Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 5206–5211
Marseille, 20-25 June 2022
© European Language Resources Association (ELRA), licensed under CC-BY-NC-4.0

Making People Laugh like a Pro: Analysing Humor Through Stand-Up Comedy

Beatrice Turano, Carlo Strapparava
University of Trento, Fondazione Bruno Kessler
Trento, Italy
beatrice.turano@studenti.unitn.it, strappa@fbk.eu

Abstract

The analysis of humor using computational tools has gained popularity in the past few years, and a lot of resources have been built for this purpose. However, most of these resources focus on standalone jokes or on occasional humorous sentences during presentations. In this paper we present a new dataset, SCRIPTS, built using stand-up comedy shows transcripts: the humor that this dataset collects is inserted in a larger narrative, composed of daily events made humorous by the ability of the comedian. This different perspective on the humor problem can allow us to think and study humor in a different way and possibly to open the path to new lines of research.

Keywords: Computational humor, stand up comedy, humor recognition

1. Introduction

Among the communicative activities that humans engage in, humor is one of the most powerful and complex: it holds the power of grabbing attention (Wanzer et al., 2010), establishing rapport with the audience (Stauffer, 1999), and plays a central role in friendship (Gray et al., 2015). Studying humor from a computational perspective is no laughing matter: together with many practical applications that could benefit from such studies, from better recommendation systems to writing-aid tools for writers and comedians (Winters, 2021), computational humor systems could help linguists and researchers collect data to build theories of humor and deepen our understanding of the cognitive processes behind it. Furthermore, it might actively contribute to deepening the bond between humans and machines (Binsted, 1995).

When it comes to Artificial Intelligence, humor is considered an "AI complete problem" (Stock and Strapparava, 2003), because the nature of humor itself is hard to grasp and, consequently, to model computationally. Moreover, factors like gender, culture and religious beliefs are understudied variables that play a role in how we perceive and appreciate humor (see, for an example, Schweizer and Ott (2016)) . Hence, while studying standalone jokes will not lead us far on the path to Natural Language Understanding, a full understanding of humor requires the study of the linguistic and cultural contexts in which it appears, - especially considering the pivotal role that the context of a joke has for the joke to be considered funny.

One way to tackle this issue is studying whole narratives built with the purpose in mind of making the audience laugh, and stand-up comedy offers that kind of narratives. Indeed, comedians performing stand-up usually tell a story that resembles, or is inspired by, real-life experiences, sometimes not comedic at all (as per the famous quote, "Comedy equals tragedy plus time"[1]), and the jokes they tell are interweaved in a routine composed of bits that are inexorably linked to one another (Brodie, 2008) and to the story being told.

In this paper, we are presenting a new dataset, called SCRIPTS, built using this kind of data for the task of humor recognition. Stand-up comedy transcripts were collected and analyzed from not only a linguistic point of view, but also from a cultural point of view: information about the speakers' culture and gender was added, in order to pave the way for studies of humor that take these features into consideration as well.

The paper is structured as follows: section 2 contains a small review of the literature on this topic and other datasets built for this purpose; the dataset object of the study is thoroughly described in section 3. The extracted features, the classification experiments and their results are reported in sections 4 and 5 respectively. Finally, some concluding remarks and a few suggestions for future works are presented in section 6.

2. Related Works

Humor recognition is the task of classifying whether a text expresses a certain degree of humor or not. In the literature, it has mostly been tackled as a binary classification task (Mihalcea and Strapparava (2005); Purandare and Litman (2006); Yang et al. (2015); Bertero and Fung (2016)), and several datasets have been built for this purpose, such as the One-liners dataset (Mihalcea and Strapparava, 2005), containing 16.000 one-liner jokes from daily joke websites, or the Pun of the Day dataset (Yang et al., 2015), constructed more

[1] The origin of this quote is obscure. Several sources would attribute it to Mark Twain; however, someone who uttered it for sure would be U.S. entertainer Steve Allen in *Cosmopolitan*, Feb 1957. The same quote appears in Woody Allen's *Crimes and Misdemeanors* (1989).

or less in the same way, with non-humorous utterances sampled from news websites.

