Süleyman Demirel Üniversitesi
Fen Bilimleri Enstitüsü Dergisi
Cilt 27, Sayı 1, 125-136, 2023

Süleyman Demirel University
Journal of Natural and Applied Sciences
Volume 27, Issue 1, 125-136, 2023

DOI: 10.19113/sdufenbed.1168260

A Novel Syntactic-Based Approach to Calculate Similarities Among Languages

Metin BİLGİN[*1]

[1] Bursa Uludağ University, Faculty of Engineering, Computer Engineering, 16059, Bursa, Turkey

(Alınış / Received: 29.08.2022, Kabul / Accepted: 28.11.2022, Online Yayınlanma / Published Online: 25.04.2023)

Keywords
Language Similarity,
Universal Dependency,
Natural Language Processing

Abstract: The approach presented in this study is about the calculation of the similarities among languages by using the new feature template to be obtained from the syntactic analysis phase. Studies were conducted on 6 different language sets from two different language families in order to show the calculability of similarity of languages with the help of the recommended new feature template. In the first study, triplet-pattern template which is obtained from the syntactic analysis of Turkey, Kazakh, and Uyghur Turkish languages from Turkic languages families belonging to the Ural-Altaic linguistic family, could be formed automatically through developed software, and also similarity analysis of the desired languages could be made thanks to a different module developed within the same software. Consequently, not only similar structural relations of the languages from the same language family but also structural differences among the languages can also be revealed. Even if the first aim is to determine the similarities among languages when developing an approach, the real aim of the desired structure is to form a system independent from the language. In order to show that the formed system has a structure independent from the language, another study was carried out. In the second study, the similarities among the languages were determined by using treebanks of English, Swedish and Norwegian from the Germen language family. When the language family is Turkic and the metrics are Jaccard, Dice, and Similarity Matching, the highest similarity is Turkish-Uyghur, and the values of the metrics are 25.21%, 40.27%, and 50.42%, respectively. When the language family is Germen, the highest similarity is Norwegian-Swedish, and the values of the metrics are 37.15%, 54.17%, and 74.3, respectively.

Diller Arasındaki Benzerliği Hesaplamak için Sözdizimsel Yeni Bir Yaklaşım

Anahtar Kelimeler
Dil Benzerliği,
Evrensel Bağımlılık,
Doğal Dil İşleme

Öz: Bu çalışmada sunulan yaklaşım, söz-dizimsel analiz safhasından elde edilecek yeni özellik şablonunun kullanılmasıyla dillerin birbirlerine olan benzerliğinin hesaplanması üzerinedir. Önerilen yeni özellik şablonu yardımıyla dillerin benzerliklerinin hesaplanabilirliğini gösterebilmek için iki farklı dil ailesine mensup 6 farklı dil kümesi üzerinde çalışmalar gerçekleştirilmiştir. İlk çalışmada Ural-Altay dil ailesine ait Türki diller ailesine mensup Türkiye, Kazak ve Uygur Türkçelerinin söz-dizimsel analizden elde edilen üçlü örüntü şablonları geliştirilen yazılım vasıtasıyla otomatik olarak çıkarılabilmekte ve aynı yazılım içerisinde geliştirilen farklı bir modül sayesinde de istenen dillerin benzerlik analizi yapılabilmektedir. Böylece ayni dil ailesine mensup dillerin yapısal olarak birbirlerine benzer ilişkilerinin gösterilmesinin yanı sıra diller arasındaki yapısal farklılıklar da ortaya çıkarılabilmektedir. Yaklaşım geliştirilirken ilk hedef Türki diller arasındaki benzerliklerin belirlenmesi olsa da oluşturulmak istenen yapının gerçek amacı dilden bağımsız bir sistem oluşturabilmektir. Oluşturulan sistemin dilden bağımsız bir yapı oluşturabildiğini gösterebilmek adına ikinci bir çalışma gerçekleştirilmiştir. İkinci çalışmada Germen dil ailesine mensup İngilizce, İsveççe ve Norveççe derlemleri kullanılarak dillerin birbirlerine olan benzerliklerin ölçümlenmesi sağlanmıştır. Dil ailesi Türkçe ve metrikler Jaccard, Dice ve Similarity matching olduğunda, en yüksek benzerlik Türkçe-Uygurca olup, metriklerin değerleri sırasıyla %25.21, %40.27 ve %50.42'dir. Dil ailesi Germen olduğunda en yüksek benzerlik Norveç-İsveççe olup, metriklerin değerleri sırasıyla %37.15, %54.17 ve %74.3'tür.

[*] *İlgili yazar: metinbilgin@uludag.edu.tr*

1. Introduction

One of the most important characteristics of human beings is to have language. By way of language, human beings can share their feelings, express their desires or signal a danger [1]. When humans develop their language aptitude, they also develop their communication skills at the same time. In this way, while they can emphasize better what they want to say, they also begin to understand better what is said. Social interactions formed thanks to the language are quite significant components in human life [2]. By virtue of the language, some things like one's age [3], sex [4, 5], political view [6], eating habits [7] could be estimated.

Despite the changes caused because of every kind of external factor, it is expected that the basic syntax of the languages from the same family should be similar. Syntax analysis is the process of extracting the internal structures of the sentences given in a natural language and it is also quite an important pre-step for semantic analysis which is the phase of inferring from a text given in a natural language.

Several studies like definition, differences, or similarity of languages have been conducted until today by not only philologists but also computer scientists together with the developing technology. The studies about language similarities are predominantly the measurements on semantic similarities of words. However, syntax is actually one of the most important fundamental differences distinguishing human beings from many other living beings [8]. In order to achieve the syntaxes, we also need to realize a stage named parsing. The parsing process is a very crucial component for artificial intelligence applications developed in recent years. For its usage area, machine translation [9, 10], information retrieval [11, 12], text recognition [13], sentiment analysis [14-16] and text clustering [17] can be given as examples.

Language similarity can be defined as typologically (word order/word complexity) or genetically/historically (Indian-Europe/China-Tibet). Examples of studies conducted on syntax and language similarities will be given in this part of the study. Potthast et al. have divided the similarity identification approaches between cross language into 5 categories as shown in Figure 1 [18]. They have conducted studies on similarities of documents belonging to 6 different languages.

