Building Annotated Parallel Corpora Using the ATIS Dataset: Two UD-style treebanks in English and Turkish

Neslihan Cesur[1,2], **Aslı Kuzgun**[1], **Mehmet Köse**[3], **Olcay Taner Yıldız**[2]
Starlang Yazılım Danışmanlık[1], Işık University[3], Özyeğin University[2],
Istanbul, Turkey
{neslihan, asli}@starlangyazilim.com, mehmetkse@gmail.com, olcay.yildiz@ozyegin.edu.tr

Abstract

In this paper, we introduce the annotation process of the Air Travel Information Systems (ATIS) Dataset as a parallel treebank in English and in Turkish. The ATIS Dataset was originally compiled as pilot data to measure the efficiency of Spoken Language Systems and it comprises human speech transcriptions of people asking for flight information on the automated inquiry systems. Our first annotated treebank, which is in English, includes 61.879 tokens (5.432 sentences) while the second treebank, which was translated into Turkish, contains 45.875 tokens for the same amount of sentences. First, both treebanks were morphologically annotated through a semi-automatic process. Later, the dependency annotations were performed by a team of linguists according to the Universal Dependencies (UD) guidelines. These two parallel annotated treebanks provide a valuable contribution to language resources thanks to the spontaneous/spoken nature of the data and the availability of cross-linguistic dependency annotation.

Keywords: Universal Dependencies, ATIS, Annotated Corpus, Parallel Corpora

1. Introduction

Large natural language corpora, whether it includes spoken or written data, are a crucial asset to natural language processing (NLP) research when it comes to building intelligent systems which can understand, manipulate and produce human language. Manually and systematically parsed, gold-standard treebanks provide important resources especially for the training and evaluation of parsers.

As most of the available corpora are monolingual, parallel corpora which contain the same content in two or more languages constitute valuable linguistic resources for supervised machine learning applications. Thanks to parallel corpora, we can build state-of-the-art multilingual parsers and evaluate parser quality using multiple languages. Parallel corpora are also beneficial for building tools such as machine translation systems and multilingual question answering systems. The ATIS parallel treebank will be part of four datasets which we hope to use in training a bilingual parser. Two of these treebanks are already available online: the PUD treebank and the Penn Treebank in English and Turkish (Kuzgun et al., 2020). The third dataset, a parallel QuestionBank is currently being annotated by our team.

The parallel ATIS treebank[1] is built as a dependency treebank in English and Turkish, in accordance with the Universal Dependencies (UD) guidelines. The treebank is comprised of annotated data from the Air Travel Information System (ATIS) Dataset (Hemphill et al., 1990). This dataset was originally collected as a pilot corpus to evaluate the progress in Spoken Language Systems. It comprises transcripts of spoken data in which customers are inquiring about flight information. As a strictly domain-specific corpus, the data mostly contains names of cities, airports, airlines and flight numbers. As the vast majority of natural language corpora are made up of samples of written language, the main advantage of the ATIS Dataset is that it contains samples of spontaneous speech. As the data is not pre-written, the corpus contains incomplete sentences and errors in speech, which differentiates it from most corpora comprising written natural language data.

Both of our treebanks include 5,432 sentences (the treebank in English has 61,879 tokens while the treebank in Turkish has 45,875 tokens due to the agglutinative nature of Turkish). In this paper, we outline the steps of our annotation process and present quantitative results through a comparison of our treebanks. The annotation of the English ATIS treebank is made up of two main stages: automatic POS-tagging (later controlled by human annotators) and dependency annotation. The Turkish treebank, however, required five stages: the semi-automatic translation of the dataset, automatic morphological analysis, morphological disambiguation, automatic POS-tagging, and dependency annotation. The morphological and syntactic annotations were carried out by three annotators with a background in linguistic studies while the translation team included three translators with a background in linguistics and translation studies. The anno-

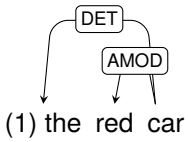
[1] Both versions of the treebank can be accessed on the website of Universal Dependencies Project.