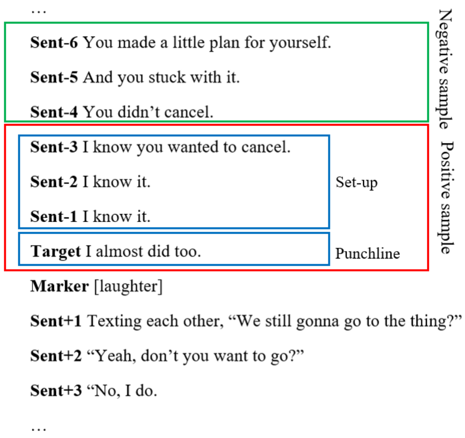
The problem of sampling the negative sentences from different distributions is that the classifiers might have learned to perform the classification based on the different topics present in the sentences, instead of focusing on the humor itself. This issue was addressed by Chen and Soo (2018), who created the negative samples for the Short Jokes and PTT Jokes datasets by sampling sentences from other websites that contained the same words of the positive samples.

The UR-FUNNY dataset (Hasan et al., 2019) focuses instead on multimodal humorous language. It contains not only text, but also the audio and video of 8257 humorous punchlines extracted from TED Talk presentations, from 1741 different speakers. The negative samples are extracted from the same distribution, by taking the same amount of non-humorous sentences from the talks. This work uses the same approach as Chen and Lee (2017), who constructed the Ted Laughter dataset and performed a classification in order to predict audience's laughter during presentations. All the works reviewed so far focused on either standalone jokes or on occasional jokes made during presentations. The new dataset described in this paper, instead, focuses on text written with the intention of building a funny narrative. Something similar has been done by Purandare and Litman (2006) and Bertero and Fung (2016), who analyzed humorous dialogues from sitcoms, but the topics covered by the humorous sentences are restricted to that specific domain.

What the SCRIPTS dataset adds to the literature is the fact that it collects humorous data with audiences' real-time laughter, so punchlines as well as setups are clearly identifiable; moreover, the dataset is enriched with cultural factors (where available) that can help us learn more about humor. To the writers' knowledge, this is the first dataset built using stand-up comedy data, a humoristic genre which is different from the ones collected in other humor datasets.

3. The Dataset

As already mentioned, SCRIPTS contains data from stand-up comedy shows. In particular, the data collected consists in stand-up comedy shows transcripts in English scraped from the same web source[2]. Only the transcripts that contained information on audience's reactions, and in particular of when the audience laughed (the "Laughter" marker), were kept, thus resulting in a total of 90 scripts from 68 different comedians. Together with the scripts, information about each comedian (Gender, Nationality, Ethnicity, Education level, Religious belief, Sexuality, Political inclination, Comedy genres), if available, was collected as well, in order to be used as features or as labels for other classification tasks.



...

Sent-6 You made a little plan for yourself.

Sent-5 And you stuck with it.

Sent-4 You didn't cancel.

Sent-3 I know you wanted to cancel.

Sent-2 I know it.

Sent-1 I know it.

Target I almost did too.

Marker [laughter]

Sent+1 Texting each other, "We still gonna go to the thing?"

Sent+2 "Yeah, don't you want to go?"

Sent+3 "No, I do.

...

Figure 1: An example of how the positive and negative samples were built. The text is from *Tom Papa: Human Mule*

Following research mentioned above (Hasan et al. (2019); Chen and Lee (2017)), the positive samples were extracted by exploiting the "Laughter" marker. Following the *setup-punchline schema*, which is present in many theories of verbal humor (Hetzron, 1991), the sentence immediately preceding the marker was considered the "punchline", while the three sentences preceding the punchline were considered the "setup", unless they were punchlines themselves. The setup and the punchline were then concatenated in order to form the positive sample, as shown in Figure 1. This choice, even though not always accurate, was led by the fact that the text format of the transcript allowed for monosyllabic or not informative sentences preceding a marker, which could have made it very difficult for the classifier to learn enough. For this reason, it was decided to try to replicate the setup-punchline schema by adding the three preceding sentences.

