

TEXT COLLECTIONS FOR EVALUATION OF RUSSIAN MORPHOLOGICAL TAGGERS

OLGA LYASHEVSKAYA^{4,5,6} – VICTOR BOCHAROV³ – ALEXEY SOROKIN^{1,2} –
TATIANA SHAVRINA^{4,7} – DMITRY GRANOVSKY³ – SVETLANA ALEXEEVA³

¹Lomonosov Moscow State University, Russia

²Moscow Institute of Physics and Technology, Russia

³OpenCorpora.org

⁴Higher School of Economics, National Research University, Moscow, Russia

⁵Vinogradov Institute of the Russian Language RAS, Moscow, Russia

⁶Russian National Corpus, Moscow, Russia

⁷General Internet-Corpus of Russian, Moscow, Russia

LYASHEVSKAYA, Olga – BOCHAROV, Victor – SOROKIN, Alexey – SHAVRINA, Tatiana – GRANOVSKY, Dmitry – ALEXEEVA, Svetlana: Text collections for evaluation of Russian morphological taggers. *Journal of Linguistics*, 2017, Vol. 68, No 2, pp. 258 – 267.

Abstract: The paper describes the preparation and development of the text collections within the framework of MorphoRuEval-2017 shared task, an evaluation campaign designed to stimulate development of the automatic morphological processing technologies for Russian. The main challenge for the organizers was to standardize all available Russian corpora with the manually verified high-quality tagging to a single format (Universal Dependencies CONLL-U). The sources of the data were the disambiguated subcorpus of the Russian National Corpus, SynTagRus, OpenCorpora.org data and GICR corpus with the resolved homonymy, all exhibiting different tagsets, rules for lemmatization, pipeline architecture, technical solutions and error systematicity. The collections includes both normative texts (the news and modern literature) and more informal discourse (social media and spoken data), the texts are available under CC BY-NC-SA 3.0 license.

Keywords: text collection, shared task, morphological tagging, universal dependencies, morphological parsing, Russian corpora

1 MOTIVATION

Comparison of existing methods for automatic text processing on every level is one of the pledges of systematic development of NLP technologies for each language. MorphoRuEval-2017 [1] is an initiative in the framework of Dialogue-Evaluation, aimed at both assessing and improving the evaluation metrics of morphological tagging and lemmatization for the Russian language, as applied to different text registers (news, social media, literary texts). As part of this shared task, the organizers faced the challenge of compiling a large training collection using different sources with annotation of good quality. It was decided to unify all the main corpus collections for Russian, coming from all the principal corpus projects – RNC [2], GICR [3], OpenCorpora.org [4], and SynTagRus [5] – all sources with different tagsets, obtained by different algorithms, and using different dictionaries. Our assumptions were that the morphological data standard for training collection should be 1) concise, 2) compatible with international shared task results, 3) suitable for

rapid and consistent annotation by a human annotator, 4) suitable for computer parsing with high accuracy, 5) easily comprehended and used by a non-linguist (the last three are taken from “Manning’s Laws” [6]). As an essential solution of the problem we have chosen a new standard of multilingual morphological tagging, Universal Dependencies¹ (UD) [7].

2 SOURCE DATA

During the shared task, the following annotated data was provided:

- 1) RNC Open: a manually disambiguated subcorpus of the Russian National Corpus – 1.35 million words, ca. 10 thousand sentences (a balanced sample of fiction, news, nonfiction, spoken data, and blogs). RNC project is regarded as the main source for research in literary language.
- 2) GICR corpus with the resolved homonymy – 1 million words. General Internet-Corpus of Russian provides rich amount of blogs and social media texts, and is used as an instrument for modern and non-normative language studies.
- 3) OpenCorpora.org data – 400 thousand tokens (news, wikipedia, nonfiction, blogs). OpenCorpora provides mainly blogs and news texts, mostly normative and modern.
- 4) SynTagRus – 900 thousand tokens (fiction, news). SynTagRus is a part of RNC, openly distributed for syntactic research.

In each corpus, information about word form, lemma, part of speech (POS), and grammatical features were provided. To unify the representation of the data, the conll-u format was chosen, as the most common, convenient, and simple, and for the unification of morphological tags, the format of the Universal Dependencies (further UD) 2.0 was used (with some specifications, see below). The text collections are now available under CC BY-NC-SA 3.0 license².

