et al., 1993), which includes the Penn Treebank Switchboard corpus (Godfrey et al., 1992), to be the closest existing dataset to AMMT. This corpus contains nearly one million transcribed words from Switchboard annotated with part of speech tags, dysfluencies, and parse trees, and it also in-

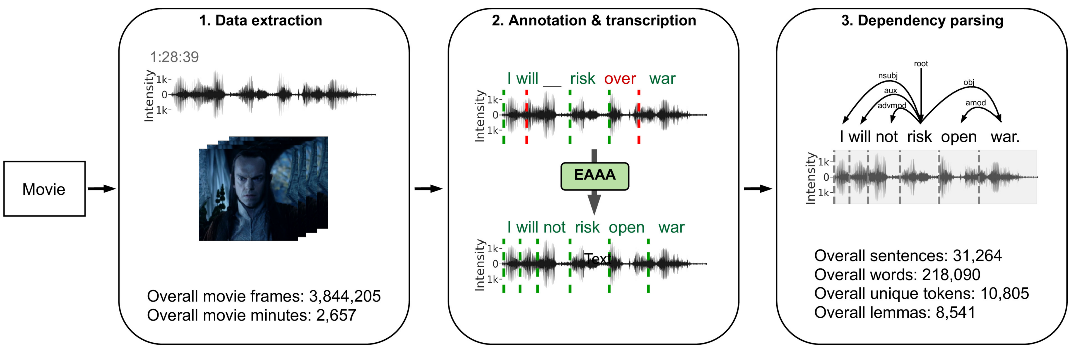

Figure 1: An overview of AMMT, our novel multimodal dataset, consisting of transcriptions and parses for 21 movies aligned at the millisecond level. EAAA is a new transcription and alignment tool introduced below.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 9531 - 9539

cludes alignment between words and audio. However, there are several key differences between this dataset and our own. AMMT provides alignment to visual as well as audio data; it is annotated with UD rather than Penn Treebank dependencies; and conversations are much shorter (Switchboard was designed to have long 10 minute conversations between strangers on the phone discussing one of a preselected list of topics). While conversations in AMMT can still be considered as prepared speech, topics are way less constrained. AMMT also includes many more speakers and its audio quality allowed us to recover almost all spoken words. For practical experiments, AMMT is significantly more entertaining for subjects, a key feature for researchers aiming to study the neuroscience of language via neural imaging. Finally, with this contribution, AMMT is being made open to the whole research community and not restricted to LDC members.

Our contributions are: 1. AMMT is the first large-scale treebank to include alignment to both audio and video. 2. AMMT includes fine-grained millisecond-level word boundaries. 3. AMMT is parsed in the UD framework and is the 3rd largest English UD treebank. 4. A new tool, Efficient Audio Alignment Annotator (EAAA), for rapid word boundary annotation in large corpora.

2 Dataset

The AMMT dataset is an English language treebank based on 21 Hollywood movies that provides transcriptions with word-level alignment to the audio-visual stream, as well as part of speech tags and dependency parses in the UD formalism. Annotations for speaker identification will be included at the time of release. Due to copyrighted source material, AMMT provides multiple 1-second-long audio-visual sample clips from every movie, and a tool chain allowing users to obtain their own copies and verify alignment with the dataset.

AMMT consists of 31,264 sentences, 218,090 words, 8,541 lemmas and 10,805 unique tokens. The counts of POS tags and dependencies are shown in appendix A. The 21 movies from which the dataset is derived are listed in table 4 along with their unique identifiers and relevant statistics.

Movies were chosen to be appropriate for many ages, with the highest rating being PG-13. They belong to a variety of movie genres (including action, adventure, animation, comedy, drama, fantasy, family, and sci-fi, according to IMDb's categorization), and their release dates range from 1995 to present. They were selected to have verbose scripts, in the top 50% of randomly sampled movies. Movies which included extensive singing such as musicals were omitted. Copies of the movies were obtained and extracted in full including opening and closing credits. Special features and after-credits scenes were omitted.

