Identifying and Handling Cross-Treebank Inconsistencies in UD: A Pilot Study

Tillmann Dönicke
Göttingen Centre for
Digital Humanities
University of Göttingen[*]
`tillmann.doenicke@`
`uni-goettingen.de`

Xiang Yu
Institute for Natural
Language Processing
University of Stuttgart
`xiangyu@ims.`
`uni-stuttgart.de`

Jonas Kuhn
Institute for Natural
Language Processing
University of Stuttgart
`jonas.kuhn@ims.`
`uni-stuttgart.de`

Abstract

The Universal Dependencies treebanks[1] are a still-growing collection of treebanks for a wide range of languages, all annotated with a common inventory of dependency relations. Yet, the usages of the relations can be categorically different even for treebanks of the same language. We present a pilot study on identifying such inconsistencies in a language-independent way and conduct an experiment which illustrates that a proper handling of inconsistencies can improve parsing performance by several percentage points.

1 Introduction

Merging several available treebanks for the same language is a simple way of obtaining more data – either for training language-specific tools or for systematic corpus studies on the language. This strategy presupposes that each of the merged treebanks adheres to the same annotation scheme; when treebanks to be merged originate from distinct research contexts (which is typically the case in practice), a harmonization process is required. The UD initiative (whose main motivation has been to obtain comparable syntactic annotations *across* languages) has the welcome side effect that whenever several corpora for the same language are included in the UD collection, the amount of consistently annotated treebank data for this language increases – at least in theory. In practice, there can be serious inconsistencies in the way the dependency relations are used in the respective treebanks. This can for instance be the effect of distinct strategies in conversion from a non-UD treebank format, or of idiosyncrasies in the linguistic treatment of a certain construction pursued in one treebanking team. Often such inconsistencies may go unnoticed – in particular when researchers do experiments on large samples of languages from UD – and may lead to problematic conclusions.

Some previous efforts have focused on discovering errors or inconsistencies *within* a treebank, by measuring the patterns based on partially lexicalized dependency relations (Boyd et al., 2008; de Marneffe et al., 2017) or calculating the fitness scores of the dependency relations in a tree (Alzetta et al., 2017). In contrast to these studies, we focus on finding inconsistencies *between* treebanks that are supposed to be annotated in the same way according to a common guideline. Our method builds on the assumption that each individual treebank is already largely consistent within itself, and seeks to identify dependency patterns with contradicting statistics across treebanks.

So far, annotation inconsistencies have been reported for some specific languages, e.g. Korean (Noh et al., 2018) and Russian (Droganova et al., 2018), but there seems to be no comprehensive study addressing these cross-treebank inconsistencies within the UD collection. In our pilot study, we experiment with a dependency-based measure for inconsistency detection that makes no specific language-typological assumptions and appears to be quite effective for detecting various types of inconsistencies. In two experiments, we demonstrate that inconsistencies detected this way can be used as the trigger for a comparatively simple, but effective conversion process on one of the treebanks: a parser trained on the converted treebank achieves higher accuracy on the treebank that is closer to the UD guidelines.

---

[*]Work performed while at University of Stuttgart.

[1]`https://universaldependencies.org/`

2 Methodology

We became aware of inconsistent usages of dependency relations when extracting relative branching directions for another task (Dönicke et al., 2020; Dönicke, 2020) and we reuse this measure here to find potential cross-treebank inconsistencies, without making any prior assumptions what could be a problematic construction in a language under consideration. The measure takes advantage of the fact that the frequency distribution of the directionality of dependencies is by and large independent of domain-specific factors (which will of course differ across treebanks). Moreover, many of the problematic inconsistencies found across treebank schemes go along with the choice of head vs. dependency status and will thus be reflected in the directionality of the arcs for a particular label – even when the label is used fairly infrequently.

The relative frequency of right-branching instances for a dependency relation R in a treebank t is

$$ v_R^t = \frac{\#\left[\,\overset{R}{\curvearrowright}\,\right](t)}{\#\left[\,\overset{R}{\curvearrowleft}\,\right](t) + \#\left[\,\overset{R}{\curvearrowright}\,\right](t)}\,. $$

Here, an expression of the form $\#[*](t)$ returns how often the construction * appears in the treebank t. The value ranges between 0 and 1; for example, $v_{\text{amod}}^t = 0$ indicates that adjectival modifiers never follow (i.e. always precede) their governor in t, 1 indicates that they always follow their governor in t, and 0.5 indicates that both orderings are equally common. This measure is very similar to the directional dependency distance of Chen and Gerdes (2017). Our measure only quantifies the average direction of a relation whereas their measure quantifies average direction and distance.

