# Estimating word co-occurrence probabilities from pretrained static embeddings using a log-bilinear model

**Anonymous ACL submission**

## Abstract

We investigate how to use pretrained static word embeddings to deliver improved estimates of bilexical co-occurrence probabilities: conditional probabilities of one word given a single other word in a specific relationship. Such probabilities play important roles in psycholinguistics, corpus linguistics, and usage-based cognitive modeling of language more generally. We propose a log-bilinear model taking pretrained vector representations of the two words as input, enabling generalization based on the distributional information contained in both vectors. We show that this model outperforms baselines in estimating probabilities of adjectives given nouns that they attributively modify, and probabilities of nominal direct objects given their head verbs, given limited training data in Arabic, English, Korean, and Spanish.

## 1 Introduction

Word co-occurrence probabilities are a key ingredient in usage-based cognitive models of language. By word co-occurrence probabilities, we mean the probability of a word $w$ given some other single word $c$, $p(w \mid c)$, where words $w$ and $c$ have some specific relationship, for example adjectives that attributively modify nouns or nouns serving as direct objects of verbs (Gries and Durrant, 2020). These co-occurrence probabilities feed into information-theoretic measures of 'thematic fit' and selectional restriction (Resnik, 1996; Lapata et al., 1999; Padó et al., 2007; Vecchi et al., 2017), are relevant in predicting human online processing difficulty (eg. McRae et al., 1998; Trueswell et al., 1994), and play a key role in language acquisition (Harris, 1954; Erickson and Thiessen, 2015). Most prominently, the widely-used pointwise mutual information (PMI) measure of association strength, $\mathrm{PMI}\,(w, c) = \log \frac{p(w|c)}{p(w)}$ (Church and Hanks, 1990), relies on these conditional probabilities as an input. PMI appears in models of grammar induction (Yuret, 1998; Clark and Fijalkow, 2020), sentence comprehension and production (Futrell et al., 2020; Ranjan et al., 2022), and quantitative theories of word order variation (Sharma et al., 2020).

Word co-occurrence probabilities are hard to estimate accurately from text data because empirical counts of a particular pair of words in a particular relation are often sparse. This limitation makes it hard to evaluate cognitive theories that operate on co-occurrence probabilities. Although high-performance pretrained language models now exist (Radford et al., 2019; Devlin et al., 2019), the probabilities of interest often cannot be read off of these models directly. This is because $w$ and $c$ might be defined by relations that cannot be straightforwardly detected in terms of linear word order or templates. For example, suppose we are interested in the distribution of adjectives attributively modifying a noun in English. We cannot ask a language model for the distribution of words immediately preceding a noun, because some of these words will not be attributive adjectives.

We propose to improve the estimation of word co-occurrence probabilities by leveraging pretrained static word embeddings to enhance generalization from potentially small training sets. Our method enables generalization based on the semantic and syntactic information contained in word embeddings for both words $w$ and $c$.

## 2 Model

**Setting** We are given a **vocabulary** of words $V$, a finite **target word set** $W \subseteq V$, a dataset of $N$ pairs of words $\{\langle w_i, c_i \rangle\}_{i=1}^{N}$ where the **target word** $w$ is an element of target word set $W$ and the **context word** $c$ is an element of the full vocabulary $V$, and a pretrained mapping from words to $D$-dimensional static embeddings $E : V \to \mathbb{R}^D$. Supposing the dataset consists of iid samples from

some distribution $p(w, c) = p(c)p(w \mid c)$, our goal is to find a conditional distribution $q(w \mid c)$ with support $W$ to approximate $p(w \mid c)$ in a way that leverages the static embeddings $E$.

**Proposed model** We propose a log-bilinear model (Mnih and Hinton, 2007, 2008) using word embeddings as input:[1]

$$q(w \mid c) = \frac{1}{Z(c)} \exp\Big\{\phi(\mathbf{w})^\top \mathbf{A} \psi(\mathbf{c})\Big\} \quad (1)$$

$$Z(c) = \sum_{w \in W} \exp\Big\{\phi(\mathbf{w})^\top \mathbf{A} \psi(\mathbf{c})\Big\}, \quad (2)$$