2.1 Transcription pipeline

The audio track was originally transcribed using the Google Cloud Speech-to-Text API (Google, 2020). It was then corrected by annotators, hired from *rev.com* and *happyscribe.com* depending on the movie, and then further extensively corrected by 7 expert annotators. Transcription followed a set of guidelines to deal with problematic audio segments and to enforce coherence. Manual transcription was performed simultaneously with word-boundary annotation using a new tool developed for this purpose, EAAA (see section 4), which was also subsequently used by annotators to perform sentence segmentation and fixing capitalization.

Transcription was verbatim without any corrections for dysfluencies or mistakes. Instructions were provided to the annotators to standardize the transcripts and eliminate problematic audio segments. Foreshortened words (*'round* vs *around*) were transcribed as they were said including the foreshortening. Abbreviations were always expanded (*dr.* vs *doctor*). Cardinal and ordinal numbers were spelled out, while long numbers were written as spoken including conjunctions such as *and* (e.g., *five hundred and five*).

Aligned Multimodal Movie Treebank	
sentences	31,264
tokens	218,090
lemmas	8,541
types	10,805
num. movies	21

Table 1: Basic statistics of the dataset

Manual transcription was carried out simultaneously with word boundary annotation using a purpose-built tool, EAAA (see section 4). EAAA presented annotators with a spectrogram for 4 second segments of a movie, along with the ability to replay and slow down any sub-segment and seek throughout the movie. As the audio was played, a line marked the location of the audio sample in

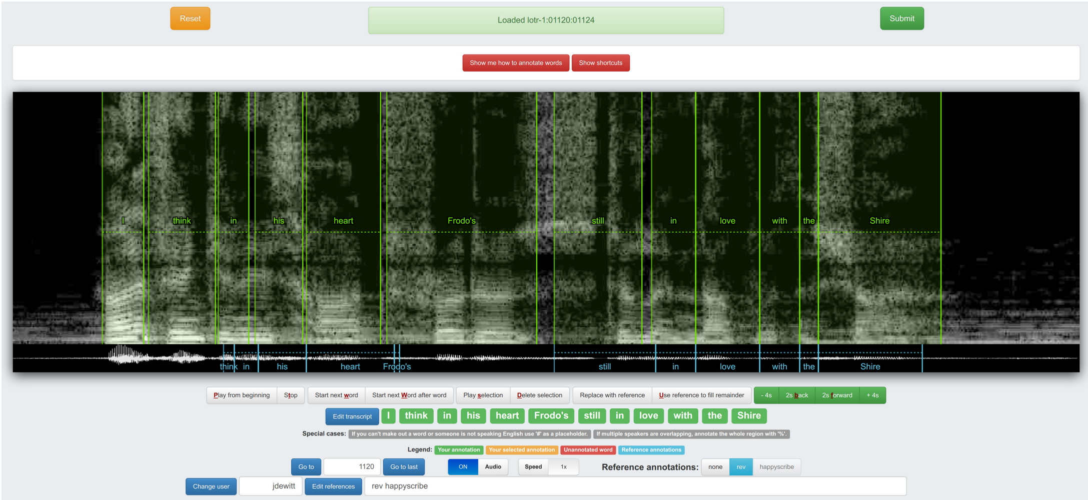

Figure 2: A screenshot of EAAA, the Efficient Audio Alignment Annotator. EAAA allows annotators to browse videos, to play audio segments, play portions of the audio segments, edit the transcript, review multiple reference annotations, and annotate and change word boundaries. EAAA also includes an in-application walkthrough as well as extensive keyboard shortcuts. The main annotation area shows a spectrogram with annotated words. Words can be dragged with a mouse and similarly word boundaries can be adjusted with the mouse. The audio for individual words can be played by clicking them, while any audio segment can be played by clicking and dragging the portion that should be played. At the bottom, in blue, one or more reference annotations are shown which can be toggled on the fly. Annotators can start with a blank slate or initialize annotations from any reference annotation. Audio speed can be controlled as necessary.

the spectrogram in real time. In some cases, annotators could hear specific words but could not clearly identify in the spectrogram where those words occurred (e.g. short words like *to*). Annotators were instructed to annotate what they heard regardless of the spectrogram, sometimes leading to such short words having zero-length intervals. Foreign sentences (e.g., Elvish in the movie *The Lord Of The Rings*) were marked but not included in the corpus, although one-off foreign words in English sentences were transcribed. All cases of singing, unintelligible speech, and multiple speakers overlapping were noted and eliminated from the dataset. Transcripts are as spoken, without correction, even when the speaker erred omitting a word or using a word inappropriately.