We found that the relative branching direction of a relation sometimes greatly differs among the treebanks of the same language. Therefore, we calculate the maximum branching direction difference (MBD) for a language ℓ as

$$ MBD(R, \ell) = \max_{t_1 \in T(\ell)} v_R^{t_1} - \min_{t_2 \in T(\ell)} v_R^{t_2} = \max_{t_1, t_2 \in T(\ell)} v_R^{t_1} - v_R^{t_2} $$

where $T(\ell)$ is the set of all treebanks for ℓ. A higher value indicates that there are at least two treebanks for ℓ in which R shows a different branching behaviour.

3 Inconsistencies among treebanks

We use the MBD to search for inconsistent usages of dependency relations in UD, version 2.5 (Zeman et al., 2019). Table 1 shows languages and relations which have high differences in branching direction between treebanks.[2] We reduced all dependency relations to their base forms (e.g. "csubj:pass" to "csubj") and excluded dependency relations that are very infrequent and tend to have very dissimilar occurrences.[3] We then manually inspected the first 20 cases to evaluate whether they are a real inconsistency or merely a false positive. For the analysis, we mainly used the statistics for individual relations on the treebanks pages (e.g. `https://universaldependencies.org/treebanks/en_partut/en_partut-dep-compound.html`) in combination with the Grew-match tool[4], maintained by Inria, to inspect treebank-specific examples.

We show three examples in the following, two real inconsistencies and one case in which a higher MBD is caused by both consistent and inconsistent use of a relation. The rest of our analysis is provided in the appendix. We are not familiar with all languages of which we analysed treebanks and it is possible that we missed something, making our analysis possibly incomplete. This is also the reason for why we could not analyse the potential inconsistencies for some of the languages (marked by a question mark in the table).

---

[2]Our implementations are available at `https://github.com/tidoe/typology-coling`.

[3]"dep", "discourse", "dislocated", "fixed", "goeswith", "list", "orphan", "parataxis", "reparandum" and "vocative".

[4]`http://match.grew.fr/`

ℓ	R	$MBD(R,\ell)$	v_R^t for all treebanks of ℓ	Incon.
Chinese	clf	1.00	GSD GSDSimp PUD CFL HK 0.00, 0.00, 0.02, 0.99, 1.00	T
Korean	aux	1.00	PUD GSD Kaist 0.00, 0.83, 1.00	T
Korean	mark	1.00	Kaist GSD 0.00, 1.00	?
Arabic	compound	0.98	PUD NYUAD 0.02, 1.00	T
Galician	cc	0.94	TreeGal CTG 0.00, 0.94	T
Turkish	compound	0.94	PUD GB IMST 0.01, 0.05, 0.95	?
English	csubj	0.90	GUM LinES EWT ESL ParTUT PUD Pronouns 0.10, 0.56, 0.62, 0.73, 0.73, 0.73, 1.00	F
Old Russian	expl	0.88	RNC TOROT 0.06, 0.93	F
Spanish	compound	0.88	PUD GSD AnCora 0.12, 0.36, 1.00	T
Chinese	compound	0.80	GSD GSDSimp PUD HK CFL 0.00, 0.00, 0.01, 0.43, 0.80	T
French	csubj	0.80	FTB Spoken PUD GSD ParTUT Sequoia 0.20, 0.20, 0.75, 0.81, 0.97, 1.0	F
Galician	nummod	0.80	TreeGal CTG 0.18, 0.98	T
Arabic	nummod	0.75	PUD NYUAD PADT 0.24, 0.47, 0.99	?
French	cop	0.73	GSD FTB PUD Sequoia Spoken ParTUT FQB 0.02, 0.03, 0.04, 0.05, 0.05, 0.08, 0.75	F
Latin	cc	0.72	ITTB PROIEL Perseus 0.04, 0.71, 0.76	T
Komi Zyrian	cop	0.69	Lattice IKDP 0.12, 0.81	?
French	compound	0.67	PUD Spoken ParTUT 0.33, 0.78, 1.00	F/T
Portuguese	compound	0.63	PUD Bosque 0.37, 1.00	T
German	compound	0.62	PUD GSD HDT LIT 0.34, 0.53, 0.94, 0.96	F
English	compound	0.59	EWT PUD ESL GUM LinES ParTUT 0.09, 0.09, 0.13, 0.13, 0.18, 0.67	F/T

Table 1: Top 20 relations with largest differences in branching direction within a language. The abbreviations written above numbers are the treebank shortcuts. The last column shows whether the difference is a true (T) or false (F) inconsistency or whether the case in unclear (?).