where $\mathbf{w} = E(w)$ and $\mathbf{c} = E(c)$ are the static embeddings of target word $w$ and context word $c$ respectively, the **target word encoder** $\phi(\cdot) : \mathbb{R}^D \to \mathbb{R}^K$ and **context word encoder** $\psi(\cdot) : \mathbb{R}^D \to \mathbb{R}^L$ are functions which may be parameterized as feed-forward neural networks with parameters denoted $\phi$ and $\psi$ respectively, and $\mathbf{A}$ is a $K \times L$ interaction matrix. The model parameters $\phi$, $\psi$, and $\mathbf{A}$ are trained to minimize the cross-entropy loss

$$J(\phi, \psi, \mathbf{A}) = -\sum_{n=1}^{N} \log q(w_n \mid c_n). \quad (3)$$

**Modeling decisions** A modeler applying this approach needs to make a number of decisions, including the choice of static word embeddings and the structure of the word encoders $\phi(\cdot)$ and $\psi(\cdot)$. It is also possible to set $\psi = \phi$, using the same function to encode both the target word and the context word; this setup can reduce the number of parameters at the cost of less flexibility in fitting the training data.

Another major modeling decision involves the target word vocabulary $W$, which determines the support of $q(w \mid c)$ and is summed over during the calculation of the partition function (Eq. 2). In some cases, the modeler may not have access to a finite set $W$ of possible target words. As long as the full vocabulary $V$ is finite, it is possible to set $W = V$ and learn a probability distribution with support on all words in $V$.

Setting $W = V$ has the advantage that it allows the modeler not to commit to any particular target word set, thus avoiding the risk of prematurely excluding legitimate target words. It has the disadvantages that (1) calculation of the partition function (Eq. 2) is slower and/or more memory intensive, and (2) the learning problem is more difficult because probability mass is initially spread over the set $V$ as opposed to a potentially much smaller set $W$.

**Implementation** In all experiments reported below, training is performed in PyTorch using the Adam algorithm (Kingma and Ba, 2015). To handle out-of-vocabulary items, we include an unknown-word symbol `UNK` in the target word set $W$ and full vocabulary $V$. If a target word $w$ in a dataset is not present in the target word set $W$, or a context word $c$ is not present in the full vocabulary $V$, then that word is mapped to `UNK`. In the embedding map, we assign `UNK` to a normalized random vector drawn from a Gaussian distribution.

## 3 Related work

Distributional similarity information has been used to improve modeling of word co-occurrence probabilities in previous work. Dagan et al. (1994, 1999) defined a kernel-based interpolated language model where probability mass is explicitly spread over similar words, with variant models along these lines found in Wang et al. (2005) and Yarlett (2008). These models leverage similarity information about target words but not context words. In contrast, Bíró et al. (2007) proposed a method which uses similarity information about the context word but not the target word. Toutanova et al. (2004) developed a method that can exploit similarity information about both target and context, using a Markov Chain algorithm incorporating distributional and WordNet similarities. None of this previous work derived word similarity information from pretrained embeddings, because such embeddings did not exist at the time.

The log-bilinear model for conditional word probabilities was introduced in a language modeling context by Mnih and Hinton (2007, 2008). Mikolov et al. (2013) influentially proposed to use the vector representations output by the word encoder as general word embeddings. The current work aims to return log-bilinear models to their language modeling roots, evaluating the capabilities of these models to estimate co-occurrence probabilities using pretrained embeddings as input, with a focus on word distributions where training data is limited.

Recently Nikkarinen et al. (2021) introduced a neural–Bayesian nonparametric estimator for prob-

[1] We have suppressed bias terms from the notation.

ability distributions on single words. Their setting has an unknown and generally infinite vocabulary $V$, and their model generalizes using a character-level LSTM. In contrast, our model assumes a pre-existing known vocabulary $V$ with embeddings, and generalizes based on those embeddings. A hybrid model may be possible in future work.

A related literature in corpus linguistics and NLP has explored the nature of restricted binary word co-occurrences, called collocations (for recent examples, see Savary et al., 2017; Kutuzov et al., 2017; Garcia et al., 2021; Espinosa Anke et al., 2021). Our work focuses narrowly on the estimation of bilexical conditional probabilities, which are often inputs to models for collocation detection.