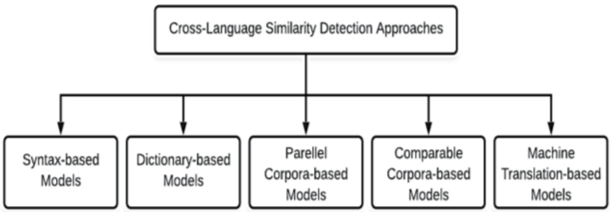
Cross-Language Similarity Detection Approaches
Syntax-based Models
Dictionary-based Models
Parellel Corpora-based Models
Comparable Corpora-based Models
Machine Translation-based Models

Figure 1. Cross-Language Similarity Detection Approaches

Crossley and McNamara have designed a system that functions in accordance with the average number of the clause in a sentence, the number of the word before the main verb in the main clause, and also syntactic similarities within the whole document [19]. Kyle has developed a syntactic tool called Tool for the Automatic Assessment of Syntactic Sophistication and Complexity (TAASSC). This is a tool calculating syntactic index with regards to the number of adjectives and adverbs per sentence [20].

Furthermore, Pennebaker estimates the syntactic similarity in her/his study through the instrument of a function which is a system using words and named as Linguistic Style Matching (LSM) [21, 22]. LSM finds out the similarity between two documents by estimating the score of the syntactic similarity with regard to the predefined word category score [23]. Gamallo et al. have developed a syntactic-based system in order to determine the similarities of the words [24]. Li et al. have performed similarity tests on the texts of 900 Chinese and 100 Lao languages in order to calculate the similarity between Chinese and Lao languages as Wordnet-based [25]. Dinh and Thanh have suggested the fuzzy-based method to define the comments of English and Vietnamese languages. This method tries to calculate the similarities by comparing every word with a fuzzy set including similar words in order to determine whether the two sentences are similar to each other or not [26].

Moreover, Steinberger et al. have used a polyglot dictionary named EUROVOC to estimate the similarities of the texts independently of the language [27]. Baron-Cedeno et al. have suggested an algorithm named as Cross-Lingual Plagiarism Analysis (CLIPA). This algorithm is based on the statistical transition model and it is a system that finds out the probability of the similarity of two languages based on Bayes principle [28].

Further, Uszkoreit et al. has conducted a study in order to find out the similar texts between two languages by calculating n-grams based on dictionary translation [29]. Maike et al. have also carried out a study on the translation from a source language to a target language by calculating the similarities in target language space [30].

2. Material and Method

2. 1. Datasets

In this part, after being given the briefing about the Universal Dependencies project (UDP), the information about used datasets is going to be presented.

UDP is a framework that includes reciprocal consistent explanations among different languages. The objectives of the UDP might be assumed to analyze the research from the perspective of a language as well as to develop multilingual decompositions and facilitate the learning process among languages. It is benefited from Stanford dependencies [31‐33], Google Universal part-of-speech tags [34] and Interset Interlingua for morphosyntactic tagsets [35] to form explanation schemas.

UDP is developed as an open platform that has many project members. Its first version was released in 2015 and 10 treebanks in 10 different languages have been formed with its 1.0 version. Together with the increasing number of participants within the years, the 2.5 version to which also the datasets in this study belong, was created. In this version, there are 157 different treebanks from 90 different languages [36]. Parallel examples of the different languages formed by using the UDP framework could be seen in Figure 2.

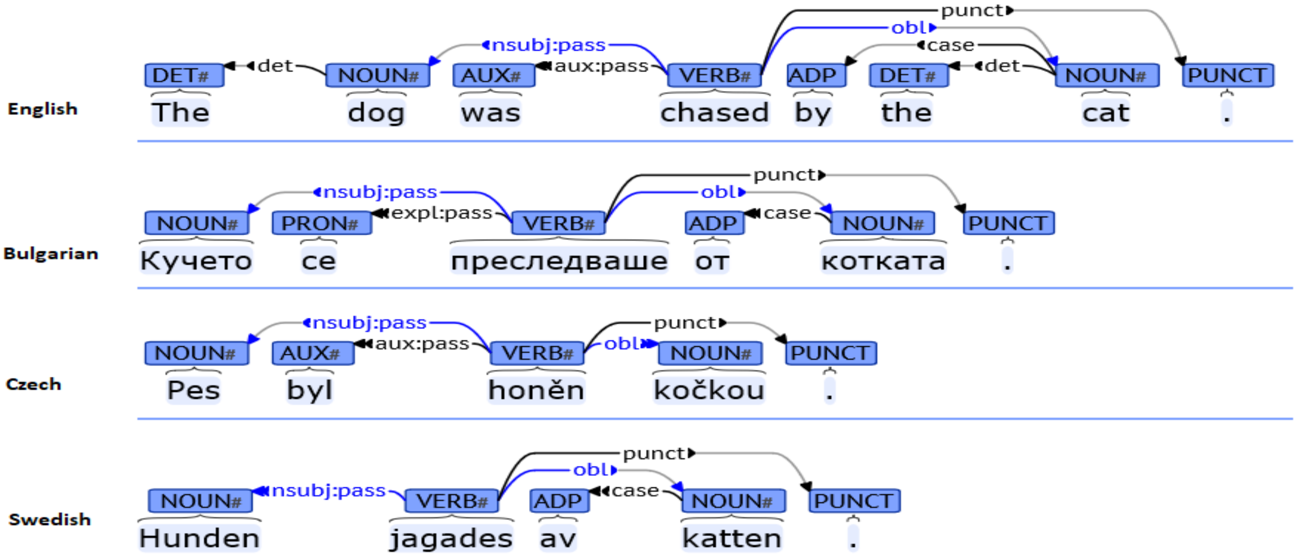
English
DET# NOUN# AUX# VERB# ADP DET# NOUN# PUNCT
The dog was chased by the cat .
det nsubj:pass aux:pass obl case det punct
Bulgarian
NOUN# PRON# VERB# ADP NOUN# PUNCT
Кучето се преследваше от котката .
nsubj:pass expl:pass obl case punct
Czech
NOUN# AUX# VERB# NOUN# PUNCT
Pes byl honěn kočkou .
nsubj:pass aux:pass obl punct
Swedish
NOUN# VERB# ADP NOUN# PUNCT
Hunden jagades av katten .
nsubj:pass obl case punct

Figure 2. Parallel samples among the languages [37].