After transcription and word boundary alignment, the text was segmented into sentences. Annotators marked the end of each sentence manually and fixed capitalization (of both proper nouns and sentences as needed). Throughout this process, some critical punctuation was introduced as annotators saw fit.

2.2 Dependency parsing pipeline, annotation and validating annotator performance

We parsed all transcriptions with Stanza (Qi et al., 2020) using the standard English model.

Metric	Precision	Recall	F1 Score	AligndAcc
Words	100.00	100.00	100.00	N/A
UPOS	99.53	99.53	99.53	99.53
UAS	98.95	98.95	98.95	98.95
LAS	98.31	98.31	98.31	98.31
CLAS	97.75	97.71	97.73	97.71
MLAS	96.74	96.70	96.72	96.70

Table 2: Inter-annotator agreement bound of AMMT syntactic annotations.

The AMMT dataset was entirely annotated by an in-house expert annotator over the course of a year. Edge cases were discussed with other three team members with strong background in linguistics and Universal Dependencies in particular. In this period of time, the expert annotator performed a total of three sequential passes *over the full dataset* with the idea of promoting internal consistency.

Separately, after this annotation process concluded, a subset of AMMT consisting in 300 sentences of length 5 through 20 uniformly sampled across movies were reannotated by an expert annotator. This expert annotator has a strong background in linguistics and did not contribute to the dataset otherwise. The length of these sentences was selected to avoid the effect of very short or very long sentences (see table 2).

The inter-annotator agreement of the annotations was with 99.53% on correct POS tagging, 98.95% on correctly placing dependencies (UAS), and 98.31% on correctly identifying the type of a dependency relation. MLAS ties together POS and LAS into a single number, 96.72%, which measures the inter-annotator agreement of the annotations (Straka, 2018).

Note that the inter-annotator score presented in table 2 is thus a measure, for this particular subset of the dataset, of the disagreement between the original expert annotator and the external expert annotator. As such it should only be considered as a bound on the actual disagreement between the two annotators.

We found word-boundary inter-annotator agreement to be remarkably high, with less than 15ms on average for all words in a single movie, *Lord Of The Rings*, annotated by 5 annotators.

2.3 Performance of existing parsers

We compared our annotations against those produced by Stanza (Qi et al., 2020) in fig. 3. Stanza was the original parser used to initialize the treebank before extensive human correction. This likely biases the results toward Stanza in subtle ways (Berzak et al., 2016) which we do not investigate here beyond section 2.2.

Note that performance on short sentences, fewer than 3 words, and long sentences, with more than 20 words, is far worse than average-case performance (see fig. 5 for the distribution of sentences in AMMT). This trend is not observed in other corpora such as the English Web Treebank (EWT) (Silveira et al., 2014), where performance increases for short sentences (although these are very infrequent) while the performance drop for long sentences is half or less than that seen in AMMT. While the distributions of POS in both corpora are slightly different (cf. appendix A), the performance drop for short sentences appears to be driven by POS tag errors, see the relative drop in POS accuracy between fig. 3(a,b,c) — perhaps such sentences require more context to be correctly interpreted. The performance drop for long sentences appears to be driven by incorrectly identified relationships, see the relative drop in UAS between fig. 3(a,b,c).