To avoid dramatic topic shifts, the negative samples were extracted from the same distribution, using a very similar process. "No-Laughter" sentences were extracted from the transcript, after making sure that they did not belong to either the setup or the punchline of any positive sample. After that, the sentences that were contiguous in the original text were concatenated in order to form the negative sample, as shown in Figure 1. Building the negative samples in a way similar to that of the positive allowed to control for the length of each sample, that precedent experiments (not reported in this paper) showed to be a relevant feature for the classification.

The final dataset contains 9647 positive and 9490 negative samples.

3.1. Statistics

As already said, 90 scripts from 68 comedians were extracted; of the 68 comedians, 10 are women. Most of the comedians are white Americans and in possession

[2] `https://scrapsfromtheloft.com/`

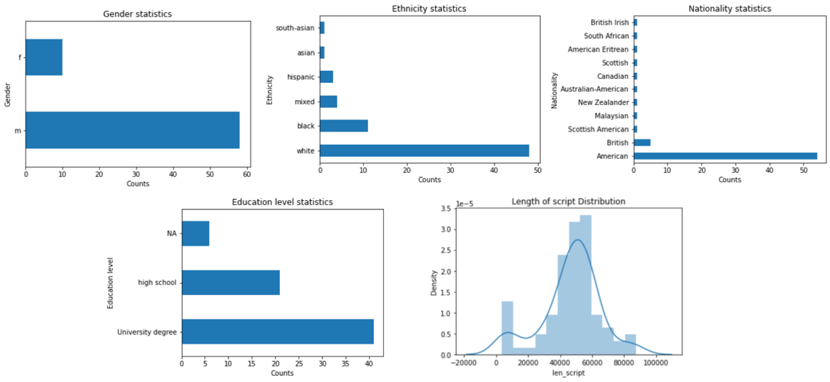


Figure 2: The graphics show some statistical data about the dataset: the gender (female/male, for brevity), ethnicity, nationality and education level of the comedian, and the distribution of the script length.

of a University degree of some kind. The length of each script was added to the dataset as well, in case further analyses want to use it as a selection criterion. This information is reported in Figure 2.

For what concerns the positive and negative samples, they are balanced in number, for a total of 19137.

4. Experiment

This section is divided as follows: firstly, a description of the features used for the classification experiments is provided. The details about the experiments are presented in the subsequent section.

4.1. Features

4.1.1. Social features

As mentioned earlier, the dataset contains some information about the comedian; some of them were included in the experiment, namely Gender, Nationality, Ethnicity and Education level (of the latter, the missing values were inferred using the mode), using one-hot encoding. The other variables had too many missing values to be included in the experiment without undermining the veracity of the data. The four features described will henceforth be referred to as "social features".

4.1.2. Linguistic features

A set of humor-related features was extracted as well, following the previous examples described in section 2. The features are drawn from previous work by Yang et al. (2015), as well as Mihalcea and Strapparava (2005); another small set of features was developed following the advice of Joe Toplyn, writer of several Late Night shows (Toplyn, 2014).

The fourteen features extracted using these sources can be divided into four main groups, and will henceforth be referred to with the umbrella term "linguistic features".

Phonetic features. Features that investigate phonetically relevant characteristics of the sentence, in particular the presence of alliteration, consonance, assonance, and stop consonants; while the first was also included in Yang et al. (2015), the other three were added following chapter 6 of Toplyn (2014). In particular, the presence of alliteration/consonance/assonance chains was investigated. An alliteration chain refers to two or more words beginning with the same phones; a consonance chain refers to two or more words containing the same consonant sounds, while an assonance chain involves vocalic sounds. By stop consonants, it is intended the number of stop consonants within the same sentence, as the use of them is considered a "Joke Maximizer" in Toplyn (2014). For each sample, the following features were extracted using the CMU Pronunciation dictionary[3]:

- the number of alliteration/assonance/consonance chains in the sentence
- the maximum length of the chains
- the number of stop consonants

Ambiguity theory features. Ambiguity plays a role in humor (Bucaria, 2004; Stock, 2003), therefore, again following Yang et al. (2015), a series of features that investigate the use of it were designed. Taking advantage of Wordnet (Fellbaum, 2010), three features were extracted:

- sense combination: as described in Yang et al. (2015), first, all Nouns, Verbs, Adjectives, Adverbs were identified using a POS tagger. Then their possible meanings were identified via Wordnet and the sense combination was computed as $log(\Pi_{i=1}^{k} n_{w_i})$.
- sense farmost: the largest path similarity between any word senses in a sentence.
- sense closest: the smallest path similarity between any word senses in a sentence.