In this study, 6 different treebanks were used, 3 of which are from Turkic language family (Turkish, Uyghur, Kazakh) and 3 from Germen language family (English, Swedish, Norwegian), which is involved in the 2.5 version and created within the scope of UDP.

The information about these used treebanks is given in Table 1. A Turkic language family is the sub-branch of the Ural-Altaic language family and Germen languages are the sub-branch of the Indo-European language family.

Table 1. Properties of The Treebanks

	Turkish	Uyghur	Kazakh	English	Swedish	Norwegian
Sentence	5635	3456	1078	16622	6026	17575
Tokens	56396	40236	10383	25489	96819	301353
Types of words	609	251	96	924	365	1302
UPOS Tags	14	16	17	17	16	17
Relation Subtypes	5	16	7	13	9	8

Turkey Turkish is from the sub-branch of Southwestern of the Turkic language family. In this study, IMST (Istanbul Technical-Middle East Technical University-Sabanci University) was used and after that, it will be named Turkish. Turkish is the renewed version of semiautomatically translated IMST Treebank [38, 39] and METU (Middle East Technical University)-Sabanci Turkish Treebank [40]. It got involved in the UDP for the first time in the 1.3 version.

Kazakh Turkish is from the sub-branch of Northwestern of Turkic language family. In the study, UD-Kazakh treebank (KTB) was used and after that, it will be named Kazakh. Kazakh was created over the taken texts from Wikipedia and news [41, 42]. It got involved in the UDP for the first time in the 1.3 version.

Uyghur Turkish is from the sub-branch of the Southeastern Turkic language family. Uyghur Dependency Treebank (UDT) is used in the study and after that it will be named Uyghur. Uyghur is UDT-based and developed in Xinjiang University located in China. It got involved in the UDP for the first time in the 1.4 version [43].

Swedish is connected to Germen family. In the study, Swedish Talbanken (ST) was used and after that, it will be named Swedish. Swedish uses the treebank developed in Lund University in 1970 as the base. The treebank has been made reusable with new explanations morphologically by Nivre and Megyesi [44]. It has been still continuing to develop within the scope of UDP as of the 1.0 version.

English belongs to Germen family. In the study, English Web Treebank (EWT) was used and after that, it will be named English. English is a dataset developed as web-based. It was taken from web-based sources like web blogs and comments. It has been still continuing to develop within the scope of UDP as of the 1.0 version [45].

Norwegian is connected to Germen family. In the study, the Nynorsk treebank was used and after that, it will be named Norwegian. Norwegian is Norwegian Dependency Treebank (NBT) based and translated to UD format automatically by Lilja Ovrelid who is from Oslo University Lilja [46]. NDT had been developed by the Text laboratory and Informatics department together in Norway national library between the years of 2011 and 2014. It has been still continuing to develop within the scope of UDP as of the 2.0 version.

UDP uses the revised version of CoNLL-X format which is called CoNLL-U. Every word is defined with10 different fields and separated with tab characters. The command line is started with # character. Sentences can be composed of one or more word lines and word lines represent the fields to be seen in the following findings. The examples of the Turkish language in the CoNLL-U format can be seen in Table 2.

Table 2. Sample sentence in Turkish (CoNLL-U)

ID	FORM	Lemma	UPOS	XPOS	Feats	Head	Deprel	Deps	Misc
#sent_id=mst-0036 #text=Nefes nefese kalmıştım.									
1	Nefes	nefes	NOUN	noun	Case=Nom|Number=Sing|Person=3	0	Root	_	_
2	nefese	nefes	NOUN	noun	Case=Dat|Number=Sing|Person=3	1	compound	_	_
3	kalmıştım	kal	VERB	verb	Aspect=Perf|Mood=Ind|Number=Sing |Person=1|Polarity=Pos|Tense=Pqp	1	compound	_	_
4	.	.	PUNCT	punct	_	1	punct	_	_

The column headings used in Table 2 can be explained as follows;

ID: Word index, integer starting at 1 for each new sentence; maybe a range for multiword tokens; may be a decimal number for empty nodes (decimal numbers can be lower than 1 but must be greater than 0).
2. FORM: Word form or punctuation symbol.
3. LEMMA: Lemma or stem of word form.
4. UPOS: Universal part-of-speech tag.
5. XPOS: Language-specific part-of-speech tag; underscore if not available.
6. FEATS: List of morphological features from the universal feature inventory or from a defined language specific extension; underscore if not available.
7. HEAD: Head of the current word, which is either a value of ID or zero (0).
8. DEPREL: Universal dependency relation to the HEAD (root iff HEAD = 0) or a defined language-specific subtype of one.
9. DEPS: Enhanced dependency graph in the form of a list of head-deprel pairs.
10. MISC: Any other annotation.

2.2. Experiments

In this part of the study, detailed information about the conducted study will be presented. The information about the created test environment, suggested pattern structure and used metrics are going to be given.

The software required for the study has been created in Visual Studio 2015 by using C# programming language. First of all, the treebank to be worked on is selected and the triplet pattern (after that it will only be named as Triplet) and frequency of the related treebank are detected automatically by the system. The same and different triplets and their numbers among the selected languages could also be found through the same software. Additionally, the similarity rates among the selected languages can be also calculated. The structure of the developed system could be seen graphically in Figure 3.

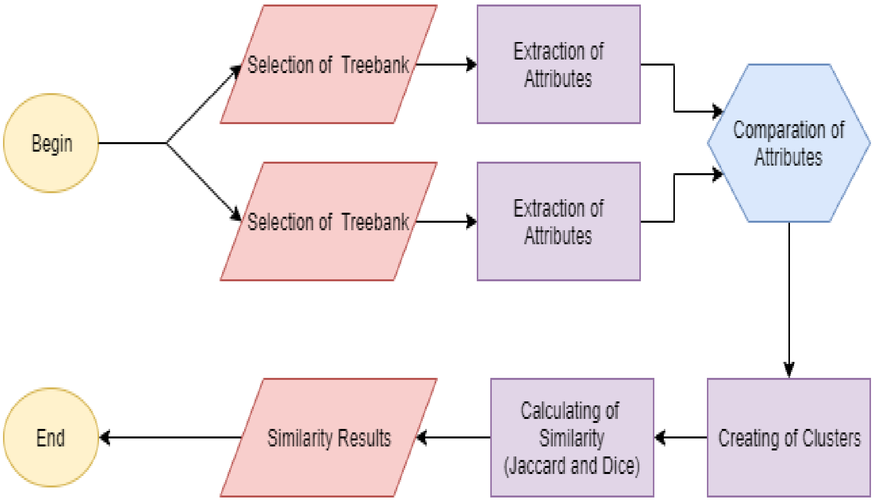

Figure 3. Devoloped System

2.2.1. Pattern Structure

Every triplet pattern is expressed as (W₁, r, W₂). W₁ refers to the first syntactic word, r to syntactic tag, and W₂ to the second syntactic word. Software that shows the relations between UPOS-DEPREL columns as triplets over 6 different treebanks used in the tests for feature extraction, is developed. As a result of this, a new feature independent from the language is tried to be defined by using word types and dependency tags instead of words. UPOS-DEPREL examples of the treebanks are given in Figure 4.