Metric	Precision	Recall	F1 Score	AligndAcc
Words	99.51	99.75	99.63	N/A
UPOS	97.64	97.88	97.76	98.13
UAS	88.02	88.24	88.13	88.46
LAS	85.68	85.89	85.78	86.10
CLAS	83.40	83.01	83.20	83.29
MLAS	81.38	80.99	81.18	81.27

(a) All sentences

Metric	Precision	Recall	F1 Score	AligndAcc
Words	99.45	99.53	99.49	N/A
UPOS	91.49	91.56	91.53	92.00
UAS	91.31	91.38	91.35	91.82
LAS	88.76	88.83	88.80	89.25
CLAS	86.49	86.06	86.28	86.71
MLAS	75.87	75.50	75.68	76.06

(b) Short sentences, fewer than 3 words

Metric	Precision	Recall	F1 Score	AligndAcc
Words	99.52	99.78	99.65	N/A
UPOS	98.44	98.70	98.57	98.92
UAS	80.47	80.68	80.57	80.86
LAS	78.78	79.00	78.89	79.17
CLAS	76.32	76.06	76.19	76.28
MLAS	74.02	73.77	73.90	73.98

(c) Long sentences, more than 20 words

Figure 3: (a) The overall accuracy of Stanza on AMMT. Performance drops significantly for (b) short sentences which are common in speech as well as for (c) long sentences.

3 Multimodal feature analysis

Exploring the utility of the corpus as a multimodal resource for grounded language and vision tasks, we quantified the co-occurrence of nouns and their corresponding objects (i.e. objects that are verbally mentioned as they appear on screen). As an approximation, we considered the 80 object classes of the Microsoft COCO dataset (Lin et al., 2014). We extracted all nouns corresponding to a COCO class (580 nouns across all movies) and manually reviewed the middle frame of a word utterance. We find an average of 36.5% noun-object agreement rate (212 co-occurring objects) across all movies ($\mu = 23.7\%$, $\sigma \approx 17.5\%$ per movie); see fig. 4.

Considering noun-object agreements across both object classes and movie types reveals variable distributions. Some nouns are highly likely to appear on screen as their corresponding noun is uttered, like Person (94.4%), types of vehicles (Car: 59.7%, Bicycle: 68.3%) and animals (Giraffe: 100%, Cow: 100%), while others have not co-occurred once despite being uttered multiple times. Moreover, unambiguous nouns (e.g. Laptop: 50%, TV: 42.8%, Toilet: 33.3%) tend to have a significantly higher agreement rate scores than words with multiple POS (e.g. Bear: 2.5%, Orange: 0%, Remote: 0%). Some movie categories are also more likely to have high noun-object agreement, such as movies aimed for a younger audience (educational and animation genres), perhaps to enable language learning through multimodality. For example *Cars-2* and *Sesame Street* present 79.2% and 74.3% agreement rate respectively, while *The Lord Of The Rings 1* and *2*, and *Avengers Infinity War* score only 17.6%, 14.2% and 5.9% respectively; see fig. 6.

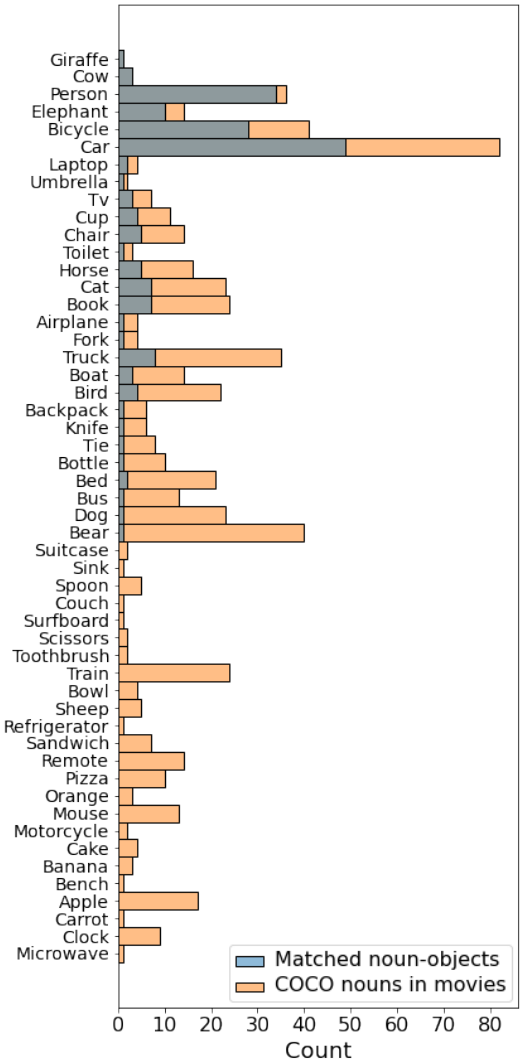

Figure 4: COCO classes noun-object agreement across the corpus (sorted by agreement rate). All nouns corresponding to one of the 80 COCO classes (orange) vs their corresponding objects in the video during the noun utterance (blue). Objects were manually detected in the middle frame of a word utterance.