BUCC 2024: The 17th Workshop on Building and Using Comparable Corpora, pages 104–110
20 May, 2024.

tation decisions specific to the data were made prior to the annotation process through open discussions.

2. The Universal Dependencies Project

In terms of the syntactic framework they follow, treebanks can either be annotated using phrase structure grammar or dependency grammar. Phrase structure grammar, whose foundations were laid by Noam Chomsky in the 1950s (Chomsky, 1957), consists of branching trees which group together constituents under labeled nodes. Dependency grammar, which was first popularized by the French linguist Lucien Tesnière in the 1950s[2] (Tesnière, 1959), is a framework which aims to mark one-to-one syntactic relations between the constituents of a given phrase or sentence. In DG, each element in a sentence is considered a node and is linked to another element through head-dependent relations. The example in (1) demonstrates the head-dependent relationship in a noun phrase. The element *car* is the head of the phrase. The determiner (DET) and the adjectival modifier (AMOD) are linked to the head *car* as dependents.



We have used the tags and rules of the Universal Dependencies (UD) project for our annotation. UD (Nivre et al., 2016) is a project which aims to provide a consistent and cross-linguistic annotation scheme for parts of speech (POS) tagging and dependency syntax. With more than 100 languages currently available for open-access, the Universal Dependencies project provides great resources for cross-lingual learning and multilingual parser development.

The first annotated treebank in Turkish is the METU-Sabancı Treebank (Oflazer et al., 2003; Atalay et al., 2003; Sulubacak et al., 2016). The Turkish Penn Treebank (Kuzgun et al., 2020) corpus is the largest Turkish dependency treebank currently available with 183,555 tokens. Moreover, The Penn Treebank corpus and the PUD treebank are the only multi-lingual treebanks which include Turkish. Amongst these annotated corpora, ATIS constitutes a crucial contribution in that it not only introduces the first treebank which is comprised of spoken natural language data but it is also another parallel treebank.

[2]Tesnière's work on syntax and dependency grammar was published posthumously.

3. Annotation Process

3.1. Translation

The ATIS dataset was loaded from an open-source repository, currently available on GitHub[3]. Before the annotation process, the ATIS Dataset was translated into Turkish by a team of seven translators. The translation was carried out on Google Sheets, which allowed the team to work simultaneously on an online platform. English sentences were listed in one column, with their corresponding Turkish translations added to the adjacent row. Figure 1 illustrates some English sentences with their Turkish counterparts. The translators adopted a semi-automatic translation strategy by translating the sentences with the help of different machine translation tools. Then, the outputs were checked and corrected by the human translators to ensure that the correspondence between the two languages was accurate. This process was important to keep the originality of the English data, including the absence of punctuation marks and the use of discourse particles such as *now* and *okay* at the beginning of sentences. As Figure 1 illustrates, the original sentences do not include question marks or periods at the end of sentences contrary to most written natural language corpora.

3.2. Morphological Analysis

Both morphological and syntactic annotations were carried out with the same interface called StarDust, introduced in (Yenice et al., 2022). StarDust is packaged as a JAR (Java ARchive) file and is implemented using the Java programming language. We opted for this interface because it provides a user-friendly interface for annotators and it can run different annotation programs such as POS-tagger, morphological analyzer and dependency annotator.

The English dataset only required POS-tag annotation whereas Turkish was a lot more complicated to analyze due to its agglutinating morphological structure. The morphological annotation of the English data consisted of POS-tagging the tokens, using the Penn Part of Speech Tags[4] (Marcus et al., 1993). Within the interface used for annotation, the POS-tag detection took place automatically. After the tags were determined, the roots of the tokens were automatically selected by the analyzer through a rule-based algorithm. The rules consisted of removing the inflections found on the token and marking the remaining part as the root. For instance, if the plural noun *flights* is marked with the tag NNS (used to indicate plural nouns), the plural marker as omitted and the remaining part is selected as

[3]https://github.com/howl-anderson/ATIS_dataset

[4]https://www.cis.upenn.edu/~bies/manuals/tagguide.pdf.