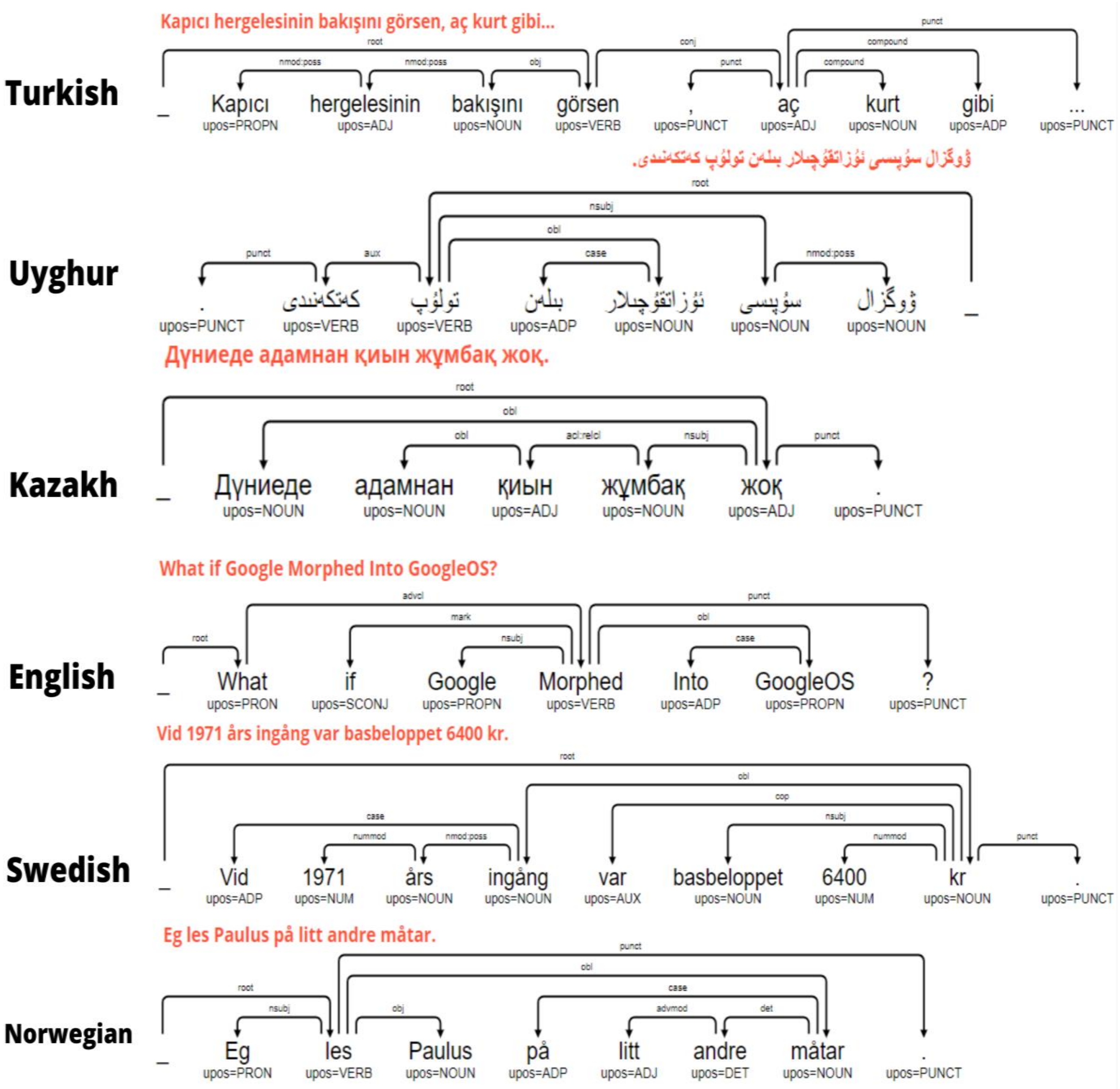

Figure 4. Samples of dependency [47].

2.2.2. Metrics

Jaccard, Dice and Simple Matching were used as similarity metrics in order to assess the study results. In equations, A and B are different languages. Jaccard Similarity is one of the main methods which is used in similarity calculations without the dictionary and evaluates the similarity statistically among sets. Its formulation is given in Equality 1.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

Dice which is also one of the similarity measurements used to measure the distance among words is in relation to Jaccard. Its formulation is given in Equality 2 and Simple Matching (SM) is given in Equality 3.

$$D(A,B) = \frac{2|A \cap B|}{|A| + |B|} \tag{2}$$

$$SM(A,B) = \frac{2|A \cap B|}{|A \cup B|} \tag{3}$$

3. Results

In this section of the study, firstly the study results of Turkic languages and then the study results of Germen languages are going to be presented. The triplet examples produced for Turkic languages by the developed system are shown in Table 3. Further, the examples of common triplets among Turkic languages could be seen in Table 4.