4 Tools

To efficiently annotate the alignment between word onsets and offsets and the audio stream, we created a new tool, the Efficient Audio Alignment Annotator (EAAA). EAAA enables annotators to start with a rough transcript and approximate alignment between words and the audio track. Annotators can simultaneously correct the transcript while annotating new words. An overview of the EAAA interface is shown in fig. 2. Tools such as Praat (Boersma, 2001) also allow for annotating audio corpora with word boundaries. Unlike Praat, EAAA is web-based making it easier for annotators to use. Data such as spectrograms and wave files seen by annotators is pre-processed on the server-side, making browsing and accessing movies with EAAA near real-time. Since EAAA is a single-purpose tool meant for transcription and fine-grained alignment, it provides custom features which significantly speed up the annotation process like keyboard shortcuts, the ability to handle audio files of any length, and a streamlined interface. EAAA also handles multiple concurrent annotators, sharing and comparing multiple annotations directly.

EAAA pre-processes movie files into 4 second segments that overlap by 2 seconds and computes spectrograms for each segment with Librosa (McFee et al., 2015). Storage is provided by a local Redis database which is not exposed to the web. In addition, EAAA includes a telemetry server which collects comprehensive information during the annotation process including every transcript change, keyboard shortcut used, and mouse press.

5 Conclusion

AMMT and EAAA are open source and AMMT will be contributed to the UD treebanks. In addition to verbatim transcriptions and a treebank, AMMT provides a tool chain to enable access and alignment to the source video and audio. Most datasets for evaluating and training parsers are focused on written rather than spoken language. With the rise of conversational agents, AMMT can serve as a more predictive benchmark in this domain.

At present, no end-to-end systems – from video-and-audio to parses – exist, even if humans often use visual information to disambiguate and contextualize auditory information. We hope that AMMT and its tooling will support further work on multimodal approaches to conversational agents, end-to-end parsing, as well as psychophysics and neuroscience with language in context.

Acknowledgements

This work was supported by the Center for Brains, Minds and Machines, NSF STC award 1231216, the NSF award 2124052, the MIT CSAIL Systems that Learn Initiative, the MIT CSAIL Machine Learning Applications Initiative, the CBMM-Siemens Graduate Fellowship, the DARPA Artificial Social Intelligence for Successful Teams (ASIST) program, the DARPA Knowledge Management at Scale and Speed (KMASS) program, the United States Air Force Research Laboratory and United States Air Force Artificial Intelligence Accelerator under Cooperative Agreement Number FA8750-19-2-1000, the Air Force Office of Scientific Research (AFOSR) under award number FA9550-21-1-0014, and the Office of Naval Research under award number N00014-20-1-2589 and award number N00014-20-1-2643. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

Ethics

The AMMT corpus was constructed using Hollywood movies. Many of these movies generated by the US/Western film industry are unbalanced in terms of cultural and sociolinguistic diversity and oftentimes rely on stereotypes. As such, the distributions of gender, race, age, socioeconomic status, etc. appearing in this corpus are biased as they are sampled from this pool.

Annotators, both in lab and online, contributed significant effort to the development of this dataset. The vast majority of the annotation effort was carried out in lab, with only limited bootstrapping from online services, due to both ethical and quality concerns. Using state-of-the-art models like speech recognizers significantly sped up every stage of the annotation, for example, making transcription only slightly slower than real time. In lab annotators were paid over $18/hour.

Limitations

Movies in AMMT were selected to be appropriate and entertaining for many ages with the highest rating being PG-13. This selection criterion limits the genres and topics covered. Also, speech in movies is prepared speech. While prepared speech is often meant to seem similar to natural speech, it limits the applicability of the corpus. Similarly, the relationships, social situations, and actions taken by agents, are constructions designed with a purpose (e.g. discoursive, entertainment) rather than examples of actual social dynamics, conflict, or growth.