	A	B	C	D
1	**TYPE**	**NO**	**ENGLISH**	**TURKISH**
2	train	1	what is the cost of a round trip flight from pittsburgh to atlanta beginning on april twenty fifth and returning on may sixth	Pittsburgh'tan Atlanta'ya 25 Nisan'da gidiş 6 Mayıs'ta dönüşü olan bir gidiş dönüş uçuşunun maliyeti nedir
3	train	2	now i need a flight leaving fort worth and arriving in denver no later than 2 pm next monday	Şimdi Fort Worth'dan ayrılan ve en geç gelecek pazartesi akşam 2'ye kadar Denver'e varacak bir uçağa ihtiyacım var
4	train	3	i need to fly from kansas city to chicago leaving next wednesday and returning the following day	Önümüzdeki çarşamba gidiş ertesi gün dönüşü olan Kansas City'den Chicago'ya giden bir uçuşa ihtiyacım var

Figure 1: The translation sheet with English sentences appearing in Column C and their corresponding Turkish translations in Column D.

the root of the token. For exceptional cases such as suppletive forms (like *are* and *were* having the root *be*), separate rules were implemented for the selection of the root. In Figure 2, black words indicate the tokens and blue words indicate the root of the tokens. According to our rules, the root of the token *are* is automatically determined as *be* and the root of *flights* is determined as *flight*. POS-tags are indicated in red and can be modified by the annotators by clicking on the token.

After the tags were checked and manually corrected by our annotators, the Penn POS-tags were automatically converted into UD-style tags, called Universal POS-tags[5]. This was also done by a rule-based algorithm. For instance, these rules automatically convert noun tags such as NN, NNS and NNP into a NOUN UD tag. The PRP (personal pronoun) tag is converted to PRON tag and so on. These UD POS-tags are visible to the annotators during the dependency annotation process as shown in Figure 3.

Figure 2: A view of the POS-tagger, showing the tokens in black, the roots in blue and the Penn POS-tags in red.

As for the morphological analysis of the Turkish treebank, a rule-based morphological analyzer by Yıldız et al. (2019) was implemented. This open-source morphological analyzer works with a lexicon and a finite state transducer. It lists out the derivations for every possible root of a given token along with every possible morphological tag of a given suffix. After this automatic morphological analysis which separated the tokens into possible roots and affixes, a manual morphological disambiguation

[5] https://universaldependencies.org/u/pos/

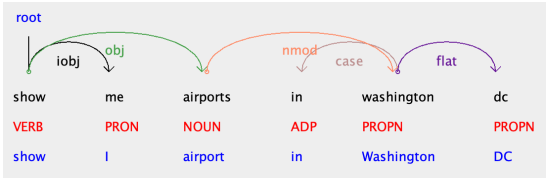


Figure 3: A view of the dependency annotator with the UD POS-tags and roots indicated below the tokens.

was carried out by our annotators in order to select the correct analysis for each token. One reason why this step is crucial for Turkish is because the same form can correspond to different morphological tags depending on the context. For instance, when a word receives the suffix *-ı*, one needs to decide whether it is an accusative marker found on direct objects or a third person possessive marker. Another example is shown in Figure 4, in which the token *sabahın* receives two possible analyses due to the suffix *-(n)In* which is added to the root *sabah* 'morning'. The suffix can either correspond to a second person possessive marker (as in *your morning*) or a genitive marker (as in *early hours of the morning*). The second option is selected according to the context in the given example.

Figure 4: A view of the Turkish morphological analyzer, showing two possible analyses for the token *sabahın*.

After the manual morphological disambiguation, the tokens are automatically assigned their UD POS-tags according to their final morphological tags. As with the sentences in English, the interface makes the UD POS-tags visible to annotators

during the dependency annotation stage, as shown in Figure 5.