Table 3. Samples Triplets (for Turkic Languages)

Turkish		Kazakh		Uyghur	
Forms	Triplet	Forms	Triplet	Forms	Triplet
en sonunda	ADV,**advmod**,ADV	оңаша жатса	ADJ,**advcl**,VERB	بەش يىلدا	NUM,**nummod**,NOUN
yerinden kalkmis	NOUN,**obl**,VERB	күйеуін құшақтап	NOUN,**obj**,VERB	مېۋە بېرىدۇ	NOUN,**obj**,VERB
mesru yollarla	ADJ,**amod**,NOUN	әдемі өмірім	ADJ,**amod**,NOUN	سەۋر قىلىپ	NOUN,**compound**,VERB
dizleri yapismisti	NOUN,**nsubj**,VERB	деп ойлайтын	VERB,**advcl**,VERB	ئويلاپ دەپتۇ	VERB,**advmod**,VERB
bakisini görsen	NOUN,obj,VERB	Бала туса	NOUN,**nsubj**,VERB	كۆچىتىنى يۇلۇۋېتىپ	NOUN,**obj**,NOUN

Table 4. Common Triplets (among Turkic Languages)

Triplet	Turkish	Kazakh	Uyghur
NOUN,**nsubj**,VERB	kişi oldu	жаны сүймеген	نەشپۈت بېرىدۇ
PRON,**nsubj**,VERB	ben şaşarım	өзінің сүйгеніне	سەن تۇرالامسەن
CCONJ,**cc**,NOUN	ve delikanlı	және прокурордың	ۋە يۇلتۇزلار
NUM,**nummod**,NOUN	yedi yıl	екі жанның	بەش يىل
NOUN,**compound**,NOUN	bölge direktörü	Бас прокурордың	نەشپۈت كۆچىتىنى

The information about the frequency and triplets which have the highest and the lowest frequencies extracted for Turkic languages in the developed system could be found in Table 5. Turkish has 766, Kazakh has 360 and Uyghur has 878 different triplets.

20 triplets that are common only between two languages and have the highest frequency are presented in Table 6.

The graphic showing the frequency status of the triplets among Turkic languages is given in Figure 5. The similarity rates among Turkish languages were calculated in three different metrics by using the suggested method and are shown in Figure 6. When examining the results, it could be seen that Turkish and Uyghur have the closest similarity score. The Kazakh language has a higher similarity rate to Uyghur than the Turkish language. While the closest two languages to each other among the three languages are Turkish and Uyghur, the most distant two languages are Turkish and Kazakh.

Table 5. Triplets and Frequences for Turkic Languages

Turkish			Kazakh			Uyghur		
Index	**Triplet**	**Frequency**	**Index**	**Triplet**	**Frequency**	**Index**	**Triplet**	**Frequency**
1	PUNCT,**punct**,VERB	6399	1	PUNCT,**punct**,VERB	1145	1	PUNCT,**punct**,VERB	4289
2	NOUN,**obl**,VERB	3084	2	NOUN,**obl**,VERB	558	2	NOUN,**obl**,VERB	2311
3	NOUN,**obj**,VERB	2090	3	NOUN,**nmod:poss**,NOUN	512	3	NOUN,**nsubj**,VERB	2030
4	PUNCT,**punct**,NOUN	2014	4	NOUN,**nsubj**,VERB	473	4	NOUN,**obj**,VERB	1664
5	ADJ,**amod**,NOUN	1842	5	ADJ,**amod**,NOUN	468	5	PUNCT,**punct**,NOUN	1533
6	NOUN,**nmod:poss**,NOUN	1701	6	NOUN,**obj**,VERB	430	6	VERB,**advcl**,VERB	1307
7	NOUN,**nmod**,NOUN	1385	7	PUNCT,**punct**,NOUN	429	7	NOUN,**amod**,NOUN	1208
8	NOUN,**nsubj**,VERB	1329	8	VERB,**advcl**,VERB	283	8	PUNCT,**punct**,AUX	955
9	ADJ,**amod**,VERB	1075	9	PUNCT,**punct**,ADJ	219	9	NOUN,**nmod:poss**,NOUN	908
10	ADV,**advmod**,VERB	1058	10	DET,**det**,NOUN	211	10	ADJ,**amod**,NOUN	850
...	...	...	...	...	...	...	...	...
757	PROPN,**obl**,ADP	1	351	NUM,**parataxis**,NOUN	1	869	VERB,**punct**,ADJ	1
758	NOUN,**parataxis**,NOUN	1	352	NOUN,**parataxis**,NUM	1	870	NUM,**punct**,NOUN	1
759	ADV,**parataxis**,VERB	1	353	VERB,**parataxis**,NUM	1	871	ADJ,**punct**,VERB	1
760	NOUN,**parataxis**,VERB	1	354	PROPN,**parataxis**,PROPN	1	872	VERB,**punct**,VERB	1
761	VERB,**punct**,ADJ	1	355	PRON,**parataxis**,VERB	1	873	VERB,**orphan**,VERB	1
762	CCONJ,**punct**,ADP	1	356	NOUN,**vocative**,ADJ	1	874	ADJ,**vocative**,ADJ	1
763	CCONJ,**punct**,VERB	1	357	PROPN,**vocative**,ADV	1	875	ADJ,**vocative**,PART	1
764	NOUN,**punct**,VERB	1	358	PROPN,**vocative**,NOUN	1	876	NOUN,**vocative**,PRON	1
765	NUM,**punct**,VERB	1	359	NOUN,**veocative**,PRON	1	877	VERB,**xcomp**,AUX	1
766	PROPN,**punct**,VERB	1	360	PRON,**xcomp**,VERB	1	878	VERB,**xcomp**,NOUN	1