## 4 Experiments

We study the ability of the embedding-based log-bilinear model to estimate conditional distributions for (1) adjectives attributively modifying nouns and (2) nominal direct objects modifying verbs, in Arabic, English, Korean, and Spanish. We compare the embedding-based log-bilinear model against baselines:

- Additive smoothing with $\alpha = 1$:

$$p_{\text{add}}(w \mid c; \alpha) \propto \text{count}(c, w) + \alpha,$$

  where $\text{count}(c, w)$ is the frequency of the pair of words $c$ and $w$ in the training data.

- An interpolated smoothed estimator:

$$p_{\text{interp}}(w \mid c) = p_{\text{add}}(w \mid c; \alpha) + \lambda p_{\text{MLE}}(w),$$

  where $p_{\text{MLE}}$ is a maximum likelihood estimate, $\lambda = \frac{1}{4}$, and $\alpha = 1$.

- A softmax distribution on target words as a function of the context word embedding $\mathbf{c}$ (as proposed by Bíró et al., 2007):

$$p_{\text{softmax}}(w \mid c) \propto \exp\left\{\theta_w^\top \psi(\mathbf{c})\right\},$$

  where $\theta_w$ is an optimized weight vector for the target word $w$. This baseline uses the context word embedding $\mathbf{c}$ but not the target word embedding $\mathbf{w}$. It is equivalent to having the target word encoder return a one-hot vector representation of target word $w$.

- Models without word encoders, achieved by setting $\phi(\cdot)$ and $\psi(\cdot)$ to identity functions. Such models decode target words from the word embeddings directly.

All baselines are subject to the same vocabulary restrictions and out-of-vocabulary policy as the full log-bilinear models. As our standard test metric, we report the average negative log likelihood (NLL) of held-out data. We report NLLs for the full test set, as well as the challenging subset of the test set consisting of word pairs where the context word was never seen during training.

Below, we describe the experimental setting for our two tasks, and then we describe the results.

### 4.1 Distribution of attributive adjectives given nouns

We examine the distribution of attributive adjectives given the nouns that they modify, for example adjectives like *red* modifying nouns like *ball* in phrases like *the red ball*.

**Data** We use Universal Dependencies (UD) 2.8[2] (Nivre et al., 2020) and the automatically-parsed Wikipedia datasets released as part of the CoNLL 2017 Shared Task (Zeman et al., 2017) as a source of attributive adjective–noun pairs. We extract all pairs of words linked by a dependency of type *amod* where the head has universal part-of-speech (UPOS) `NOUN` and the dependent has UPOS `ADJ`. We represent the pair using the downcased wordforms of the adjective and noun.

For each language, we use the fastText aligned word vectors (Bojanowski et al., 2017; Joulin et al., 2018),[3] limiting the vocabulary set $V$ to the top 200,000 vectors by frequency. For the target word vocabulary $W$, we take the 10,000 most frequent wordforms among all attributive adjectives extracted from the entire CoNLL Wikipedia dataset.

As training sets, we use 99,000 adjective–noun pairs drawn randomly from the Wikipedia datasets for each language, so training set size is fixed across languages. We use an additional 1,000 pairs from the Wikipedia datasets as development sets for hyperparameter tuning and early stopping, and for test sets we extract all pairs from the relevant UD corpora.[4] Pairs where the target word $w$ is not in the target word vocabulary $W$ are removed from the development and test sets.

[2] http://hdl.handle.net/11234/1-3687

[3] https://fasttext.cc/docs/en/aligned-vectors.html

[4] For English, we concatenate EWT and GUM. For Arabic, we concatenate NYUAD and PADT. For Spanish, we concatenate AnCora and GSD. For Korean, we concatenate Kaist and GSD.