3.3. Dependency Annotation

After the morphological analysis/disambiguation of Turkish tokens and the assignment of Universal POS-tags of both treebanks, the dependency annotations were carried out by the same three annotators. The annotations took place on the same open-source interface (Yenice et al., 2022) which was used for the morphological analysis and POS-tagging. During this stage, the annotators determined the heads and dependents in a given sentence or phrase and labeled them with the appropriate UD dependency tags. Images 3 and 5 show how the arrows depart from the head and point to the dependent. Each relation is marked with a separate color and the corresponding tag is shown in the arch of the arrow. Moreover, Figure 6 shows a larger overview of the interface including the tag box. When the annotator drags the dependent towards the head, the tag box pops up. As the interface allows for a connection between the layers of the POS-tagger and dependency annotator, the possible UD tags which are available for a specific relation are automatically restricted to enable a faster selection. Moreover, errors in annotation violating the UD rules are automatically detected and indicated at the bottom of the screen.

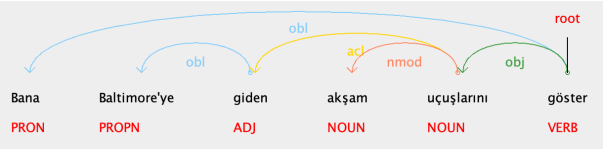


Figure 5: A view of the dependency annotator, showing how heads are connected to their dependents with arrows.

A total of 32 UD tags were used in the syntactic annotation process. Table 1 and 2 illustrate the 10 most frequently used UD tags in the Turkish and English ATIS treebanks, respectively.

We observe that the frequency for the NMOD (nominal modifier) tag is higher than the ROOT tag in both languages. This shows that most of the sentences contain more than one nominal modifier. Even though nominal modifiers are common in most treebanks, the significant number in our treebank points to the frequency of phrasal elements such as *from Burbank*, *in Washington*, etc. Also, the fact that the CASE tag (which marks adpositions) is more common than the ROOT tag in English points to the abundance of prepositions indicating location and direction such as *in*, *to* and *from*, which is also shown in Table 4. The dependency representations below show examples from

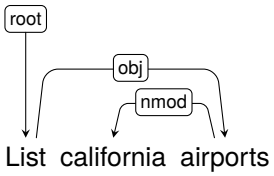
Dependency Relation	Frequency
NMOD	11.626
ROOT	5.431
OBL	4.193
FLAT	3.266
AMOD	2.928
OBJ	2.744
ACL	2.038
NSUBJ	1.824
COMPOUND	1.470
DET	1.305

Table 1: Top 10 most frequent dependency tags and their frequencies in the Turkish ATIS treebank

Dependency Relation	Frequency
CASE	13.131
NMOD	8.568
ROOT	5.432
DET	4.738
NSUBJ	3.323
OBJ	3.274
FLAT	3.148
OBL	3.130
COMPOUND	2.377
AMOD	1.787

Table 2: Top 10 most frequent dependency tags and their frequencies in the English ATIS treebank

the English ATIS treebank. The first example shows a case where the NMOD tag is used within a noun phrase. The nominal modifier *california* -which is the dependent- is linked to the head of the phrase, *airports*. The verb *list* is marked as the ROOT and the head of the noun phase *airports* is marked as the object (OBJ) of the main verb. The second example shows a noun phrase with the head noun *flights*. The phrases *from Las Vegas* and *to Burbank* are attached to the head noun as nominal modifiers. We also see the use of the two most common words in the English dataset, *to* and *from*, attached to different noun heads with the CASE tag.



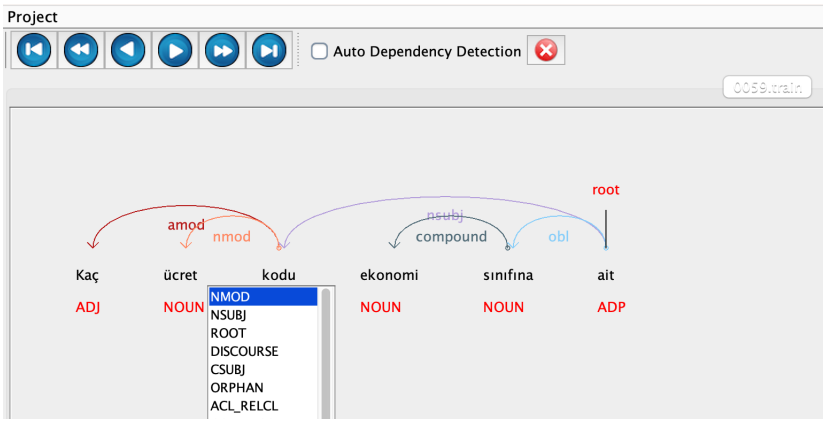
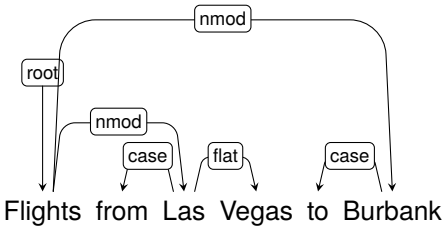