Inconsistent Chinese classifiers (clf) Figure 1 shows that the annotation for classifiers is categorically different among the Chinese treebanks. The treebanks CFL and HK follow the UD guidelines[5] and attach the classifier to the numeral whereas the other treebanks GSD and PUD attach it to the noun. Note that the annotation of numerals is also different; however, since the branching direction stays the same this cannot be detected by our method.

Figure 1: Chinese annotation scheme for noun phrases with a classifier in CFL/HK (left) and GSD/ GSDSimp/PUD (right).

Inconsistent Korean auxiliaries (aux) All three Korean treebanks have a different usage of the "aux" relation (Figure 2), which is already documented by Noh et al. (2018): Kaist is most closely to the UD guidelines, using the "aux" relation together with the "AUX" tag in around half of its instances (but with a lot of different tags such as "NOUN" and "SCONJ" in the other half), whereas GSD generally uses "flat" and "VERB" to connect auxiliary verbs with their main verbs. In both treebanks, "aux" connects

main and auxiliary verbs from left to right. PUD on the other side uses "aux" to connect main and auxiliary verbs from right to left and tags both with "VERB".
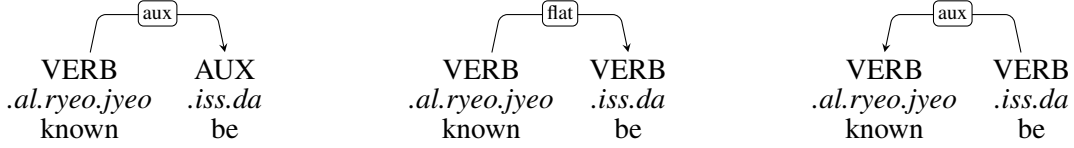

Figure 2: Korean annotation scheme for auxiliary verbs in Kaist (left), GSD (middle) and PUD (right).

In-/Consistent English compounds (compound)   The use of "compound" is only partially inconsistent in English. "compound" connects common-noun compounds right-to-left, whereas "compound:prt" connects verbs and their particles left-to-right, and the ParTUT contains a greater proportion of "compound:prt" instances than the other treebanks. This difference between the treebanks is detected by our method but not an inconsistency. However, the difference is additionally increased by many proper-name compounds in ParTUT that are connected left-to-right with "compound". In the other English treebanks, similar compounds usually show a "flat" relation. Figure 3 shows examples for nominal compounds.
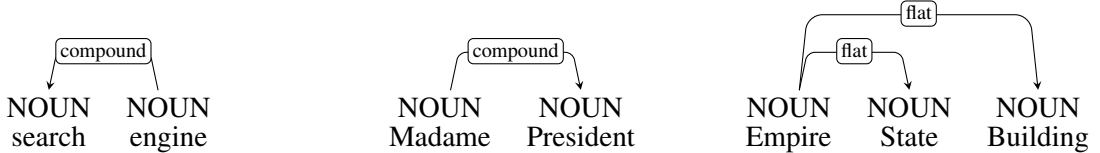

Figure 3: English annotation examples for common-noun compounds (left), proper-name compounds in ParTUT (middle) and proper-name compounds in the other treebanks (right).

4   Experiments

	zh_hk	zh_pud
zh_gsd.orig	48.14	54.97
zh_gsd.clf	50.25	51.51

(a) Traditional Chinese.

	zh_cfl
zh_gsdsimp.orig	43.36
zh_gsdsimp.clf	46.78

(b) Simplified Chinese.

	ko_pud	ko_kaist
ko_gsd.orig	19.10	26.91
ko_gsd.aux	20.13	30.08

(c) Korean.

Table 2: LAS on the test treebanks from models trained on original and converted training treebanks.