Table 6. Common only Triplets (between two languages) and frequencies

Kazak and Uyghur			**Kazak and Turkish**			**Uyghur and Turkish**		
Index	**Triplet**	**Frequency**	**Index**	**Triplet**	**Frequency**	**Index**	**Triplet**	**Frequency**
1	NOUN,**amod**,NOUN	1208	1	PROPN,**conj**,PROPN	90	1	PRON,**det**,NOUN	566
2	NOUN,**advmod**,VERB	723	2	NOUN,**compound**,PROPN	73	2	VERB,**compound_lvc**,VERB	214
3	AUX,**aux**,VERB	486	3	PRON,**obl**,ADJ	72	3	VERB,**obj**,VERB	184
4	ADJ,**advmod**,VERB	346	4	PROPN,**compoud**,PROPN	53	4	ADJ,**compound**,VERB	143
5	VERB,**aux**,VERB	250	5	CCONJ,**cc**,PROPN	42	5	NOUN,**compound:redup**,NOUN	120
6	NOUN,**advcl**,VERB	98	6	PROPN,**nsubj**,ADJ	32	6	VERB,**nmod**,NOUN	97
7	NOUN,**advmod**,NOUN	91	7	NOUN,**obl**,NUM	31	7	VERB,**compound**,VERB	77
8	ADJ,**obl**,VERB	61	8	ADJ,**amod**,NUM	28	8	VERB,**nsubj**,VERB	74
9	VERB,**discourse**,VERB	50	9	ADJ,**compound**,NUM	28	9	PUNCT,**punct**,CCONJ	65
10	ADV,**cc**,VERB	49	10	VERB,**conj**,PRON	26	10	ADP,**advmod**,VERB	47
11	NOUN,**acl**,NOUN	44	11	PROPN,**conj**,NOUN	23	11	NOUN,**case**,VERB	45
12	SCONJ,**cc**,VERB	44	12	PROPN,**obl**,ADJ	22	12	NOUN,**discourse**,VERB	44
13	PRON,**advmod**,VERB	43	13	ADJ,**conj**,PROPN	17	13	NOUN,**flat**,NOUN	44
14	NOUN,**obl**,NOUN	37	14	PRON,**nsubj**,PRON	14	14	VERB,**nmod:poss**,NOUN	36
15	NOUN,**nummod**,NOUN	36	15	PROPN,**nmod**,ADJ	13	15	VERB,**compound**,NOUN	34
16	NOUN,**vocative**,VERB	36	16	ADV,**advmod**,PROPN	12	16	VERB,**acl**,VERB	25
17	PRON,**advmod**,ADJ	33	17	DET,**nsubj**,NOUN	12	17	ADJ,**compound:redup**,ADJ	21
18	AUX,**aux**,NOUN	30	18	AUX,**cop**,PROPN	11	18	NUM,**nummod**,VERB	21
19	NUM,**advmod**,VERB	25	19	CCONJ,**cc**,PRON	10	19	ADJ,**compound:redup**,NOUN	20
20	ADV,**amod**,NOUN	20	20	ADJ,**nsubj**,PRON	10	20	NOUN,**cop**,VERB	20

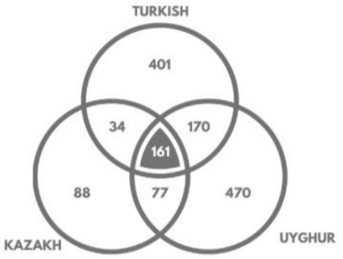

Figure 5. The frequency of the triplets among Turkic Languages

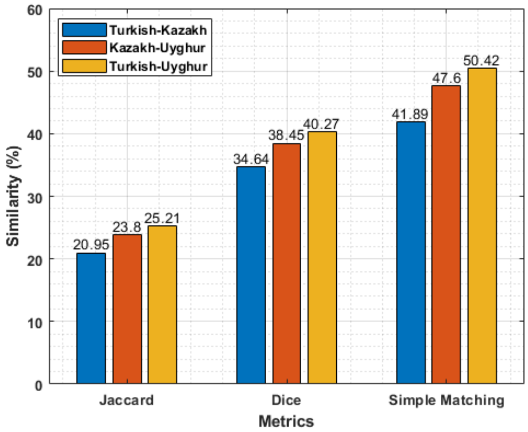

Figure 6. The results of Turkic Languages

The study results of Germen languages are going to be given in this section of the study. In this part, similar studies conducted for Turkic languages were repeated for Germen languages. At the end of the study, Swedish has formed 702 triplets, English has formed 1258 triplets and Norwegian has formed 915 triplets. The graphic showing the frequency status of the triplets among Germen languages is given in Figure 7. The similarity rates among Germen languages were calculated in three different metrics by using the suggested method and are shown in Figure 8.

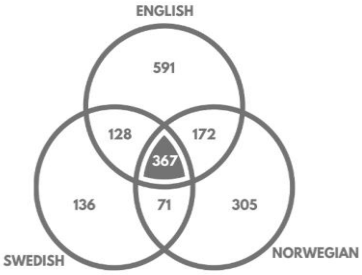

Figure 7. The frequency of the triplets among Germen Languages

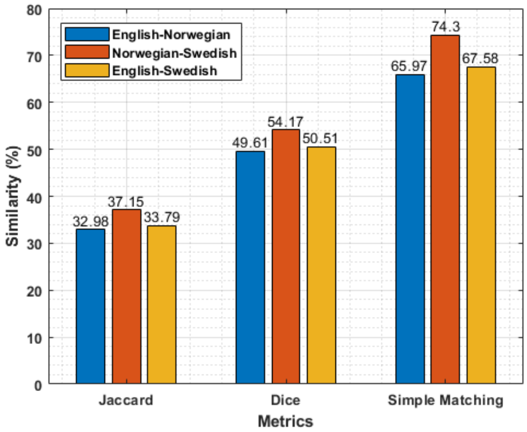

Figure 8. The results of Germen Languages

4.Discussion and Conclusion

New feature extraction is suggested within the scope of the study in order to find out the similarities among languages. The feature extraction suggested in order to find the similarities of 3 different languages from 2 different language families, was tested through software developed within this study. Based on the features obtained, not only the similarity of languages was calculated but also acquired patterns were analyzed. The following findings have been found for Turkic languages;

- The number of triplets common among three languages is 161 and the triplet with the highest frequency is PUNCT, punct, VERB with 4289. It means that a verb is connected to punctuation mark with a punct tag. The second triplet with the highest frequency is NOUN, obl, VERB with 2311. It means that a noun is connected to a verb with an obl tag. The third one with the highest frequency is NOUN, nsubj, VERB with 2030 and it means that a noun is connected to a verb by nsubj tag.
- The number of triplets common only between Turkish and Kazakh is 34 and the triplet with the highest frequency is PROPN, conj, PROPN with 90. It means that a proper noun is linked to a proper noun by a conj tag. The second triplet with the highest frequency is NOUN, compound, PROPN with 73 and it means that a noun is connected to a proper noun by compound tag. The third one with the highest frequency is PRON, obl, ADJ with 72. This refers that a pronoun is linked to an adjective by an obl tag.
- The number of triplet common only between Turkish and Uyghur is 170 and the triplet with the highest frequency is PRON, det, NOUN with 566. Meaning that a noun is linked to a noun by an amod tag. The second triplet with the highest frequency is VERB, compound_lvc, VERB with 214. It indicates that a verb is connected to a verb by compound_lvc tag. The third one with the highest frequency is VERB, obj, VERB with

184 and it means that a verb is connected to a verb by obj tag.