Figure 6: An overview of the dependency annotation interface, illustrating the control buttons and the tag box.



What is interesting is that we do not find the PUNCT tag in neither of the lists even though it usually appears quite frequently in general-purpose datasets as each sentence usually ends with a punctuation mark. As we have indicated in Section 3.1 and Figure 1 the original dataset lacks the usual punctuation marks. Their absence in our treebanks points to the fact that ATIS is a dataset of spoken natural language data. As the transcripts do not include punctuation, we observe a discrepancy compared to treebanks with written language data. The table below shows inter-annotator agreement scores for both treebanks. DEP shows the percentage of dependencies linked to the correct head with the correct tag. TO shows the percentage of dependencies linked to the correct head. TAG shows the percentage of dependencies which were marked with the correct tag.

	DEP	TO	TYPE
Turkish	78	84	82
English	82	91	86

Table 3: Inter-annotator agreement scores for both languages

Overall, the percentages for English are higher for each score type. The linking between heads and their dependents is more accurate than the selected tag in both languages.

4. A Quantitative Analysis of the Treebanks

As we have already stated, the Turkish ATIS Treebank comprises a total of 45,875 tokens while the English ATIS Treebank has 61,879 tokens. Considering that the number of annotated sentences are the same, the significant difference between token numbers point to the distinct morphological nature of the two languages.

Moreover, due to the same morphological pattern, we observe that the English dataset has less unique surface forms[6] (932 unique surface forms) than the Turkish dataset (2,133 unique surface forms), despite containing more tokens. This means that 4.64% of the Turkish dataset consists of unique forms while for English, this percentage is around 1.5. One reason for this discrepancy can be found in prepositional phrases. A state/city name in English can only appear in its bare form in English (such as *Denver* or *Boston*). The directionality information is conveyed through prepositions such as *to* and *from*. However, in Turkish, the directionality is expressed using nominal case markings such as the dative form (*Denver'a*), the locative form (*Denver'da*) and the ablative form (*Denver'dan*). If we consider that these case markings are suffixed to each location name, we end up with a greater number of unique forms in Turkish. For each location name, only 1 unique word is added in English (the bare form of the location name) while in Turkish, four unique forms (considering only the nominative, locative, dative and ablative forms) are created.

Another significant comparison can be made regarding the domain-specific nature of the treebanks. Compared to a dataset including a wider range of topics, a domain-specific treebank is ex-

[6]Each occurrence of a distinct word form is counted as a *unique surface form*. For example, *flight* and *flights* are two unique surface forms in English.

pected to contain less unique surface forms. We can clearly observe this fact when we compare the Turkish ATIS treebank to the KeNet dependency treebank which is also a part of the Universal Dependencies Project. KeNet contains a total of 149,524 tokens[7] amongst which 49,156 are unique forms. This means that while 4.64% of the Turkish ATIS treebank is comprised of unique forms, the KeNet data comprises up to 32.84% of unique forms. This significant difference indicates that as a domain-specific treebank, ATIS shows much less variety in terms of words and word forms. Another domain-specific dependency treebank in Turkish, the Tourism treebank, contains a total of 71,322 tokens and 4,961 unique surface forms which makes up the 6.96% of the dataset. This number is slightly larger than what we have found for the Turkish ATIS treebank. This shows that amongst the Turkish treebanks, the new ATIS dataset is the most specific one with the least amount of diversity in words and word forms.