Parsing performance of a model trained on one treebank of a language and tested on another treebank of the same language is generally affected negatively by the domain difference, but it will suffer even more seriously in case some of the arc labels are used inconsistently across the treebanks (even though the labels are taken from the fixed UD label set). To empirically demonstrate the impact of annotation inconsistency and verify that our method could alleviate the problem, we perform a pilot experiment on classifiers in Chinese and auxiliaries in Korean, as they are the top two in Table 1.

We designed a set of rules to convert the inconsistent relations in the training treebanks towards the UD guidelines[6]. The target test sets are left as they are, including the ones that are against the general UD guideline (the PUD Treebanks for both languages), since we cannot guarantee the correctness of the automatic conversion. For Chinese classifiers and Korean auxiliaries, there are almost perfect one-to-one correspondences between the annotation schemes which allows a simple conversion of ZH_GSD/GSDSimp and KO_GSD, which is shown in Figure 4.
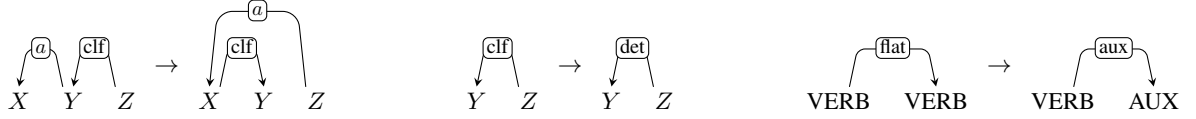
---

[6] https://universaldependencies.org/u/dep/

Figure 4: Conversion rules for instances of "clf" with modifier (left) and without modifier (middle) in Chinese, and for instances of "flat" to "aux" in Korean. X, Y, Z and a are variables for any POS tag or dependency relation, respectively.

We use the UDPipe parser (Straka et al., 2016) in our experiments with the default settings. It is based on the Arc-Standard projective transition system using a neural network classifier. It uses predicted POS tags and morphological features by the tagger.

Table 2 shows the Labeled Attachment Scores (LAS) on the test sets by models trained on the original and converted training sets, noted by the suffixes ".orig" and the respective converted relations. For traditional Chinese, converting the classifier relation brings 2.11 points improvements on ZH_HK, while decreases by 3.56 points on ZH_PUD, since ZH_PUD annotates the classifier similar to the original ZH_GSD. For simplified Chinese, the converted model also outperforms the original by 3.42. For Korean, PUD is improved by 1.03 and Kaist by 3.17.

We then break down the parser accuracy on each individual relation. As expected, the F1 score of "clf" in the Chinese treebanks improves from 0 to 69.96 for HK, from 0 to 84.34 for CFL, and from 65.05 to 0 for PUD (since PUD is annotated the same as the original GSD). For Korean, the F1 score of "aux" changes from 0.93 to 63.53 for Kaist, but remains 0 for PUD, since PUD annotates "aux" different from both GSD and Kaist, as explained in Section 3.

Generally, it is clear that by identifying and converting the inconsistencies, the parser model can learn to handle the relations consistent to the target treebank. It improves the parsing accuracy on the particular relations as well as the overall performance.

5   Conclusion and future work

In this pilot study, we show that the measure of relative branching direction applied on two UD treebanks for the same language can be used as a simple, but effective means for detecting candidates for inconsistencies in the usage of UD relation labels – independent of prior assumptions of which relations are particularly problematic. We manually inspected a list of relations with high divergence of branching direction, and as a pilot experiment converted some identified relations and analyze the performance of a parser in the out-of-domain scenario. The experiment confirms the benefit of our identification and conversion of the inconsistent annotation.

Concurrently with this work, Aggarwal and Zeman (2020) propose another method that addresses the identification of cross-treebank inconsistencies by measuring the differences of POS-tag trigrams. However, it focuses on a different type of inconsistency than what our method can identify. Both types of inconsistency can occur independently and a joined approach of the two complementary methods could make further contribution to error detection/correction in the UD treebanks.

References

Akshay Aggarwal and Daniel Zeman. 2020. Estimating POS annotation consistency of different treebanks in a language. In *Proceedings of the 19th International Workshop on Treebanks and Linguistic Theories (TLT 2020)*.

Chiara Alzetta, Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2017. Dangerous relations in dependency treebanks. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, pages 201–210, Prague, Czech Republic.

Adriane Boyd, Markus Dickinson, and W. Detmar Meurers. 2008. On detecting errors in dependency treebanks. *Research on Language and Computation*, 6(2):113–137.