- The number of triplets common only between Kazakh and Uyghur is 77 and the triplet with the highest frequency is NOUN, amod, NOUN with 1208. It means that a pronoun is linked to a noun by a det tag. The second triplet with the highest frequency is NOUN, advmod, VERB with 723 and it means that a noun is connected to a verb by advmod tag. The third one with the highest frequency is AUX, aux, VERB with 486. It refers that an auxiliary verb being linked to a verb with an aux tag.
- There are only 401 triplets belonging to Turkish, 88 to Kazakh, and 470 to Uyghur.
- In the 766 different triplets belonging to Turkish, the triplet with the highest frequency is PUNCT, punct, VERB one with 6399, the second one is NOUN, obl, VERB with 3084 and the third one is NOUN, obj, VERB with 2090. Those values show that the dependencies in Turkish are generally between a punctuation mark or noun and a verb.
- In the 360 triplets belonging to Kazakh, the triplet with the highest frequency is PUNCT, punct, VERB one with 1145, the second one is NOUN, obl, VERB with 558 and the third one is NOUN, nmod:poss, NOUN with 512. Those values show that the dependencies in Kazakh are generally between a punctuation mark or noun and a verb or noun.
- In the 878 triplets belonging to Uyghur, the triplet with the highest frequency is PUNCT, punct, VERB one with 4289, the second one is NOUN, obl, VERB with 2311 and the third one is NOUN, nsubj, VERB with 2030. Those values show that the dependencies in Uyghur are generally between a punctuation mark or noun and a verb.
- It is found that the first two triplets with the highest frequency among the three languages are common in all languages. A similar situation is 7 for the first 10 triplets with the highest frequency. Even though the rankings for language change with regard to the frequency values, 7 of the first 10 triplets are common.
- The number of triplets passed only once within the treebank is 116 in Kazakh, 335 in Uyghur, and 221 in Turkish. As a result of the study conducted for the Turkic languages, it might be argued that the structure of Turkish is closer to Uyghur. Additionally, the structural similarity between Kazakh and Uyghur is more than Turkish.

The following findings have been found for Germen languages;

- 1258 triplets for English, 702 for Swedish, and 915 for Norwegian have been found.

Declaration of Ethical Code

In this study, we undertake that all the rules required to be followed within the scope of the "Higher Education Institutions Scientific Research and Publication Ethics Directive" are complied with, and that none of the actions stated under the heading "Actions Against Scientific Research and Publication Ethics" are not carried out.

References

[1] Searle, J.R. 1975. Indirect speech acts. ss. 59-82. Speech Act, Brill Press, New York, USA, 406s.

[2] Taylor, P.J., Thomas S. 2008. Linguistic style matching and negotiation outcome. Negotiation and Conflict Management Research, 1(3), 263-281.

[3] Pennebaker, J.W., Stone, L.D. 2003. Words of wisdom: language use over the life span. Journal of personality and social psychology, 85(2), 291.

[4] Groom, C.J., Pennebaker, J.W. 2005. The language of love: Sex, sexual orientation, and language use in online personal advertisements. Sex Roles, 52, 447–461.

[5] Laserna, C.M., Seih, Y.T., Pennebaker, J.W. 2014. Um... who like says you know filler word use as a function of age, gender, and personality. Journal of Language and Social Psychology, 33(3), 328-338.

[6] Dehghani, M., Sagae, K., Sachdeva, S., Gratch, J. 2014. Analyzing political rhetoric in conservative and liberal weblogs related to the construction of the ground zero mosque. Journal of Information Technology Politics, 11(1), 1–14.

[7] Skoyen, J.A., Randall, A.K., Mehl, M.R., Butler, E.A. 2014. We overeat, but i can stay thin: Pronoun use and body weight in couples who eat to regulate emotion. Journal of Social and Clinical Psychology, 33(8), 743.

[8] Berwick, R.C., Friederici, A.D., Chomsky, N., Bolhuis, J.J. 2013. Evolution, brain, and the nature of language. Trends in cognitive sciences, 17(2), 89–98.

[9] Miceli-Barone, A.V., Attardi, G. 2015. Non-projective dependency-based pre-reordering with recurrent neural network for machine translation. Proceedings of the Ninth Workshop on Syntax, Semantics and Structure in Statistical Translation, July 26-31, Beijing, China, 846-856.

[10] Xiao, T., Zhu, J., Zhang, C., Liu, T. 2016. Syntactic skeleton-based translation. the thirtieth AAAI conference on artificial intelligence, February 12-17, Phoenix, Arizona, USA, 2856-2862.

[11] Song, M., Kim, W.C., Lee, D., Heo, G.E., Kang, K.Y. 2015. PKDE4J: entity and relation extraction for public knowledge discover. Journal of Biomedical Informatics, 57, 320–332.

[12] Yu, M., Gormley, M.R., Dredze, M. 2015. Combining word embeddings and feature embeddings for fine-grained relation extraction. 2015 Conference of the North American chapter of the Association for Computational Linguistics: human language technologies, May 31 – June 5, Denver, Colorado, USA, 1374-1379.

[13] Pado´ S, S., Noh, T.G., Stern, A., Wang, R., Zanoli, R. 2015. Design and realization of a modular architecture for textual entailment. Natural Language Engineering, 21(2), 67-200.

[14] Joshi, M., Penstein-Ros´e, C. 2009. Generalizing dependency features for opinion mining. Association for Computational Linguistics, 4 August, Suntec, Singapore, 313-316.

[15] Vilares, D., Alonso, M.A., Gomez-Rodriguez, C. 2015. A linguistic approach for determining the topics of Spanish Twitter messages. Journal of Information Science, 41(02), 127–145.

[16] Vilares, D., Alonso, M.A., Gomez-Rodriguez, C. 2015. A syntactic approach for opinion mining on Spanish reviews. Natural Language Engineering, 21(01), 139–163.

[17] Errecalde, M.L., Cagnina, L.C., Rosso, P. 2016. Silhouette + attraction: A simple and effective method for text clustering. Natural Language Engineering 22(5), 687-726.

[18] Potthast, M., Barron-Cedeno, A., Stein, B., Rosso, P. 2011. Cross-language plagiarism detection. Language Resources and Evaluation, 45(1), 45-62.

[19] Crossley, S.A., McNamara, D.S. 2017. Adaptive educational technologies for literacy instruction. Routledge, New York, 310s.

[20] Kyle, K. 2016. Measuring syntactic development in l2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication, Georgia State University, College of Arts and Sciences, PhD Dissertation, 201s.