Effort was made to make transcriptions verbatim to maintain the regional or cultural variation in speech present in the original movies, e.g., by directly including foreshortened words. However, such variation is generally selected against in the creation of Hollywood movies and so is poorly represented in this dataset.

The current version of the corpus is monolingual (English) although two Spanish movies were partially processed and may be included in a future version.

References

Lars Ahrenberg. 2007. Lines: An English-Swedish Parallel Treebank. In *Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA 2007)*.

Yevgeni Berzak, Yan Huang, Andrei Barbu, Anna Korhonen, and Boris Katz. 2016. Anchoring and agreement in syntactic annotations. *Conference on Empirical Methods in Natural Language Processing (EMNLP) 2016*.

Paul Boersma. 2001. Praat, a system for doing phonetics by computer. *Glot. Int.*, 5(9):341–345.

Andrew Caines, Michael McCarthy, and Paula Buttery. 2017. Parsing transcripts of speech. *Conference on Empirical Methods in Natural Language Processing (EMNLP) 2017*.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. SWITCHBOARD: Telephone Speech Corpus for Research and Development. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1.

Google. 2020. Speech-to-text: Automatic speech recognition — Google Cloud.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision*.

Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn Treebank.

Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and Music Signal Analysis in Python. In *Proceedings of the 14th Python in Science Conference*, volume 8.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*.

Joakim Nivre, Jens Nilsson, and Johan Hall. 2006. Talbanken05: A Swedish Treebank with Phrase Structure and Dependency Annotation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Natalia Silveira, Timothy Dozat, Marie-Catherine De Marneffe, Samuel R Bowman, Miriam Connor, John Bauer, and Christopher D Manning. 2014. A Gold Standard Dependency Corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*.

A Appendix

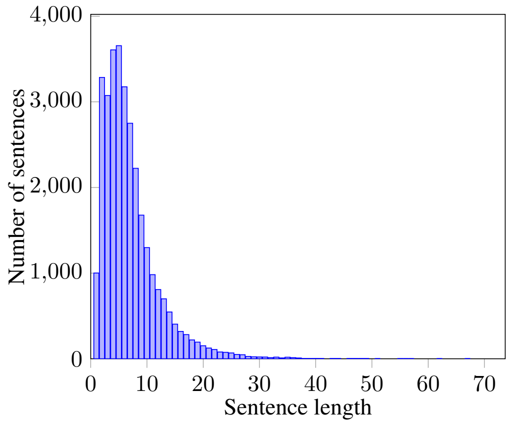

Figure 5: Distribution of sentence lengths in AMMT. Most sentences are quite short. The mean sentence length is 6.97 words long. Compare to standard corpora derived from written sources like the English Web Treebank (15.33 words/sentence) long and the Penn Treebank (23.73 words/sentence in the test set).

POS	Count	Dependencies	Count
ADJ	9829	nsubj	25050
ADP	12464	advmod	14003
ADV	13688	obj	12825
AUX	18965	det	12325
CCONJ	3746	case	11274
DET	12984	aux	9286
INTJ	6275	cop	7830
NOUN	25457	obl	6653
NUM	1835	mark	5693
PART	7202	amod	4958
PRON	36370	xcomp	4306
PROPN	8679	nmod:poss	3996
PUNCT	30301	discourse	3912
SCONJ	2140	cc	3682
SYM	10	compound	3335
VERB	28139	conj	3322
X	6	vocative	3134

Table 3: The distribution of POS tags (left), and the most common dependencies (right). There is a long tail of dependencies.