Xinying Chen and Kim Gerdes. 2017. Classifying languages by dependency structure. Typologies of delexicalized universal dependency treebanks. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 54–63, Pisa, Italy, September. Linköping University Electronic Press.

Marie-Catherine de Marneffe, Matias Grioni, Jenna Kanerva, and Filip Ginter. 2017. Assessing the annotation consistency of the Universal Dependencies corpora. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 108–115, Pisa,Italy, September. Linköping University Electronic Press.

Tillmann Dönicke, Xiang Yu, and Jonas Kuhn. 2020. Real-valued logics for typological universals: Framework and application. In *Proceedings of the 28th International Conference on Computational Linguistics*.

Tillmann Dönicke. 2020. Evaluation of complex typological universals with language vectors and real-valued logics. Master's thesis, University of Stuttgart.

Kira Droganova, Olga Lyashevskaya, and Daniel Zeman. 2018. Data conversion and consistency of monolingual corpora: Russian UD treebanks. In *Proceedings of the 17th international workshop on treebanks and linguistic theories (TLT 2018)*, pages 52–65.

Youngbin Noh, Jiyoon Han, Tae Hwan Oh, and Hansaem Kim. 2018. Enhancing universal dependencies for Korean. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 108–116, Brussels, Belgium, November. Association for Computational Linguistics.

Milan Straka, Jan Hajič, and Jana Straková. 2016. UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Noëmi Aepli, Željko Agić, Lars Ahrenberg, Gabrielė Aleksandravičiūtė, Lene Antonsen, Katya Aplonova, Maria Jesus Aranzabe, Gashaw Arutie, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Victoria Basmov, Colin Batchelor, John Bauer, Sandra Bellato, Kepa Bengoetxea, Yevgeni Berzak, Irshad Ahmad Bhat, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Rogier Blokland, Victoria Bobicev, Loïc Boizou, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Kristina Brokaitė, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Savas Cetin, Fabricio Chalub, Jinho Choi, Yongseok Cho, Jayeol Chun, Alessandra T. Cignarella, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Marine Courtin, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Elvis de Souza, Arantza Diaz de Ilarraza, Carly Dickerson, Bamba Dione, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Hanne Eckhoff, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Tomaž Erjavec, Aline Etienne, Wograine Evelyn, Richárd Farkas, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Kim Gerdes, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Bernadeta Griciūtė, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Nizar Habash, Jan Hajič, Jan Hajič jr., Mika Hämäläinen, Linh Hà Mỹ, Na-Rae Han, Kim Harris, Dag Haug, Johannes Heinecke, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Jena Hwang, Takumi Ikeda, Radu Ion, Elena Irimia, Ọlájídé Ishola, Tomáš Jelínek, Anders Johannsen, Fredrik Jørgensen, Markus Juutinen, Hüner Kaşıkara, Andre Kaasen, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Boris Katz, Tolga Kayadelen, Jessica Kenney, Václava Kettnerová, Jesse Kirchner, Elena Klementieva, Arne Köhn, Kamil Kopacewicz, Natalia Kotsyba, Jolanta Kovalevskaitė, Simon Krek, Sookyoung Kwak, Veronika Laippala, Lorenzo Lambertino, Lucia Lam, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phng Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Cheuk Ying Li, Josie Li, Keying Li, KyungTae Lim, Maria Liovina, Yuan Li, Nikola Ljubešić, Olga Loginova, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Sarah McGuinness, Gustavo Mendonça, Niko Miekka, Margarita Misirpashayeva, Anna Missilä, Cătălin Mititelu, Maria Mitrofan, Yusuke Miyao, Simonetta Montemagni, Amir More, Laura Moreno Romero, Keiko Sophie Mori, Tomohiko Morioka, Shinsuke Mori, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Robert Munro, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Lng Nguyễn Thị, Huyền Nguyễn Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Adéyọ Olúòkun, Mai Omura, Petya Osenova, Robert Östling, Lilja Øvrelid, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Angelika Peljak-Łapińska, Siyao Peng, Cenel-Augusto Perez, Guy Perrier, Daria Petrova, Slav Petrov, Jason Phelan,