[3] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

Incongruity structure features. A humorous effect often arises from two contrasting pieces of information put next to one another. How to measure contrast? Again deriving from Yang et al. (2015), we used Word2Vec (Mikolov et al., 2013) in order to measure the meaning distance between each word pairs in a sentence. We used Google pretrained 300-dimensional vectors, without training new vectors for this specific task. The two features designed are:

- disconnection: the maximum meaning distance of word pairs in a sentence
- repetition: the minimum meaning distance of word pairs in a sentence

Humor features. This set is composed of two features. The first one, antonymy, derived from Mihalcea and Strapparava (2005), draws from the fact that often a humorous effect can be obtained by using antonymic words or phrases (again derived from incongruity theories of humor), and it is extracted using Wordnet. The other is derived from Toplyn's Joke maximizers (Toplyn, 2014), in particular "Be specific": according to the author, using specific references instead of generic ones increases the humorous effect of a joke. We interpreted it as the use of specific names of people, organizations and other entities that we extracted using a combination of SpaCy (Honnibal and Montani, 2017) and Stanza's (Qi et al., 2020) NER algorithms. The features extracted are:

- antonymy: the number of antonyms present in a sentence.
- specificity: the number of specific entities mentioned in a sentence

4.1.3. Vector representations

Lastly, the Bag of Words and TF-IDF representations of each sample were collected as well: in order to extract these features, the text was tokenized and lowercased and punctuation and special characters removed.

4.2. Models

The goal of the experiments performed was to establish a performance baseline for the dataset. The task is to identify humorous sentences, and it is structured as a binary classification task. The positive samples collected were labeled with a 1 ("humorous"), while the negative ones with a 0 ("non-humorous").

All the features described were normalized before the experiments. 75% of the dataset was used for training, and the remaining 25% for testing.

Two different classes of models were built, one starting from the Bag of Words and the other from TF-IDF representations of the samples. To these base models, the sets of social features (SF) and linguistic features (LF) were concatenated either alone or together, thus resulting in a total of eight different types of feature vectors (see Table 1 for clarity). Three different classification algorithms were used: Logistic Regression (LR), Naive Bayes (NB) and Random Forest (RF). The choice for these algorithms was dictated by the fact that they constituted the baselines for many of the works reviewed in section 2 (see i.e. Chen and Lee (2017), Hasan et al. (2019)). All the models were built using the Scikit-learn (Pedregosa et al., 2011) package for Python.

5. Results and Discussion

Accuracy, precision, recall and F1 scores for every model and every classifier are presented in Table 1. The results show a clear pattern of performance: adding the social and linguistic features to both the Bag of Words and TF-IDF models improved the classifiers' ability to identify humorous sentences. There is no great difference between the Bag of Words and TF-IDF models; however, the latter seem to provide better results.

The Naive Bayes classifier had the worst performance on both types of models, yielding the lowest scores with Bag of Words with social features; interestingly enough, all four Naive Bayes BoW models present a decrease in performance when the social features are added, a pattern that did not repeat for any of the other models and classifiers.

The Logistic Regression classifier presented a different pattern, showing a slight improvement with the addiction of the social features and a greater one when the linguistic features were added too (accuracy=71%).

The Random Forest classifier yielded the best results with both types of models, reaching an accuracy of 72% when all the features were included.

These results well match what can be found in the literature, where baseline performances tend to be established using one of these classifiers as a conventional model (see for example Yang et al. (2015), Chen and Lee (2017) or Hasan et al. (2019)).