| | Attributive adjectives given nouns | | | | | | Direct objects given verbs | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Softmax | | Log-Bilinear | | | | Softmax | | Log-Bilinear | |
| Data | Add. | Interp. | No Enc. | Enc. | No Enc. | Enc. | Add. | Interp. | No Enc. | Enc. | No Enc. | Enc. |
| Arabic | 8.50 | 7.05 | 8.31 | 8.04 | **5.79** | 5.89 | 9.78 | 9.78 | 9.17 | 9.00 | 8.63 | **8.47** |
| *Unseen c* | 9.15 | 9.60 | 8.31 | 8.52 | **6.93** | 6.98 | 9.71 | 9.84 | 9.03 | 8.86 | 9.09 | **8.76** |
| English | 8.75 | 7.17 | 7.15 | 7.16 | **6.40** | 6.41 | 9.64 | 8.99 | 8.64 | 8.58 | 8.16 | **8.04** |
| *Unseen c* | 9.01 | 8.40 | 7.21 | 7.22 | 6.99 | **6.96** | 9.89 | 9.96 | 8.62 | 8.56 | 8.39 | **8.35** |
| Spanish | 8.70 | 7.49 | 8.13 | 8.10 | **6.27** | **6.27** | 9.70 | 9.10 | 8.64 | 8.52 | 7.96 | **7.84** |
| *Unseen c* | 9.17 | 9.50 | 8.15 | 8.21 | 7.16 | **7.09** | 9.80 | 9.62 | 8.48 | 8.48 | 8.35 | **8.18** |
| Korean | 7.96 | 5.39 | 5.51 | 5.61 | **4.81** | 4.82 | 9.71 | 9.76 | 9.20 | 9.18 | 8.34 | **7.99** |
| *Unseen c* | 7.16 | 5.92 | 5.45 | 5.48 | 5.44 | **5.40** | 9.67 | 9.91 | 9.16 | 9.14 | 9.58 | **8.76** |

Table 1: Average NLLs of adjectives given nouns and direct objects given verbs in UD corpora for models and baselines. 'Add.' is the additive smoothing baseline. 'Enc.' and 'No Enc.' refer to models with and without word encoders, respectively. *Unseen c* indicates performance on pairs where the context (the head noun for adjectives given nouns, and the head verb for direct objects given verbs) was never observed at train time.

**Training and hyperparameters** Each model is trained for the number of iterations that gives minimum loss on the Wikipedia dev set. The word encoders are feed-forward neural networks with one hidden layer of 300 units and an output layer of 300 units, with ReLU activation. In training, we use batch size 32; we also experimented with batch size 512 but this resulted in rapid overfitting.

### 4.2 Distribution of nominal direct objects given verbs

We examine the distribution of nominal direct objects given verbs; for example, from a sentence such as *I kicked the red ball*, we would be interested in the probability of the direct object *ball* given its head noun *kicked*. All procedures here are the same as for the distribution of attributive adjectives given nouns except where noted.

**Data** We extracted direct objects as all pairs of words linked in a dependency of type *obj* where the head has UPOS `VERB` and the dependent has UPOS `NOUN`. Because nouns are more open-class than adjectives, we used a target word vocabulary of size 20,000.

### 4.3 Results

Results are shown in Table 1. The log-bilinear models outperform all others. In several cases (see for example Spanish and Korean adjectives), only the log-bilinear model is capable of outperforming the interpolated baseline.

When predicting adjectives from nouns, the log-bilinear models without word encoders sometimes outperform those with word encoders. These is perhaps not surprising: the input word embeddings were trained to be used in a log-bilinear skip-gram probability model, so they already form useful representations for word prediction.

Overall performance on predicting objects from verbs is worse than when predicting adjectives from nouns. This reflects the harder nature of the task and the larger support size required to model nouns rather than adjectives.

### 4.4 Additional experiments

We also trained full log-bilinear models with a number of other settings. We found that tying the word and context encoders does not substantially change performance, but that fine-tuning the input word embeddings leads to severe overfitting. Removing the target word vocabulary restriction (setting $W = V$) also substantially negatively impacts performance: for adjectives, the best test set NLL is 6.57 for Arabic, 6.75 for English, 6.95 for Spanish, and 4.89 for Korean.

## 5 Conclusion

We evaluated log-bilinear modeling as means to leverage pretrained word embeddings for the estimation of co-occurrence probabilities in different syntactic configurations. We found that this method delivers accurate probability estimates across languages, outperforming baselines. This method will be useful in all applications requiring such probabilities. Upon publication we will release code to train and run these models.

## References


István Bíró, Zoltán Szamonek, and Csaba Szepesvári. 2007. Sequence prediction exploiting similarity information. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, pages 1576–1581.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Alexander Clark and Nathanaël Fijalkow. 2020. Consistent Unsupervised Estimators for Anchored PCFGs. *Transactions of the Association for Computational Linguistics*, 8:409–422.

Ido Dagan, Lillian Lee, and Fernando CN Pereira. 1999. Similarity-based models of word cooccurrence probabilities. *Machine Learning*, 34(1):43–69.