Jussi Piitulainen, Tommi A Pirinen, Emily Pitler, Barbara Plank, Thierry Poibeau, Larisa Ponomareva, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Peng Qi, Andriela Rääbis, Alexandre Rademaker, Loganathan Ramasamy, Taraka Rama, Carlos Ramisch, Vinit Ravishankar, Livy Real, Siva Reddy, Georg Rehm, Ivan Riabov, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Luisa Rocha, Mykhailo Romanenko, Rudolf Rosa, Davide Rovati, Valentin Roșca, Olga Rudina, Jack Rueter, Shoval Sadde, Benoît Sagot, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Dage Särg, Baiba Saulīte, Yanin Sawanakunanon, Nathan Schneider, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Muh Shohibussirri, Dmitry Sichinava, Aline Silveira, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Isabela Soares-Bastos, Carolyn Spadine, Antonio Stella, Milan Straka, Jana Strnadová, Alane Suhr, Umut Sulubacak, Shingo Suzuki, Zsolt Szántó, Dima Taji, Yuta Takahashi, Fabio Tamburini, Takaaki Tanaka, Isabelle Tellier, Guillaume Thomas, Liisi Torga, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Andrius Utka, Sowmya Vajjala, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Eric Villemonte de la Clergerie, Veronika Vincze, Lars Wallin, Abigail Walsh, Jing Xian Wang, Jonathan North Washington, Maximilan Wendt, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska, Mary Yako, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Zhuoran Yu, Zdeněk Žabokrtský, Amir Zeldes, Manying Zhang, and Hanzhi Zhu. 2019. Universal dependencies 2.5. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Appendix A. Analysis of other potential cross-treebank inconsistencies

Inconsistent Arabic compounds (compound) In the NYUAD treebank, the "compound" relation is only used to connect numerals (NUM tags) from left to right (in the PUD treebank and the PADT, words tagged as NUM are commonly connected by "nummod"). Whilst the "compound" relation is not used in the PADT at all, there are two usages in the PUD treebank: First, "compound" to connect nouns, also from left to right, but rarely used. Second, "compound:prt" to connect words, mainly verbs (VERB), with their particles (PART). Here, the particle almost always precedes its governor. The labelling of particles is, obviously, handled differently in the PADT and the NYUAD treebank.

Inconsistent Galician coordinating conjunctions (cc) Whilst the TreeGal treebank follows the UD guidelines[7] and combines the "cc" and the "conj" relation, the CTG treebank uses "cc" differently (as shown in Figure 5) and does not use "conj". For conjuncts that are nouns, the "nmod" relation is used instead, which makes the "nmod" relation ambiguous between a nominal modifier and a nominal conjunct. Conjuncts which are not nouns are connected similarly using other relations.
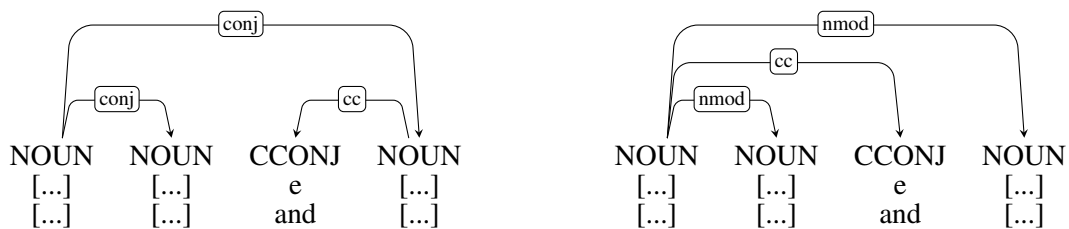

Figure 5: Galician annotation scheme for coordinating conjunctions in TreeGal (left) and CTG (right).

Consistent English clausal subjects (csubj) The presence of an expletive pronoun changes the direction of the "csubj" relation, see Figure 6. The GUM treebank simply contains more sentences with expletives.

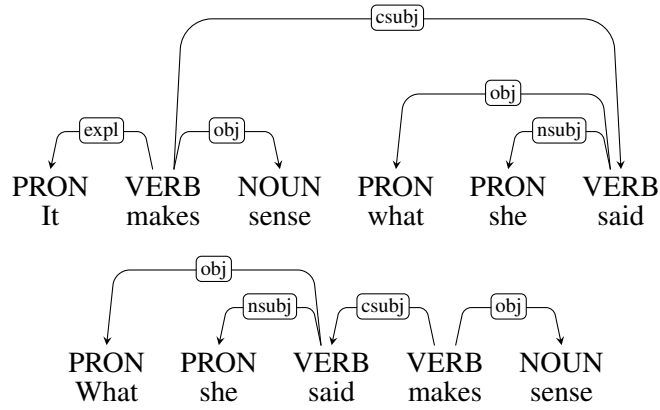[7] https://universaldependencies.org/u/dep/cc.html

Figure 6: English annotation scheme for clausal subjects with expletive (above) and without expletive (below).