Overall, the results of this preliminary study are promising, even though they focus merely on humor recognition. A dataset like this can be used in future classification experiments with other purposes, for example to distinguish between different types of humor, possibly by exploiting the feature "Comedy genres", which was not used in this study, or to teach classifiers to distinguish a humorous versus non-humorous monologue. Moreover, we believe that this data could also be used to fine-tune large language models for not only humor recognition, but perhaps, why not, also for humor generation purposes. As we already stated in other sections of this paper, computational humor is employable in a vast number of applications, and it could play a pivotal role in the growingly tighter bond between humans and machines. Finally, a whole series of classification experiments could be done using the social features included in the dataset, experiments that would be interesting from, for example, a sociological standpoint: for instance, one would want to investigate what are the most prominent features of female versus

	Logistic Regression				Naive Bayes				Random Forest			
	acc.	prec.	recall	F1	acc.	prec.	recall	F1	acc.	prec.	recall	F1
BoW	0.58	0.58	0.58	0.58	0.59	0.60	0.59	0.58	0.71	0.72	0.71	0.71
BoW+SocialFeats	0.61	0.61	0.61	0.61	0.57	0.57	0.57	0.56	0.72	0.73	0.72	0.71
BoW+LingFeats	0.69	0.69	0.69	0.69	0.61	0.61	0.61	0.61	0.71	0.73	0.71	0.70
BoW+Social+LingFeats	0.70	0.70	0.70	0.70	0.58	0.59	0.58	0.57	0.72	0.73	0.72	0.71
TFIDF	0.63	0.63	0.63	0.63	0.61	0.61	0.61	0.61	0.71	0.73	0.71	0.71
TFIDF+SocialFeats	0.65	0.65	0.65	0.65	0.64	0.64	0.64	0.64	0.71	0.73	0.71	0.71
TDIDF+LingFeats	0.70	0.71	0.70	0.70	0.66	0.66	0.66	0.66	0.71	0.73	0.71	0.71
TFIDF+Social+LingFeats	0.71	0.72	0.71	0.71	0.67	0.67	0.67	0.67	0.72	0.73	0.72	0.71

Table 1: Performance of the classifiers on the different models.

male comedy, if they differ in terms of words, topics, or how the humorous effect is built syntactically. The same is valid for different nationalities, ethnicities, religious beliefs, and so forth: this kind of experiments could make a first attempt to understand the extent to which our culture, society, and personal values and beliefs shape what we like joking about - and possibly what we like laughing at.

6. Conclusion

In this paper we presented SCRIPTS, a dataset for the automatic identification of humorous sentences. The dataset contains humorous and non-humorous sentences taken from a corpus of stand-up comedy scripts; the corpus was annotated in such a way to have not only information specific to humor, but also sociological information about the comedian performing the script: indeed, humor is “a quintessential social phenomenon” (Kuipers, 2008) and, therefore, any study of humor should include information about the society where that humor is used, perceived, and appreciated. This information was then added as a feature to the dataset, in order to train a classifier to recognize humorous and non-humorous sentences. The classification was performed using not only conventional Bag of Words and TF-IDF representations in combination with the social features, but also a set of linguistic features, and the results show that mostly the latter have an impact in improving the accuracy, recall and precision of the classifiers, except for Naïve Bayes.

Nevertheless, this study comes with a series of limitations that cannot be ignored. First of all, a lot of the sociological data that was meant to be collected was actually missing, and it will probably take some more time to fill that information, given the sensitive nature of it. Secondly, the transcripts collected all come from the same web source, but only those with the “Laughter” markers were kept: a lot of the scripts available were therefore excluded from analysis. This resulted in an extremely low variety of comedians: indeed, most women’s transcripts did not have the marker and therefore could not be included, and it was the same for transcripts belonging to comedians whose Nationality/Ethinicity was not American/white. A possible future direction for work would be to annotate the available texts or to find the same texts, but with the markers, via another source. Third, neural models were not employed in the classification, and it remains to be seen whether they would provide a stronger performance baseline; and fourth, the final matrix, although normalized, was a sparse matrix and no factorization technique was employed.