We have also provided for the comparison the following plain text collections: 30 million words from LiveJournal, 30 million words from Facebook, Twitter and VKontakte, and 300 million words from Librusec.

3 MORPHOLOGICAL STANDARD

3.1 Background

Historically, the first morphological standards of the publicly available Russian corpora were, generally taken, based on Zalizniak’s grammatical dictionary [7] and its spin-offs, and adopted the output of a few programs for Russian morphological analysis (Dialing/AOT, Mystem, ETAP, Starling). As a prominent example, the POS list of the RNC standard [2] included 13 classes of Zalizniak and three more specific subcategories for adverbs and predicates, while the inventory of grammatical features incorporated the so called “secondary forms” such as locative II (e.g. (v) *les-u* ‘in

¹ <http://universaldependencies.org/>

² All materials accessible at <https://github.com/dialogue-evaluation/morpho-RuEval-2017>

the forest’, as opposed to the locative (*o*) *les-e* ‘about the forest’) and comparative II (e.g. *po-skoreje* ‘faster’). In the SynTagRus treebank, a number of additional distinctions were motivated by the needs of machine translation (e.g. grammatical gender of the personal pronoun *ja* ‘I’).

Later on, a successful attempt was made to compile a tagset compatible with the international multilingual specifications developed with emphasis on the statistical processing, Multext-East [9]. The manually disambiguated portion of the RNC was converted into this format and used for training, and a number of models for TreeTagger, TnT, SVMTagger were provided (see <http://corpus.leeds.ac.uk/mocky/>). The variants of the Multext-East are currently exploited in the Russian Internet Corpus, HANKO, ruTenTen, Araneum, and GICR corpora.

Yet another multilingual standard was adopted for the Russian morphology in UD-Russian and UD-Russian-SynTagRus annotation schema [10]. It is mostly compatible with the RNC standard and annotation practice, but the feature set is reduced by dropping distinctions between the “primary” and “secondary” forms, whereas the POS list is expanded to the new categories of proper nouns, auxiliaries, subordinate conjunctions, symbols, and punctuation marks to agree with unified Universal Dependencies standard [13].

Unlike the above-mentioned standards, the OpenCorpora tagset was developed specifically to be convenient for manual disambiguation of grammatical forms taken into account that annotation is made by crowdsourcing. Since some distinctions made in reference grammars and dictionaries were considered difficult to be explained to the crowd and to be applied to real data by the crowd, a number of adjustments were made. For example, the comparative forms of adjectives and adverbs were collapsed in a single POS category. Participles, gerunds, infinitives, and finite verb forms were treated as four separate parts of speech since this reflected a classification used in some secondary school programs of the Russian language.

As a result, the morphological annotation of existing Russian corpora differs in the following respects:

- (a) If the annotation is token-based (simplex forms), or the periphrastic forms are tagged as well (cf. analytical future tense forms such as *budem schitat* ‘(we) will assume’);
- (b) If the multiword units are tokenized as one token or several tokens, particular multiword expressions are treated as single units, if any;
- (c) The number and borders of the POS categories;
- (d) The structure of the inflectional categories and their values;
- (e) The structure of the lemma-classifying categories and values (e.g. transitivity, personal names, etc.);
- (f) Presence/absence of additional tags which signals the disambiguation status; not-in-dictionary-ness; violation of grammatical norms, etc.;
- (g) Lexical attachment: for example, the animacy tag may be obligatorily assigned to the pronoun *kto* ‘who’ in some corpora and be omitted in others;
- (h) Lemmatization rules are affected by the structure of POS-tags and grammatical tags, on the one hand, and by some internal agreements within the standard, on

the other hand. For example, the superlative forms can get (i) the lemma of the base adjective and the superlative degree tag (cf. SynTagRus); (ii) the lemma with the superlative affix and no degree tag (cf. RNC, GICR); or (iii) the lemma with the superlative affix and the superlative degree tag (cf. OpenCorpora). In RNC, the perfective and imperfective verbs are assigned two different lemmas, whereas in SynTagRus, the perfective verb will usually get the lemma of the imperfective aspectual counterpart.