[21] Niederhoffer, K.G., Pennebaker, J.W. 2002. Linguistic style matching in social interaction. Journal of Language and Social Psychology, 21(4), 337-360.

[22] Ireland, M.E., Pennebaker, J.W. 2010. Language style matching in writing: synchrony in essays, correspondence, and poetry. Journal of personality and social psychology, 99(3), 549-571.

[23] Pennebaker, J.W., Francis, M.E., Booth, R.J. 2001. Linguistic inquiry and word count: Liwc 2001, Mahway: Lawrence Erlbaum Associates, 71.

[24] Gamallo, P., Gasperin, C., Agustini, A., Lopes, G.P. 2001. Syntactic-based methods for measuring word similarity. International Conference on Text, Speech and Dialogue, 11-13 September, Berlin, Germany, 116-125.

[25] Li, S., Zhou, L., Zhang, J., Zhou, F., Guo, J., Huo, W. 2018. Chinese-Lao Cross-Language Test Similarity Computing Based on WordNet. International Conference on Mechatronics and Intelligent Robotics, 19-20 May, Kunming, China, 459-464

[26] Dinh, D., Thanh, N.L. 2019. English–Vietnamese cross-language paraphrase identification using hybrid feature classes. Journal of Heuristics, 1-17.

[27] Steinberger, R., Pouliquen, Hagman, J. 2002. Cross-lingual document similarity calculation using the multilingual thesaurus EUROVOC. 3rd Conference on Computational Linguistics and Intelligent Text Processing, 17-23 February, Mexico City, Mexico, 415-424.

[28] Barron-Cedeno, A., Rosso, P., Pinto, D., Juan, A. 2008. On cross-lingual plagiarism analysis using a statistical model. International Workshop on Uncovering Plagiarism, 22 July, Patras, Greece, 9-14.

[29] Uszkoreit, J., Ponte, J.M., Popat, A.C., Dubiner, M. 2010. Large scale parallel document mining for machine translation. 23rd International Conference on Computational Linguistics, 23-27 August, Beijing, China, 1101-1109.

[30] Maike, E., Andrew, F., Kotaro, N. 2011. Calculating wikipedia article similarity using machine translation evaluation metrics. International Conference on Advanced Information Networking and Applications, 22-25 March, Biopolis, Singapore, 620-625.

[31] De Marneffe, M.C., MacCartney, B., Manning, C.D. 2006. Generating typed dependency parses from phrase structure parses. the Fifth International Conference on Language Resources and Evaluation, 24-26 May , Genoa, Italy, 449-454.

[32] De Marneffe, M.C., Manning, C.D. 2008. The Stanford typed dependencies representation. Cross-Framework and Cross-Domain Parser Evaluation, 23 August , Manchester, United Kingdom, 1-8.

[33] De Marneffe, M.C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, Manning, J.V. 2014. Universal Stanford Dependencies: A cross-linguistic typology. Ninth International Conference on Language Resources and Evaluation, 26-31 May, Reykjavik, Iceland, 4585-4592.

[34] Petrov, S., Das, D., McDonald, R. 2012. A universal part-of-speech tagset. The Eight

International Conference on Language Resources and Evaluation, 21-27 May, Istanbul, Turkey, 2089-2096.

[35] Zeman, D. 2008. Reusable Tagset Conversion Using Tagset Drivers. The Sixth International Conference on Language Resources and Evaluation, 28-30 May, Marrakech, Morocco, 213-218.

[36] Zeman, D., Nivre, J., Abrams, M. 2022. Universal Dependencies 2.5. http://hdl.handle.net/11234/1-3105. (Accessed 25 June 2022.)

[37] UD Project 2019. Universal Dependency Project. https://universaldependencies.org/introduction.html. (Accessed 25 June 2022).

[38] Sulubacak, U., Eryiğit, G. 2018. Implementing Universal Dependency, Morphology and Multiword Expression Annotation Standards for Turkish Language Processing. Turkish Journal of Electrical Engineering & Computer Sciences, 26(3), 1662-1672.

[39] Sulubacak, U., Gökırmak, M., Tyers, F., Çöltekin, Ç., Nivre, J., Eryiğit, G. 2016. Universal Dependencies for Turkish. The 26th International Conference on Computational Linguistics, 11-16 December, Osaka, Japan, 3444-3454.

[40] Oflazer, K., Say, B., Hakkani-Tür, D.Z., Tür, G. 2003. Building a Turkish Treebank. Ss 261-277. Tree-banks: Building and Using Parsed Corpora, Academic Publishers, Dordrecht, 407s.

[41] Tyers, F.M., Washington, J. 2015. Towards a free/open-source universal-dependency treebank for Kazakh. The 3rd International Conference on Turkic Languages Processing, 8-10 April, Kazan, Tataristan, Russia, 108-120.

[42] Makazhanov, A., Sultangazina, A., Makhambetov, O., Yessenbayev, Z. 2015. Syntactic annotation of kazakh: Following the universal dependencies guidelines. The 3rd International Conference on Turkic Languages Processing, 8-10 April, Kazan, Tatarstan, Russia, 338-350.

[43] Mushajiang, W, Yibulayin, T., Abiderexiti, K., Liu, Y. 2016. Universal dependencies for Uyghur. The Third International Workshop on Worldwide Language Service Infra-structure and Second Workshop on Open Infrastructures and Analysis Frameworks for Human Language Technologies, 12 December, Osaka, Japan, 44-50.

[44] Nivre, J., Megyesi, B. 2007. Bootstrapping a Swedish treebank using cross-corpus harmonization and annotation projection. The 6th International Workshop on Treebanks and Linguistic Theories, 7-8 December, Bergen, Norway, 97-102.

[45] Silveira, N., Dozat, T., De Marneffe, M.C., Bowman, S., Connor, M., Bauer, J., Manning, C. 2014. A Gold Standard Dependency Corpus for English. The Ninth International Conference on Language Resources and Evaluation, 26-31 May, Reykjavik, Iceland, 2897-2904.

[46] Velldal, E., Ovrelid, L., Hohle, P. 2017. Joint UD Parsing of Norwegian Bokmal and Nynorsk. The 21st Nordic Conference on Computational Linguistics, 22-24 May, Gothenburg, Sweden, 1-10.

[47] Dependency Graph 2019. Grew-Match. http://match.grew.fr/ (Accessed 25 June 2022).