Consistent Old Russian expletives (expl) The only expletives in the Old Russian treebanks are forms of *sebe/sja* 'oneself'. The TOROT contains over 1,000 examples and in 93% of the cases the expletive follows the verb; the RNC treebank addresses those forms with the subtype relation "expl:rv" and contains only nine examples, in all but one of them the expletive precedes the verb. Furthermore, in the RNC treebank the base "expl" relation is used to connect particles (PART) with their governors and the particle usually precedes its governor.

Inconsistent Spanish compounds (compound) First of all, the PUD treebank only contains "compound:prt" whereas the GSD treebank and the AnCora treebank only contain "compound". The usage varies in several points, for example with temporal words and reflexive words:

Figure 7: Spanish annotation scheme for temporal words and *pasado* 'last' in AnCora (left) and GSD/ PUD (right).
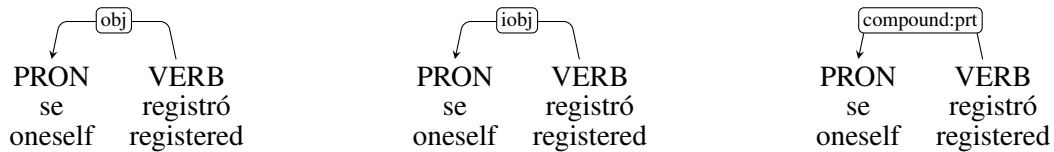

Figure 8: Spanish annotation scheme for reflexive verbs in AnCora (left), GSD (middle) and PUD (right).

Inconsistent Chinese compounds (compound) All Chinese treebanks use the base "compound" relation and the direction is generally from right to left. However, the HK treebank and the CFL treebank also use "compound:dir", "compound:ext", "compound:vo" and "compound:vv" where the direction is generally from left to right. The extended use of "compound" (by subtypes) in some but not all Chinese treebanks is an inconsistency.

Consistent French clausal subjects (csubj) In French, the clausal subjects show the same expletive alternation as in English (see above).

Inconsistent Galician numeral modifiers (nummod) The CTG treebank attaches numerals to the determiner instead of the noun (whereas both determiner and numeral should be attached to the noun according to the UD guidelines):

Figure 9: Galician annotation scheme for numeral modifiers in TreeGal (left) and CTG (right).

Consistent French copulae (cop) The FQB contains a lot of questions where fronted forms of *quel* 'which' are the governor of the copula, i.e. the fronted forms change the average direction of the "cop" relation.

Inconsistent Latin coordinating conjunctions (cc) The conjunction is categorically attached to different conjuncts in the Latin treebanks, analogous to Galician (see above).

In-/Consistent French compounds (compound) The "compound" relation is consistently used for common-noun compounds in the PUD treebank, the Spoken treebank and the ParTUT. An inconsistency, however, which cannot be detected by dependency direction difference, is that the other French treebanks mark compounds by other relations, e.g. "nmod".

Inconsistent Portuguese compounds (compound) The Portuguese treebanks show a different use of "compound" and "compound:prt", similar to the Spanish compounds.

Consistent German compounds (compound) The German treebanks differentiate between "compound" and "compound:prt". The former is used for (hyphenated) nominal compounds and usually spans from right to left; the latter is used for verbal particles and mainly goes the opposite direction, from left to right.
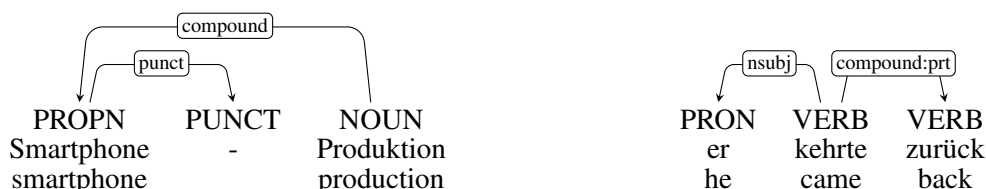

Figure 10: German annotation scheme for nominal compounds (left) and particle verbs (right).