3.2 Unified Representation

There is no clear benchmark for morphological tagging for Russian. Apart from Universal Dependencies for Russian, those advantages were already mentioned in Section 1, there are already several competing standards, such as AOT tagset, NLC tagset, Dialog-2010 tagset, positional tagset for Russian, etc³. In this way, with one's desire to evaluate morphological tagging quality, one should inevitably face the problem of unification. With respect to the work of our colleagues at the MorphoEval-2010 [11], we carefully summarized all the inconsistencies and tag matches of our data set (described in Section 4). Within our standard, we unified the mismatches, concerning closed-class mismatches (predicatives, particles, determiners, conjunctions and adpositions), yet some of the cases of open-class lexemes left as is (see Section 5).

4 CONVERSION AND EVALUATION OF THE DATA

4.1 The Tagsets of Four Corpora

RNC Open

RNC Open is a subcorpus of the manually disambiguated corpus made available for the offline processing under a non-commercial license. The texts of social media (blogs) were prepared specifically for the MorphoRuEval. The “deficient” tagsets [12] (those lacking some non-determined categories such as gender in pluralia tantum nouns) were normalized. Besides, in the cases where more than one possible grammatical parsing was present in the annotation, we left only one, usually the most frequent and pragmatically neutral. All grammatical categories which were not included in the MorphoRuEval list (such as transitivity, voice, indeclinability, anomalous and distort forms, etc.) are provided with their values in a separate field (in the UD notation).

OpenCorpora

OpenCorpora project works on crowdsourcing morphological annotation. All Russian language native speakers are encouraged to participate and volunteers' knowledge or ability to do this work isn't assessed before they start. Works on annotation aren't paid directly, nor indirectly. Participants are motivated by the fact that they create a freely available resource. About 5 thousand people have participated so far.

In order to maintain annotation quality three- or fourfold overlap is provided and all disagreements are verified manually by moderators with linguistic education. The part of OpenCorpora dataset which was used in MorphoRuEval-2017 shared task consists of randomly selected sentences only partially verified. Decision in all not moderated cases is taken by majority voting.

³<https://github.com/kmike/russian-tagsets>

SynTagRus

SynTagRus was one of the first Russian treebanks automatically converted into UD standard [10]. For the MorphoRuEval shared task, the data were reannotated in UD v.1.4 standard and then the morphological tags were converted into the unified standard. Unlike UD-SynTagRus, all the limitations and solutions of the shared task are applied to the data.

GICR

GICR corpus with resolved homonymy is a first GICR open-source subcorpus, tagged with the aid of Abbyy Compreno technologies. Natural Language Compiler tagset was converted to MSD-Russian with the following specifications: GICR now contains a special category for parenthesis, predicatives and digits, that led to extension of the common MSD format. These specifications were reduced to UD standard according to the instructions, with an exception of parenthesis – they were left in the training data with the tag “H”. The procedure of conversion is not straightforward since, for example, GICR contain several classes of pronouns, which must be arranged to different classes in SynTagRus. For example, adjective pronouns (*ego* ‘his’, *kotoryj* ‘which’) become determiners and adverbial pronouns (*kak-to* ‘somehow’ *vsegda* ‘always’) become adverbs, Several GICR adjectives *drugoj* ‘other’, *kaghdyj* ‘every’ were also considered as determiners.

4.2 POS

Table 1 demonstrates mapping of POS-tags in four corpora and MorphoRuEval unified list. For the reference, the column for the UD 2.1 POS-tags is also provided.

Part of speech	RNC	GICR	Open Corpora	SynTagRus (UD 1.4)	UD 2.1	Morpho-RuEval
(common) noun	S	N	NOUN	NOUN	NOUN	NOUN
proper noun	S	N	NOUN	--	PROPN	PROPN
initial letter	INIT	=	NOUN + Init	=	=	=
pronoun	SPRO	P	NPRO	PRON	PRON	PRON
numeral	NUM	M	NUMR	NUM	NUM	NUM
adjective	A	A	ADJF	ADJ	ADJ	ADJ
adjective (short form)	=	=	ADJS	=	=	=
adjectival numeral	ANUM	=	ADJF / ADJS + Anum	ADJ	ADJ	=
adjectival pronoun / determiner	APRO	P	ADJF / ADJS + Apro	DET	DET	DET
participle, full form	V	A	PRTF	VERB	VERB	ADJ
participle, short form	=	A	PRTS	=	=	=
verb	V	V	VERB	VERB	VERB	VERB
Infinitive verb	=	=	INFN	=	=	=
gerund	=	=	GRND	=	=	=
auxiliary	=	=	--	=	AUX	=
adverb	ADV	R	ADVB	ADV	ADV	ADV
adverbial pronoun	ADVPRO	P	ADVB + Ques / Dmns	=	=	=