The effects of domain-specificity can also be observed in the most frequent surface forms. Word frequency lists of more generic datasets usually pattern with the most frequently used words of the given language. These usually include determiners, prepositions, auxiliaries and conjunctions. However, due to their restricted content, domain-specific treebanks include content words relating to the topic of the dataset. Table 4 is a list of the most common 15 words in the ATIS treebanks. We observe different forms of the word *flight* (*uçuşlar* which means *flights* and its accusative form *uçuşları*) in both treebanks. We also find several state/city names (*Boston*, *Denver*, *San Francisco*) and question words. Such specific content words and proper names would not appear as frequently in a dataset containing more generic content. The rest of the words include pronouns (*ben*, *bana*, *I*, *me*), determiners (*the*, *bir*) and prepositions expressing directionality (*to* and *from*).

	Turkish ATIS	English ATIS
1	uçuşları	to
2	San	from
3	olan	flights
4	göster	the
5	uçuş	on
6	bir	what
7	istiyorum	flight
8	uçuşlar	me
9	var	I
10	ve	San
11	bana	Boston
12	Boston'dan	show
13	hangi	a
14	en	Denver
15	Francisco'ya	in

Table 4: Top 10 most frequent surface forms in both ATIS Treebanks

5. Conclusion

This paper was an overview of the morphological and syntactic annotation process of a parallel treebank in English and in Turkish.

Our two annotated treebanks constitute a valuable contribution to the Universal Dependencies project as they are the only annotated dependency treebanks which include solely spoken language data. They also show certain distinct characteristics regarding their domain-specific nature, including a decreased variety in unique forms and a more specific set of most frequent words compared to generic datasets.

Another valuable aspect of our treebanks is that they are bilingual. As we have seen above, this type of treebanks allow for a typological comparison between languages. We have discussed the gap between the number of tokens and the percentage of unique words in order to show that such treebanks offer quantitative measures which point to morphological distinctions between languages. In addition to typological analysis, parallel treebanks can be used for the training of multilingual parsers. In this regard, the ATIS treebanks would be especially useful in training parsers for the analysis of spoken natural language and interpreting simple commands.

[7]The number of tokens indicated here does not include punctuation marks considering that the KeNet dataset includes a great number of punctuation while ATIS does not make use of a significant amount.

6. Bibliographical References

Nart B Atalay, Kemal Oflazer, and Bilge Say. 2003. The annotation process in the turkish treebank. In *Proceedings of 4th International Workshop on Linguistically Interpreted Corpora (LINC-03) at EACL 2003*.

Noam Chomsky. 1957. *Syntactic structures*. Mouton § Co.

Charles T Hemphill, John J Godfrey, and George R Doddington. 1990. The atis spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*.

Aslı Kuzgun, Neslihan Cesur, Bilge Nas Arıcan, Merve Özçelik, Büşra Marşan, Neslihan Kara, Deniz Baran Aslan, and Olcay Taner Yıldız. 2020.

On building the largest and cross-linguistic turkish dependency corpus. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 1–6. IEEE.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Kemal Oflazer, Bilge Say, Dilek Zeynep Hakkani-Tür, and Gökhan Tür. 2003. Building a turkish treebank. *Treebanks: Building and using parsed corpora*, pages 261–277.

Umut Sulubacak, Gülşen Eryiğit, and Tuğba Pamay. 2016. Imst: A revisited turkish dependency treebank. In *Proceedings of TurCLing 2016, the 1st international conference on Turkic computational linguistics*. Ege University Press.

Lucien Tesnière. 1959. *Eléments de syntaxe structurale*. Klincksieck.

Arife B Yenice, Neslihan Cesur, Aslı Kuzgun, and Olcay Taner Yıldız. 2022. Introducing stardust: A ud-based dependency annotation tool. In *Proceedings of the 16th Lingusitic Annotation Workshop (LAW-XVI) within LREC2022*, pages 79–84.

Olcay Taner Yıldız, Begüm Avar, and Gökhan Ercan. 2019. An open, extendible, and fast turkish morphological analyzer. In *International Conference Recent Advances in Natural Language Processing*.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium. Association for Computational Linguistics.