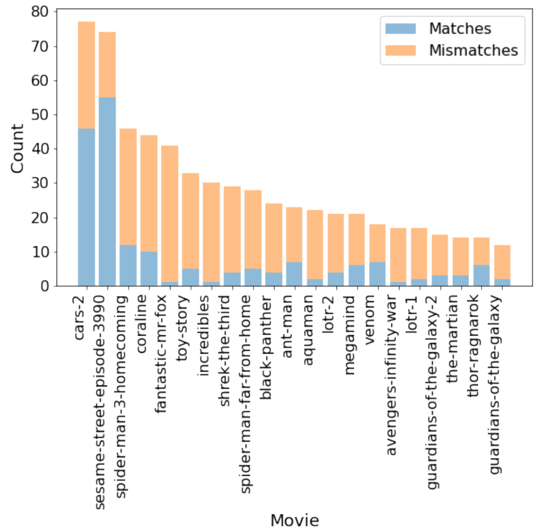

Figure 6: COCO classes noun-object agreements per movie (sorted by number of nouns). All nouns corresponding to one of the 80 COCO classes (orange) vs their corresponding objects in the video during the noun utterance (blue) per movie.

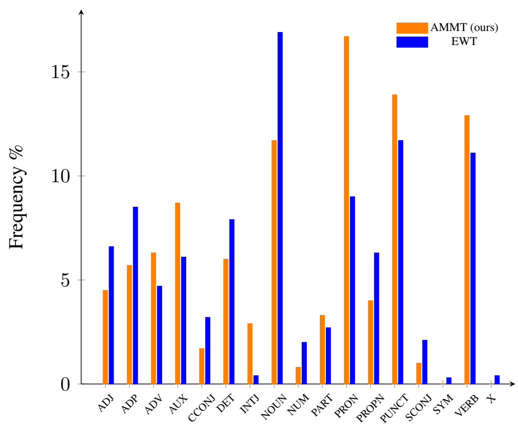

Figure 7: Comparing POS frequency in EWT, a treebank derived from text on the web, and AMMT, our new benchmark derived from spoken language. Among many differences, note that in AMMT, nouns are much less common and pronouns are far more common.

Movie	Year	IMDb ID	Time (s)	Sentences	Tokens	Types	Rating	Frames
Ant-Man	2015	tt0478970	7027	1412	9846	1956	PG-13	168507
Aquaman	2018	tt1477834	8601	1003	7218	1563	PG-13	206251
Avengers: Infinity War	2018	tt4154756	8961	1372	8479	1780	PG-13	214884
Black Panther	2018	tt1825683	8073	1139	7571	1628	PG-13	193590
Cars 2	2011	tt1216475	6377	1801	11404	2060	G	152920
Coraline	2009	tt0327597	6036	933	5428	1251	PG	144743
Fantastic Mr. Fox	2009	tt0432283	5205	1162	8457	1892	PG	124815
Guardians of the Galaxy 1	2014	tt2015381	7251	1104	8241	1799	PG-13	173878
Guardians of the Galaxy 2	2017	tt3896198	8146	1180	9332	1839	PG-13	195341
The Incredibles	2003	tt0317705	6926	1408	9369	1966	PG	166085
Lord of the Rings 1	2001	tt0120737	13699	1424	10538	2011	PG-13	328502
Lord of the Rings 2	2002	tt0167261	14131	1620	11017	2085	PG-13	338861
Megamind	2010	tt1001526	5735	1351	8833	1748	PG	137525
Sesame Street Ep. 3990	2016	tt13725852	3440	718	4218	804	TV-Y	103096
Shrek the Third	2007	tt0413267	5568	999	7192	1586	PG	133520
Spiderman: Far From Home	2019	tt6320628	7764	1705	12004	1988	PG-13	186180
Spiderman: Homecoming	2017	tt2250912	8008	1993	12258	2107	PG-13	192031
The Martian	2015	tt3659388	9081	1421	11360	2210	PG-13	217762
Thor: Ragnarok	2017	tt3501632	7831	1471	9651	1806	PG-13	187787
Toy Story 1	1995	tt0114709	4863	1240	7194	1545	G	116614
Venom	2018	tt1270797	6727	1301	7859	1527	PG-13	161313

Table 4: Name, unique identifier (IMDb ID), and statistics for the 21 movies from which AMMT is derived. Movies were selected to be appropriate for most ages enabling a wide range of experiments. Movies are not randomly sampled; they were selected for their verbose scripts and subjects entertainment during experiments. For more on IMDb identifiers, see https://developer.imdb.com/documentation/key-concepts#imdb-ids