parenthetically used discourse markers	PARENTH	H	ADVB	ADV	ADV	H
preposition / postposition	PR	S	PREP	ADP	ADP	ADP
conjunction	CONJ	C	CONJ	CONJ	CCONJ	CONJ
subordinate conjunction	=	=	CONJ	=	SCONJ	=
particle	PART	Q	PRCL	PART	PART	PART
interjection	INTJ	I	INTJ	INTJ	INTJ	INTJ
symbol	SYM	X	SYMB	SYM	SYM	X
foreign words, non-words	NONLEX	X	LATN	X	X	X
punctuation mark	--	-	PNCT	PUNCT	PUNCT	PUNCT
comparative	--	A, R	COMP	--	--	ADJ, ADV
predicative, predicative pronoun	PRAEDIC, PRAEDIC PRO	W	PRED	--	--	ADJ, ADV, VERB

Tab. 1. POS-tags

In RNC, the nouns were divided into NOUNs and PROPNS using the grammatical features of personal names, patronymics, toponyms, etc.; inanimate nouns were checked manually. The participles, which were tagged as VERB in the original standard, were assigned the tag ADJ, but their lemma remains the form of the infinitive, and an additional tag. The adjectival numerals were converted to ADJ except *odin* ‘one’, which semantically belongs to the class of cardinal numerals (marked as NUM following the UD standards). The classes of SPRO and APRO roughly correspond to PRON and DET, respectively. We compiled word lists to define these categories, and all words outside the lists were treated as nouns and adjectives. Conversion of predicatives is shown below.

In GICR, there is a special category for parenthetical constructions (H), which cannot be simply mapped onto adverbs or predicatives, as they are often a complex token combination. H is left in the training data, but not considered in evaluation. Proper nouns were also mapped onto simple NOUN during conversion to UD, that also led to testing procedure constraints discussed in Section 5.

OpenCorpora uses its own morphological tagset developed to be convenient for manual annotation purposes. In order to convert this tagset to Universal Dependencies an “OpenCorpora to UD” module has been added to Russian-tagsets project⁴.

There is a number of deviations from MorphoRuEval-2017 guidelines in morphological annotation of OpenCorpora subset:

- the concept of auxiliary verb doesn’t exist in OpenCorpora on morphological level and VERB / AUX disambiguation isn’t performed. The verb *byt’*, ‘be’ is always annotated with VERB tag;
- OpenCorpora treats comparative as a separate part of speech. Universal dependencies guideline considers comparative as a form of an adjective or an adverb. In UD version of OpenCorpora subset all comparatives are annotated with ADJ tag.

⁴<https://github.com/kmike/russian-tagsets>

SynTagRus shows the closest match with regard to POS tags, except proper names, participles, and symbols. The proper names are tagged as NOUNs, the participle forms were converted to ADJ, and SYM was converted to X.

5 REMAINING DISCREPANCIES

Concerning the fact that the irreducible standard difference can affect the training results of the track participants, we refused to use the part-of-speech SYM (symbol) and AUX (auxiliary verb), and coordinate and subordinate conjunctions are both marked as CONJ. Here are the left ones in our collection: noun (NOUN), proper name (PROPN), adjective (ADJ), pronoun (PRON) numeral (NUM), verb (VERB), adverb (ADV), determinant (DET), conjunction (CONJ), preposition (ADP), particle (PART), interjection (INTJ). Also on the data are marked punctuation marks (PUNCT) and non-word tokens (X).

The following categories are marked and unified for different parts of speech:

1. Noun: gender, number, case, animacy
2. Proper name: gender, number, case
3. Adjective: gender, number, case, brevity of form, degree of comparison
4. Pronoun: gender, number, case, person
5. Numeral: gender, case, graphic form
6. Verb: inclination, person, tense, number, gender
7. Adverb: degree of comparison
8. Determinant: gender, number, case
9. Conjunction, preposition, particle, parenthesis, interjection, other: none

Accepted values:

Case: nominative – Nom, genitive – Gen, dative – Dat, accusative – Acc, locative – Loc, instrumental – Ins

Gender: masculine – Masc, feminine – Fem, neuter – Neut

Number: singular – Sing, plural – Plur

Animacy: animated – Anim, inanimated – Inan

Tense: past – Past, present or future – Notpast

Person: first – 1, second – 2, third – 3

VerbForm: infinitive – Inf, finite – Fin, gerund – Conv

Mood: indicative – Ind, imperative – Imp

Variant: short form – Brev (if the form is complete, no mark is placed)

Degree: positive or superlative – Pos, comparable – Cmp

NumForm: numeric token – Digit (if the token is written in alphabetic form, no mark is placed).

In order to increase the annotation agreement in the collections converted from different sources and simplify semiautomatic verification of annotation correctness, the following decisions were made:

- 1) DET is a closed class which includes 44 pronouns used primarily in the attributive position, exceeding official list of 30 determiners – such cases as *vsyak* ‘any’ (*vernacular*), *ihniy* ‘their’ (*vernacular*) were also included.

- 2) Predicative words. Modal words such as *mozžno* ‘can’, *nelzja* ‘cannot’ are considered as adverbs. The word *net* ‘no, not’ is considered as a third-person form of a verb. The predicative words homonymous to the short neuter forms of adjectives are coded as adjectives. Unlike adverbs, the short adjectives always form a part of the predicate.

That condition was checked automatically by extracting the subject and predicate from each sentence and verified manually afterwards. Except for several words, our algorithm discriminates between adverbs and short adjectives in the same way as the one use in UD-SynTagRus does.

- 3) The lemma of the verb is its infinitive form in a particular aspect (perfective or imperfective). The gerund forms constitute a part of the verb paradigm. Verbs in passive voice keep their passive suffix *-sja* in their infinitive form as well.
- 4) The participles are treated as adjectives and their lemma is the full nominative masculine singular form. This form is reconstructed using dictionary lookup and suffix transformations.
- 5) The ordinal numerals are considered adjectives.
- 6) The tense forms of the verb are divided into Past and Notpast (present or future).
- 7) The analytic (multi-word) forms of verbs, adjectives, and adverbs are not coded. For example, the analytic future tense form is annotated as two separate tokens: the future form of the verb *byt* ‘be’ and the infinitive.
- 8) For all prepositions including phonetic variants *c/co*, *6/bo* its lemma coincides with the word itself.
- 9) NOUN and PROPN were evaluated as a single tag.
- 10) CONJ and SCONJ\CCONJ were also regarded to one tag.
- 11) Differences between UD 1.4 and UD 2.0 were not penalized.

Several of categories received the status of “not rated”: they may be present or not in the output of the system under evaluation:

- * animacy (nouns, pronouns);
- * aspect, voice, and transitivity (verbs);
- * POS tags of prepositions, conjunctions, particles, interjections, and X (others).

6 CONCLUSION

The dataset collected shows one of the most challenging issue in the Russian NLP domain: there exist a lot of competing standards, associated with different existing pipelines and different theoretical views on Russian morphology. From the point of view of technological development and increasing interest among developers to the field of NLP, the mentioned data sources will inevitably be unified to one format. One can only hope that this format will be widely used and won’t become just one of N+1 competing standards, as in comparison with the previous shared tasks, this unification is more detailed. The main merits of the work described are:

- the original data set which was annotated in a single format consistent with UD guidelines was prepared and presented;
- techniques and principles which correspond to the UD standard, at the same time considering current situation with disparate standards for the Russian;
- the comprehensive guidelines for testing procedure and evaluation in this format.

All materials of MorphoRuEval-2017 including training and test set are now available at the competition's github. We welcome NLP-researchers and specialists in machine learning to use this collection and we hope that the collection will stay practical and relevant for a long time.

ACKNOWLEDGEMENTS

Authors are grateful to all the colleagues who were also involved in the MorphoRuEval organization committee and provided data: Kira Droganova, Alena Fenogenova, Ilia Karpov. We also thank GICR and OpenCorpora.org teams for preparing their segments. Dmitry Sichinava and Svetlana Savchuk has done a lot to make the RNC Open collection available. We also express our gratitude to the participants of the MorphoRuEval tracks who took part in the discussion list and provided their suggestions and comments.

References

- [1] Sorokin, A., Shavrina, T., Lyashevskaya, O., Bocharov, V., Alexeeva, S., Droganova, K., and Fenogenova, A. (forthcoming). MorphoRuEval-2017: an evaluation track for the automatic morphological analysis methods for Russian. In *Computational linguistics and intellectual technologies. Proceedings of International Workshop Dialogue'2017*, Moscow.
- [2] Lyashevskaya, O. N., Plungian, V. A., and Sichinava, D. V. (2005). O morfologicheskom standarte Korpusa sovremennogo russkogo jazyka [Morphological standard of the Corpus of contemporary Russian]. In *Nacional'nyj korpus russkogo jazyka: 2003–2005* [Russian National Corpus: 2003–2005], pages 111–135, Moscow. Accessible at: <http://ruscorpora.ru/sbornik2005/08lashevs.pdf>.
- [3] Selegey, D., Shavrina, T., Selegey, V., and Sharoff, S. (2016). Automatic morphological tagging of Russian social media corpora: training and testing. In *Computational linguistics and intellectual technologies. Proceedings of International Workshop Dialogue'2016*, Moscow.
- [4] Bocharov, V. V., Alexeeva, S. V., Granovsky, D. V., Protopopova, E. V., Stepanova, M. E., and Surikov, A. V. (2013). Crowdsourcing morphological annotation. In *Computational linguistics and intellectual technologies. Proceedings of International Workshop Dialogue'2013*, Vol. 12 (19), Moscow.
- [5] Boguslavsky, I. (2014). SynTagRus—a Deeply Annotated Corpus of Russian. In Blumenthal, P., Novakova, I., and Siepmann, D., editors, *Les émotions dans le discours-Emotions in Discourse*, pages 367–380, Peter Lang, Frankfurt am Main, Germany.
- [6] Nivre, J. (2016). *Reflections on Universal Dependencies*. Department of Linguistics and Philology, Uppsala University.
- [7] Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, Ch. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of LREC 2016*, pages 1659–1666, Portorož, Slovenia.
- [8] Zalizniak, A. A. (1977/2003). *Grammaticheskij slovar' russkogo jazyka* [A Grammatical Dictionary of Russian.] Moscow.
- [9] Sharoff, S., Kopotev, M., Erjavec, T., Feldman, A., and Divjak, D. (2008). Designing and evaluating Russian tagsets. In *Proceedings of LREC 2008*, Marrakech, Morocco.
- [10] Lyashevskaya, O., Droganova, K., Zeman, D., Alexeeva, M., Gavrilova, T., Mustafina, N., and Shakurova, E. (2016). Universal Dependencies for Russian: a New Syntactic Dependencies Tagset. In *Series: Linguistics, WP BRP 44/LNG/2016*.
- [11] Toldova, S., Sokolova, E., Astafiyeva, I., Gareyshina, A., Koroleva, A., Privoznov, D., Sidorova, E., Tupikina, L., and Lyashevskaya, O. (2012). Ocenka metodov avtomaticheskogo analiza teksta

- 2011-2012: Sintaksicheskie parsery russkogo jazyka [NLP evaluation 2011-2012: Russian syntactic parsers.] In *Computational linguistics and intellectual technologies. Proceedings of International Workshop Dialogue 2012*. Vol. 11 (18), pages 797–809, RGGU, Moscow.
- [12] Lyashevskaya, O. (2016). The grammatical tagset of Russian. In Lyashevskaya, O. *Korpusnye instrumenty v leksiko-grammaticheskikh issledovaniyakh russkogo jazyka* [Corpus approach to Russian grammar and lexicon], pages 435–456, Languages of Slavic culture press, Moscow.
- [13] McDonald, R., Nivre, J., Quirbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K., Petrov, S., Zhang, H., Täckström, O., Bedini, C., Bertomeu Castelló, N., and Lee, J. (2013). Universal Dependency Annotation for Multilingual Parsing. In *Proceedings of ACL*. Accessible at: <https://ryanmcd.github.io/papers/treebanksACL2013.pdf>.