Ido Dagan, Fernando CN Pereira, and Lillian Lee. 1994. Similarity-based estimation of word cooccurrence probabilities. In *32nd Annual Meeting of the Association for Computational Linguistics*, Las Cruces, New Mexico, USA. 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. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Lucy C Erickson and Erik D Thiessen. 2015. Statistical learning of language: Theory, validity, and predictions of a statistical learning account of language acquisition. *Developmental Review*, 37:66–108.

Luis Espinosa Anke, Joan Codina-Filba, and Leo Wanner. 2021. Evaluating language models for the retrieval and categorization of lexical collocations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1406–1417, Online. Association for Computational Linguistics.

Richard Futrell, Edward Gibson, and Roger P Levy. 2020. Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science*, 44(3):e12814.

Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021. Assessing the representations of idiomaticity in vector models with a noun compound dataset labeled at type and token levels. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2730–2741, Online. Association for Computational Linguistics.

Stefan Th Gries and Philip Durrant. 2020. Analyzing co-occurrence data. In *A Practical Handbook of Corpus Linguistics*, pages 141–159. Springer.

Zellig S. Harris. 1954. Distributional structure. *Word*, 10(23):146–162.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2984, Brussels, Belgium. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Andrey Kutuzov, Elizaveta Kuzmenko, and Lidia Pivovarova. 2017. Clustering of Russian adjective-noun constructions using word embeddings. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, pages 3–13, Valencia, Spain. Association for Computational Linguistics.

Maria Lapata, Scott McDonald, and Frank Keller. 1999. Determinants of adjective-noun plausibility. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 30–36, Bergen, Norway. Association for Computational Linguistics.

Ken McRae, Michael J Spivey-Knowlton, and Michael K Tanenhaus. 1998. Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, 38(3):283–312.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Andriy Mnih and Geoffrey E Hinton. 2007. Three new graphical models for statistical language modelling. In *ICML '07: Proceedings of the 24th International Conference on Machine Learning*, pages 641–648.

Andriy Mnih and Geoffrey E Hinton. 2008. A scalable hierarchical distributed language model. *Advances in Neural Information Processing Systems*, 21:1081–1088.

Irene Nikkarinen, Tiago Pimentel, Damián E Blasi, and Ryan Cotterell. 2021. Modeling the unigram distribution. *arXiv preprint arXiv:2106.02289*.

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*, pages 4034–4043, Marseille, France. European Language Resources Association.

Sebastian Padó, Ulrike Padó, and Katrin Erk. 2007. Flexible, corpus-based modelling of human plausibility judgements. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 400–409, Prague, Czech Republic. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Sidharth Ranjan, Rajakrishnan Rajkumar, and Sumeet Agarwal. 2022. Locality and expectation effects in hindi preverbal constituent ordering. *Cognition*, 223:104959.

Philip Resnik. 1996. Selectional constraints: An information-theoretic model and its computational realization. *Cognition*, 61(1-2):127–159.

Agata Savary, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, and Antoine Doucet. 2017. The PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 31–47, Valencia, Spain. Association for Computational Linguistics.

Kartik Sharma, Richard Futrell, and Samar Husain. 2020. What determines the order of verbal dependents in Hindi? Effects of efficiency in comprehension and production. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 1–10, Online. Association for Computational Linguistics.

Kristina Toutanova, Christopher D Manning, and Andrew Y Ng. 2004. Learning random walk models for inducing word dependency distributions. In *Proceedings of the Twenty-First International Conference on Machine Learning*, page 103.

John C Trueswell, Michael K Tanenhaus, and Susan M Garnsey. 1994. Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language*, 33(3):285–318.

Eva M Vecchi, Marco Marelli, Roberto Zamparelli, and Marco Baroni. 2017. Spicy adjectives and nominal donkeys: Capturing semantic deviance using compositionality in distributional spaces. *Cognitive Science*, 41(1):102–136.

Qin Iris Wang, Dale Schuurmans, and Dekang Lin. 2005. Strictly lexical dependency parsing. In *Proceedings of the Ninth International Workshop on Parsing Technology*, pages 152–159, Vancouver, British Columbia. Association for Computational Linguistics.

Daniel G Yarlett. 2008. *Similarity-based generalization in language*. Ph.D. thesis, Stanford University.

Deniz Yuret. 1998. *Discovery of linguistic relations using lexical attraction*. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria de Paiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonça, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2017. CoNLL 2017 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. Association for Computational Linguistics.