A future version of this work, therefore, sees the employment of a neural classifier (like a Convolutional Network, as in Chen and Lee (2017) and Chen and Soo (2018), or a Transformer as in Annamoradnejad and Zoghi (2020)) and the addition of at least another feature, adult slang, following Mihalcea and Strapparava (2005); from Yang et al. (2015), it would be interesting to perform the Humor Anchor Extraction task on this dataset, given that the punchlines are weaved in a larger narrative.

Secondly, adding more of the social features described and using them for classification, in combination with topic models, could potentially be interesting. We believe indeed that information about the people that make us laugh can tell a lot about the people who laugh at their jokes: learning what is humorous for someone according to its gender, religious beliefs, culture, or political tendencies might not be just a curiosity for cocktail parties, but a first step towards a deeper understanding of this phenomenon and a intelligent and audience-adaptive usage of it.

7. Bibliographical References

Annamoradnejad, I. and Zoghi, G. (2020). Colbert: Using bert sentence embedding for humor detection. *arXiv preprint arXiv:2004.12765*.

Bertero, D. and Fung, P. (2016). A long short-term memory framework for predicting humor in dialogues. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 130–135, San Diego, California, June. Association for Computational Linguistics.

Binsted, K. (1995). Using humour to make natural language interfaces more friendly. In *Proceedings of* *the AI, ALife and Entertainment Workshop, Intern. Joint Conf. on AI.*

Brodie, I. (2008). Stand-up comedy as a genre of intimacy. *Ethnologies*, 30:153, 01.

Bucaria, C. (2004). Lexical and syntactic ambiguity as a source of humor: The case of newspaper headlines. *Humor*, 17(3):279–309.

Chen, L. and Lee, C. M. (2017). Predicting audience's laughter using convolutional neural network. *arXiv preprint arXiv:1702.02584*.

Chen, P.-Y. and Soo, V.-W. (2018). Humor recognition using deep learning. 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 113–117.

Gray, A. W., Parkinson, B., and Dunbar, R. I. (2015). Laughter's influence on the intimacy of self-disclosure. *Human Nature*, 26(1):28–43.

Hasan, M. K., Rahman, W., Bagher Zadeh, A., Zhong, J., Tanveer, M. I., Morency, L.-P., and Hoque, M. E. (2019). Ur-funny: A multimodal language dataset for understanding humor. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Hetzron, R. (1991). On the structure of punchlines. *Humor*, 4(1):61–108.

Honnibal, M. and Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Kuipers, G. (2008). *The sociology of humor*. De Gruyter Mouton.

Mihalcea, R. and Strapparava, C. (2005). Making computers laugh: Investigations in automatic humor recognition. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 531–538.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Purandare, A. and Litman, D. (2006). Humor: Prosody analysis and automatic recognition for f* r* i* e* n* d* s. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 208–215.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (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*.

Schweizer, B. and Ott, K.-H. (2016). Faith and laughter: Do atheists and practicing christians have different senses of humor? *Humor*, 29(3):413–438.

Stauffer, D. (1999). Let the good times roll: Building a fun culture. *Harvard Management Update*, 4(10):4–6.

Stock, O. and Strapparava, C. (2003). HAHAcronym: Humorous agents for humorous acronyms. *Humor - International Journal of Humor Research*, 16(3):297–314.

Stock, O. (2003). Password swordfish: Verbal humour in the interface. *Humor - International Journal of Humor Research*, 16(3):281–295.

Toplyn, J. (2014). *Comedy Writing for Late-night Tv: How to Write Monologue Jokes, Desk Pieces, Sketches, Parodies, Audience Pieces, Remotes, and Other Short-form Comedy*. Twenty Lane Media, LLC.

Wanzer, M. B., Frymier, A. B., and Irwin, J. (2010). An explanation of the relationship between instructor humor and student learning: Instructional humor processing theory. *Communication education*, 59(1):1–18.

Winters, T. (2021). Computers learning humor is no joke. *Harvard Data Science Review*, 4. https://hdsr.mitpress.mit.edu/pub/wi9yky5c.

Yang, D., Lavie, A., Dyer, C., and Hovy, E. (2015). Humor recognition and humor anchor extraction. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2367–2376.

8. Language Resource References

Fellbaum, C. (2010). Wordnet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer.