Comparative analysis of neural network-based morphological parsers of Korean language

Alisa Mosolova
RSUH
Moscow, Russia
alisa-mosolova@mail.ru

Abstract
This study analyzes the work of the UDPipe and Stanza neural parsers' models trained on corpora of Universal Dependencies project, the purpose of which is to develop a universal morphological markup for the corpora of various languages. The paper considers the applicability of the universal tags' set to the part-of-speech system of Korean, and compares the UD, KAIST and Sejong tagsets used by the parsers for Korean. The part-of-speech classification proposed in these tagsets is analyzed from the point of view of Korean grammar. The paper also evaluates the performance of the UDPipe and Stanza parsers.
Keywords: Korean, Universal Dependencies, morphological tagger, parts of speech

Сопоставительный анализ морфопарсеров корейского языка на нейронных сетях

Мосолова Алиса Олеговна
РГГУ
Москва, Россия
alisa-mosolova@mail.ru

Аннотация
В данном исследовании анализируется работа моделей нейронных парсеров UDPipe и Stanza, обученных на корпусах проекта Universal Dependencies, целью которого является разработка универсальной морфоразметки для корпусов различных языков. В статье рассматривается применимость набора универсальных тегов к системе частей речи корейского языка, а также сравниваются тегсеты UD, KAIST и Sejong, используемые парсерами для корейского языка. Частеречное деление, предлагаемое в этих наборах тегов, анализируется с точки зрения корейской грамматики. В работе также оценивается качество работы парсеров UDPipe и Stanza.
Ключевые слова: корейский язык, Universal Dependencies, морфологический парсер, части речи.

1. Введение

Морфологический парсер — программа, автоматически осуществляющая морфоразметку. Прежде чем приступать к собственно морфоразметке, парсер токенизирует текст, то есть разбивает его на токены — единицы морфоразметки, чаще всего соответствующие словоформе. Следующим этапом морфоразметки является приписывание токенам частеречных тегов, а также приведение токена к лемме, то есть начальной форме слова. В случае с морфоразметкой корейского текста токен не приводится к лемме, а разбивается на морфемы.

В ходе исследования анализировалась работа двух нейронных парсеров UDPipe и Stanza. Эти модели были обучены на корпусах, созданных в рамках проекта Universal Dependencies, целью которого является разработка универсальной морфоразметки для корпусов различных языков [de Marneffe et al 2006]. Нейронные парсеры были выбраны как

наиболее современный инструмент в области автоматического анализа текста [Straka 2018]. В работе рассматривается применимость набора универсальных тегов к частеречной системе корейского языка, а также оценивается качество работы парсеров UDPipe и Stanza для корейского языка.

В разделе 2 описываются корпуса, на которых обучались модели парсера; в разделе 3 рассматриваются части речи корейского языка, которые затруднительно описать с помощью универсального тегсета; в разделе 4 даётся оценка качества парсеров; в разделе 5 отдельно рассматривается качество разметки имён собственных.

2. Корпуса корейского языка в UD

В рамках проекта Universal Dependencies для корейского языка было создано несколько корпусов, в том числе UD Korean GSD и UD Korean KAIST. Объём UD Korean GSD — 80 322 токена, UD Korean KAIST — 350 090 токенов. В UD Korean GSD вошли новостные тексты и записи из веб-блогов, в UD Korean KAIST — художественные и новостные тексты, а также научные статьи. Процесс разработки этих корпусов описывается в [Chun et al. 2018]. Модели UDPipe KAIST и Stanza GSD, обученные на этих корпусах, при разметке проставляют для каждого токена два типа частеречных тегов: UPOS и XPOS. Основное различие между тегами UPOS и XPOS для корейского в том, что считается единицей разметки: для тегов UPOS такими единицами являются ограниченные пробелами оджоли, а для XPOS — морфемы. То есть одному токену приписывается только один UPOS-тег и один или несколько XPOS-тегов. В таблице 1 приводится пример разбора модели UDPipe KAIST:

Таблица 1

	токен	лемма	UPOS-тег	XPOS-тег
разбор морфопарсера	고양이를	고양이+을	NOUN	ncn+jco
транслитерация	koyangilul	koyangi+lul		
глоссы	‘кошку’	кошка ACC		

XPOS-теги корректнее называть не частеречными, а морфемными: тег ncn приписывается корню существительного, а jco — частице винительного падежа.

Корпус UD Korean GSD был получен в результате автоматического преобразования корпуса Google UD Korean Treebank [McDonald et al. 2013] в соответствии со стандартами обновленных правил разметки в формате UD. Кроме того, разметка была дополнена с помощью морфопарсера KOMA [Lee and Rim 2009]. KOMA использует тегсет проекта 21st Century Sejong [Kang, Kim 2004, Kim 2006]. Задачей этого проекта, запущенного в 1998 году, было создание корейского корпуса, сопоставимого с Британским национальным корпусом. Морфопарсер, обученный на UD Korean GSD, проставляет для каждого токена в качестве XPOS-тегов теги проекта 21st Century Sejong (далее — теги Sejong). В тегсете содержится 45 тегов [Choi 2013].

Корпус UD Korean KAIST — результат конвертации в соответствии с принципами разметки UD корпуса KAIST, который разрабатывался с 1992 года [Choi et al. 1994]. В процессе конвертации несколько тегов KAIST для каждого оджоля заменялись на один тег UD. Морфопарсер, обученный на UD Korean KAIST, проставляет для каждого токена в качестве XPOS-тегов теги проекта KAIST (далее — теги KAIST). В тегсете содержится 54 тега [Choi 2013].

Соответствие тегов, за исключением относящихся к символам и пунктуации, приведено в приложении 4. Также в таблице приведены примеры словоформ и морфем, которые размечаются соответствующими тегами, и соотношение тегов с частями речи по грамматике корейского языка [Martin 1992].

3. Применимость тегсетов к частеречной системе корейского языка

Тегсеты UD, KAIST и Sejong были проанализированы с точки зрения соответствия частеречной системе корейского языка, предложенной в грамматике [Martin 1992]. Схема основных частей речи по [Martin 1992] приводится в приложении 1. Было выявлено три класса частей речи, для которых сложно подобрать подходящий тег: атрибутивы, бытийные и локативные связки.

3.1. Бытийные и локативные связки

В корейском языке класс предикативов делится на две больших группы: глаголы и прилагательные. Различия между ними проявляются в ограничениях на образование некоторых форм и при выборе алломорфов некоторых суффиксов, например, образующих формы косвенной речи (см. таблицу 2). На основании парадигм предикативов можно также выделить ещё две группы: бытийные связки и локативные, которые в грамматике [Martin 1992] называются квазиглаголами.

Таблица 2

Наклонение	Бытийные связки	Прилагательные	Локативные связки	Глаголы
Изъявительное	*-la*	*-ta*		*-nunta*/*-nta*
Вопросительное	*-unya*/*-nya*		*-nunya*	
Императив				*-ula*/*-la*
Гортатив				*-ca*

Таблица 2 демонстрирует, что бытийные и локативные связки морфологически отличаются и от глаголов, и от прилагательных, поэтому при отсутствии специальных тегов возникает проблема с отнесением связок в ту или иную категорию. Рассмотрим, как парсеры помечают каждую из связок:

Таблица 3

		Stanza GSD		UDPipe KAIST	
		теги Sejong	теги UD	теги UD	теги KAIST
бытийная связка		VCP	VERB	VERB	jp
отрицательная бытийная связка	финитная форма	VCN	ADJ	ADJ	paa
	нефинитная форма		VERB		
локативная связка	финитная форма	VV	ADJ	ADJ	paa
	нефинитная форма		VERB		
отрицательная локативная связка	финитная форма	VA	ADJ	ADJ	paa
	нефинитная форма		VERB		

Наименее удачно оказываются размечены тегами UD отрицательные связки при использовании модели Stanza GSD: одна и та же лексема в финитной и нефинитной форме размечается тегами разных частей речи.

В разметке UD для других языков бытийная связка обычно обозначается тегом AUX (вспомогательный глагол), поэтому выбор других тегов представляется нецелесообразным с учётом задач проекта:

Таблица 4

	лемма	частеречный тег
русский	быть	AUX
английский	be	AUX
нидерландский	zijn	AUX
литовский	būti	AUX
арабский	kaana	AUX

3.2. Атрибутивы

Атрибутивы в корейской грамматике – существительные, которые могут выступать исключительно или в основном в приименной позиции: ***yeys*** *chinkwu* ‘**старый** друг’
Лексема *yeys* — пример качественного атрибутива, кроме них можно также выделить дейктические атрибутивы, такие как *i* ‘этот’ и *ku* ‘тот’. В таблице 5 приводится соответствие тегов для этих двух групп атрибутивов:

Таблица 5

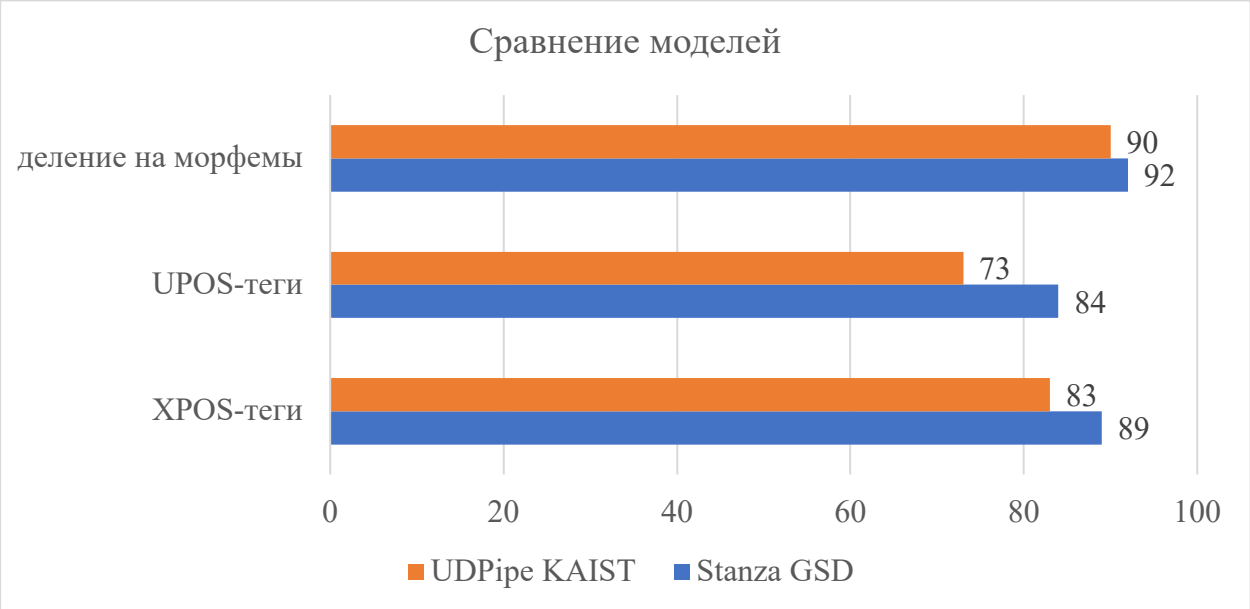
тип атрибутива	UD	KAIST	Sejong
качественный атрибутив	ADJ	mma	MM
дейктический атрибутив	DET	mmd	

В тегсете UD для класса атрибутивов не предусмотрено специального тега, в результате чего атрибутивные демонстративы размечаются как детерминативы (тег DET), а качественные атрибутивы, не отличающиеся от них морфологически, размечаются как прилагательные (ADJ), хотя прилагательные в корейском сильно отличаются от качественных атрибутивов с точки зрения и морфологии, и синтаксиса. Таким образом, набор UPOS-тегов не позволяет корректно отразить особенности этой части речи.

4. Оценка качества работы парсеров

Для того, чтобы оценить качество работы парсера, был выбран отрывок из 456 оджолей из текста «Счастье, о котором мечтает одна “везучая девушка”» («“Un cohun” celmunika kkwumkkwunun hayngpok»), опубликованного в журнале Корейского фонда «Koreana». Модель Stanza GSD выделила в тексте 551 токен, а модель UDPipe KAIST — 510. В отдельные токены выделялись не только оджоли, но и знаки препинания, кроме того, Stanza GSD иногда выделяет как токены пробелы. Далее будут анализироваться разборы только тех токенов, которые были выделены корректно: 537 для модели Stanza GSD и 483 для модели UDPipe KAIST.

Диаграмма 1



На диаграмме 1 видно, что модель Stanza GSD справилась со всеми задачами лучше, чем UDPipe KAIST, а самой сложной задачей для обеих моделей оказалось выставление UPOS-тегов. Далее будут подробнее разобраны ошибки, которые допускала каждая из моделей при выставлении этих тегов.

4.2. Stanza GSD

Основные типы ошибок при выставлении UPOS-тегов с примерами представлены в таблице:

Таблица 6

Тип ошибки	Токен	Правильно	Ошибка	Кол-во ошибок	Комментарий
вспомогательный предикатив размечен как глагол или прилагательное	않았다	AUX	VERB	15	Ни один из встретившихся в тексте вспомогательных глаголов не был разобран верно.
	anh-ass-ta	не_делать-PST-FIN	‘не делал’		
прилагательное размечено как глагол	많아	ADJ	VERB	4	На выборке из 100 прилагательных парсер показал качество в 68%.
	manh-a	много-INF	‘много’		
имя собственное размечено как имя нарицательное	서울	PROPN	NOUN	17	См. раздел 5
	sewul	Сеул	‘Сеул’		
имя (NOUN/ PROPN/PRON) размечено как наречие	카페에서	NOUN	ADV	36	Из 82 токенов, состоящих из имени и частицы адвербиального падежа и/или выделительной частицы, верно размечено 56%.
	khaph eyse	кафе LOC	‘в кафе’		
другое				14	
Всего ошибок		86	Верно размечено		84%

4.3. UDPipe KAIST

Корпус UD Korean KAIST был получен в результате конвертации корпуса KAIST. Для того, чтобы провести конвертацию, нужно было прописать соответствия тегов KAIST и UD, а также установить, какой из нескольких тегов, приписанных оджолю в корпусе имеет более высокий приоритет при принятии решения. Например, тегу px соответствует тег AUX (пример 1), а тегам частиц чоса — тег ADP (пример 3). Однако, тегом px может размечаться второй корень сложного глагола (пример 2), соответственно, в сочетании с тегом pvg он должен давать UPOS-тег VERB. Аналогично в примере 4 приоритет тега ncn выше, чем jco, поэтому оджолю приписывается UPOS-тег NOUN.

Таблица 7

№	токен	лемма	XPOS-теги	UPOS-тег
1	않다	않+다	px+ef	AUX
	anhta	anh+ta		
	‘не делать’	не_делать-FIN		
2	찾아가다	찾+아+가+다	pvg+ecx+px+ef	VERB
	chacakata	chac+a+ka+ta		
	‘навещать’	искать-INF-идти-FIN		
3	를	를	jco	ADP
	lul	lul		
		ACC		
4	고양이를	고양이+을	ncn+jco	NOUN
	koyangilul	koyangi+lul		
	‘кошку’	кошка ACC		

Конвертация была проведена с ошибками: некоторым тегам был приписан слишком высокий приоритет, а для некоторых тегов или их комбинаций были неверно установлены соответствия. В таблице 8 приведены примеры выявленных ошибок — в первую очередь их вызвала попытка определять UPOS-тег через XPOS-теги аффиксов, а не корней. Таблица с другими выявленными ошибками приводится в приложении 2.

Таблица 8

XPOS-теги	Неверное соответствие UPOS-тегу	Верное соответствие UPOS-тегу	Пример слова	
			слово	XPOS-теги
xp	PART	тег для приставки не может определять UPOS-тег	헛소리	xp+ncn
			hes-soli	ложный-звук
			‘чушь’	
именной тег +jca	ADV	соответствующий именной тег (NOUN/ PROPN/PRON/NUM)	서울에서	nq+jca
			sewul eyse	Сеул LOC
			‘в Сеуле’	
тег предикатива +ecc	CCONJ	соответствующий тег предикатива (ADJ/VERB/AUX)	하고	pvg+ecc
			ha-ko	делать-CONV
			‘делая’	

Всего модель допустила 130 ошибок при выставлении UPOS-тегов, соответственно, верно теги были проставлены для 73% токенов. Так как все UPOS-теги определяются последовательностью XPOS-тегов, все ошибки модели UDPipe KAIST вызваны или ошибкой при постановке XPOS-тега, или ошибкой при конвертации корпуса, на котором модель училась. Из 130 ошибок ко второму типу относится лишь 26. Среди них есть те же типы ошибок, что и в разборе Stanza GSD: например, UDPipe KAIST так же не всегда верно определяет имена собственные и иногда размечает прилагательные как глаголы.

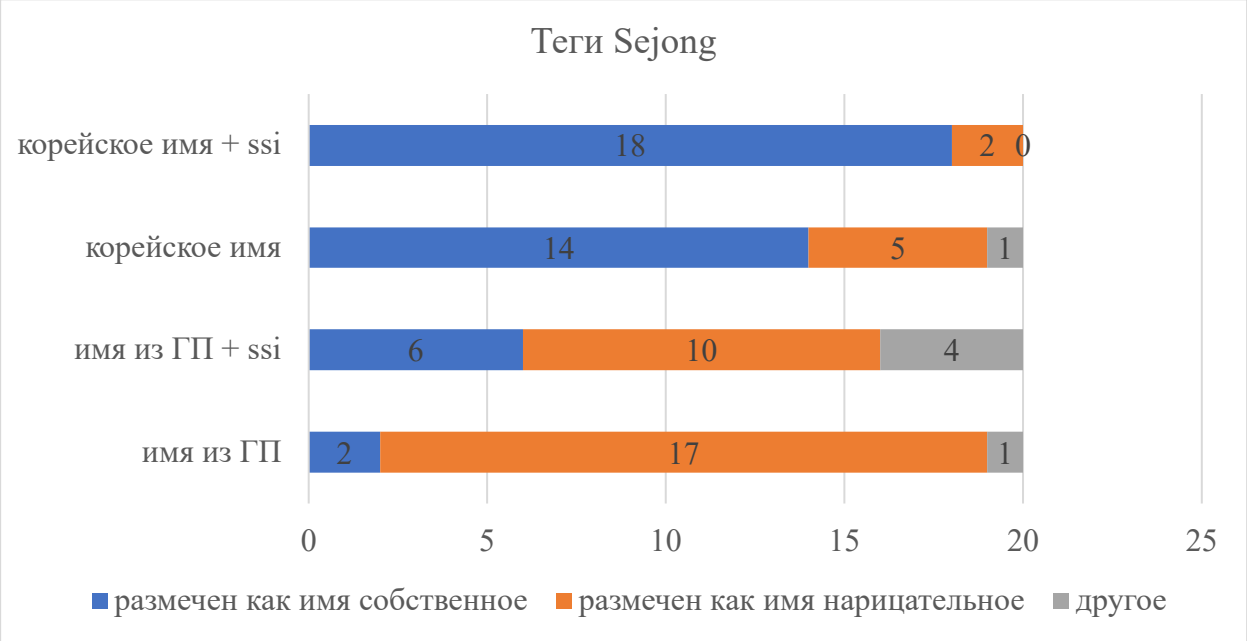
5. Качество разметки имён собственных

Во всех трёх тегсетах есть отдельный тег для имён собственных: PROPN в UD, NNP в Sejong и nq в KAIST. Ошибки при разметке имён собственных допускают обе модели. Чтобы проверить, что влияет на качество разметки имён собственных, было отобрано 20 корейских имён и 20 имён персонажей из книги «Гарри Поттер». С каждым именем было найдено два предложения: со служебным именем *ssi* ‘господин/госпожа’ после имени собственного и без него. Было предположено, что корейские имена будут распознаны парсером лучше, чем имена из «Гарри Поттера», а также парсеру будет легче определить, что в оджоль входит имя собственное, если за ним будет следовать служебное имя *ssi*, так как оно может стоять только после имён собственных.

Гипотеза частично подтвердилась: теги KAIST чаще приписывались корейским именам правильно, если после них стояло *ssi*, а теги Sejong чаще приписывались правильно корейским именам, но не именам из «Гарри Поттера».

На диаграмме 2 представлены результаты разметки имён собственных моделью Stanza GSD XPOS-тегами. Другие результаты проверки разметки тестовых предложений представлены в приложении 3. Важно учитывать, что результаты расстановки UPOS-тегов моделью UDPipe KAIST напрямую определяются расстановкой XPOS-тегов.

Диаграмма 2



6. Заключение

В работе было проведено сравнение тегсетов UD, KAIST и Sejong, используемых парсерами UDPipe и Stanza для корейского языка. Частеречное деление, предлагаемое в этих тегсетах, было проанализировано с точки зрения грамматики корейского языка [Martin 1992]. Проблему для универсального тегсета составляют атрибутивы и бытийные и локативные связки, остальные же части речи могут быть описаны с его помощью.

Также было проведено сравнение качества морфоразметки парсеров UDPipe и Stanza для корейского языка: более точной оказалась разметка модели Stanza GSD. Обе модели не всегда корректно размечали имена собственные и прилагательные.

Основной причиной ошибок модели UDPipe KAIST оказалась некорректная конвертация корпуса KAIST в соответствующий формату Universal Dependencies корпус UD Korean KAIST, на котором модель обучалась. Однако в том случае, если разметка корпуса будет исправлена, переобученная модель UDPipe KAIST может показать более высокие результаты, чем Stanza GSD. Кроме того, возможно, включение в корпуса, на которых будут обучаться модели, текстов с именами собственными, не являющимися корейскими именами, может помочь повысить качество разметки таких токенов.

Библиография

1. Choi J. D. Preparing Korean data for the shared task on parsing morphologically rich languages //arXiv preprint arXiv:1309.1649. – 2013.
2. Choi K. S. et al. KAIST tree bank project for Korean: Present and future development //Proceedings of the International Workshop on Sharable Natural Language Resources. – 1994. – С. 7-14.
3. Chun J. et al. Building universal dependency treebanks in Korean //Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). – 2018.
4. De Marneffe M. C. et al. Generating typed dependency parses from phrase structure parses //Lrec. – 2006. – Т. 6. – С. 449-454.
5. Kang B., Kim H. Sejong Korean Corpora in the Making //LREC. – 2004. – Т. 2004. – С. 1747-1750.
6. Kim H. Korean national corpus in the 21st century Sejong project //Proceedings of the 13th NIJL international symposium. – National Institute for Japanese Language Tokyo, 2006. – С. 49-54.
7. Lee D. G., Rim H. C. Probabilistic modeling of Korean morphology //IEEE transactions on audio, speech, and language processing. – 2009. – Т. 17. – №. 5. – С. 945-955.
8. Martin S. E. A reference grammar of Korean: A complete guide to the grammar and history of the Korean language. – Tuttle Publishing, 1992.
9. McDonald R. et al. Universal dependency annotation for multilingual parsing //Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). – 2013. – С. 92-97.
10. Straka M. UDPipe 2.0 prototype at CoNLL 2018 UD shared task //Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. – 2018. – С. 197-207.

Приложение 1. Части речи по Мартину

Приложение 2. Ошибки при конвертации корпуса KAIST

XPOS-теги	Неверное соответствие UPOS-тегу	Верное соответствие UPOS-тегу	Пример слова	
			слово	XPOS-теги
xp	PART	тег для приставки не может определять UPOS-тег	헛소리	xp+ncn
			hes-soli	ложный-звук
			‘чушь’	
именной тег +jca	ADV	соответствующий именной тег (NOUN/PROPN/ PRON/NUM)	서울에서	nq+jca
			sewul eyse	Сеул LOC
			‘в Сеуле’	
именной тег +jcj	CCONJ	соответствующий именной тег (NOUN/PROPN/ PRON/NUM)	한국과	nq+jcj
			hankwuk kwa	Корея COM_2
			‘Корея и’	
именной тег +jct	CCONJ	соответствующий именной тег (NOUN/PROPN/ PRON/NUM)	대본과	ncn+jct
			taypon-kwa	сценарий-COM_1
			‘со сценарием’	
именной тег +jcv	INTJ	соответствующий именной тег (NOUN/PROPN/ PRON/NUM)	애들아	ncn+xsn+jcv
			ay-tul a	ребёнок-PL VOC
			‘дети’	
тег предикатива +ecc	CCONJ	соответствующий тег предикатива (ADJ/VERB/AUX)	하고	pvg+ecc
			ha-ko	делать-CONV
			‘делая’	
тег предикатива +ecs	SCONJ	соответствующий тег предикатива (ADJ/VERB/AUX)	하면	pvg+ecs
			hamyen	делать-CONV
			‘когда делаю’	
тег предикатива +ecx+px	AUX	соответствующий тег предикатива (ADJ/VERB)	찾아가다	pvg+ecx+px+ef
			chacakata	искать-INF-идти-FIN
			‘навещать’	
тег предикатива +ef+jcr	VERB	соответствующий тег предикатива (ADJ/VERB/AUX)	크다고	paa+ef+jcr
			khu-ta-ko	большой-FIN-QUOT
			‘…, что большой’	
тег предикатива +ef+jxf	VERB	соответствующий тег предикатива (ADJ/VERB/AUX)	커요	paa+ef+jxf
			khu-e-yo	большой-FIN-DISC
			‘большой’	
ncps+xsm	VERB	ADJ	요란한	ncps+xsm+etm
			yolanhan	шум+делать+PART.PST
			‘шумный’	
pad	VERB	ADJ	그렇습니다	pad+ef
			kulehsupnita	такой-FIN
			‘такой’	

Приложение 3. Имена собственные и нарицательные

UDPipe KAIST	тег UD			
	имя из ГП	имя из ГП + *ssi*	корейское имя	корейское имя + *ssi*
размечен как имя собственное	2	5	2	8
размечен как имя нарицательное	11	6	11	11
другое	7	9	7	1
	тег KAIST			
	имя из ГП	имя из ГП + *ssi*	корейское имя	корейское имя + *ssi*
размечен как имя собственное	4	5	2	8
размечен как имя нарицательное	14	6	15	11
другое	2	9	3	1

Stanza GSD	тег UD			
	имя из ГП	имя из ГП + *ssi*	корейское имя	корейское имя + *ssi*
размечен как имя собственное	0	0	0	0
размечен как имя нарицательное	16	17	19	20
другое	4	3	1	0
	тег Sejong			
	имя из ГП	имя из ГП + *ssi*	корейское имя	корейское имя + *ssi*
размечен как имя собственное	2	6	14	18
размечен как имя нарицательное	17	10	5	2
другое	1	4	1	0

Приложение 4. Сравнительная таблица тегсетов

UD		Sejong		KAIST		Часть речи по Мартину	Примеры/ комментарии
тег	значение	тег	значение	тег	значение		
NOUN	существительное	NNB	служебное существительное	nbn	служебное несчётное существительное	квазисвободное существительное	것 kes ‘вещь, факт’
						несчётное функциональное имя	씨 ssi ‘господин, госпожа’
				nbu	служебное счётное существительное	счетное слово	달 tal ‘месяц’
		NNG	имя нарицательное	ncn	непредикативное имя	свободное существительное	고양이 koyangi ‘кошка’
				ncpa	предикативное имя действия	глагольное существительное	공부하다 **kongpu-**ha-ta **учёба**-делать-FIN ‘учиться’
				ncps	предикативное имя состояния	адъективное существительное	가능하다 **kanung**-ha-ta возможность-делать-FIN ‘возможный’
PROPN	имя собственное	NNP	имя собственное	nq	имя собственное	имя собственное	알버스 alpesu ‘Альбус’
NUM	числительное	NR	числительное	nnc	количественное числительное	количественное числительное	한 han ‘один’
				nno	порядковое числительное	порядковое числительное	둘째 twulccay ‘второй’
PRON	местоимение	NP	местоимение	npd	местоименный демонстратив	дейктическое имя	뭐 mwe ‘что’
				npp	личное местоимение		나 na ‘я’
INTJ	междометие	IC	междометие	ii	междометие	междометие	아니 ani ‘нет’

SCONJ	подчинительный союз						
CCONJ	сочинительный союз	MAJ	соединительное наречие	maj	соединительное наречие	соединительное наречие	그리고 kuriko ‘и’
ADV	наречие	MAG	обычное наречие	mad	наречный демонстратив	дейктическое наречие	그리 kuli ‘туда’
				mag	обычное наречие	наречие	지금 cikum ‘сейчас’
DET	детерминатив	MM	атрибутив	mmd	атрибутивный демонстратив	дейктический атрибутив	그 ku ‘тот’
ADJ	прилагательное			mma	качественный атрибутив	атрибутив	옛 yeys ‘старый’
		VA	прилагательное	pad	адъективный демонстратив	прилагательное	그렇다 kurehta ‘такой’
				paa	качественное прилагательное		크다 khuta ‘большой’
		VCN	отрицательная бытийная связка			связка	아니다 anita ‘не (быть)’
VERB	глагол						
		VCP	бытийная связка	jp	бытийная связка		이다 ita ‘быть’
		VV	глагол	pvd	глагольный демонстратив	глагол	그러다 kureta ‘делать так’
				pvg	обычный глагол		가다 kata ‘идти’
AUX	вспомогательный предикатив	VX	вспомогательный предикатив	px	вспомогательный предикатив	вспомогательный глагол	보다 pota ‘пробовать’
						вспомогательное прилагательное	싶다 siphta ‘хотеть’
		SH	китайское слово				飛行 ‘полёт’
X	неизвестное слово	SL	иностранное слово	f	иностранное слово		University ‘университет’
		NF	неизвестное существительное				
		NV	неизвестный предикатив				

		NA	неизвестное слово				
ADP	адлог						
PART	частица						
		JKS	номинатив$_1$	jcs	номинатив$_1$	частица	고양이가 호랑이가 되었습니다. koyangi-ka holangi-ka toy-ess-supnita кошка-NOM$_1$ тигр-NOM$_2$ стать-PST-FIN Кошка стала тигром.
		JKC	номинатив$_2$	jcc	номинатив$_2$		
		JKG	генитив	jcm	генитив		
		JKO	аккузатив	jco	аккузатив		
		JKV	вокатив	jcv	вокатив		
		JKB	адвербиальный падеж	jca	адвербиальный падеж		датив, локатив, инструменталис
				jct	комитатив		언니가 엄마와 같다. enni-ka emma-wa kath-ta старшая_сестра-NOM$_1$ мама-COM$_1$ похожий-FIN Старшая сестра похожа на маму.
		JC	соединительный падеж	jcj	соединительный падеж		엄마와 언니 emma-wa enni мама-COM$_2$ старшая_сестра мама и старшая сестра
		JKQ	цитирование	jcr	цитирование		먹고 싶다**고** 말했다. mek-ko siph-ta-**ko** mal-ha-ess-ta есть-CONV FIN-**CONV** слово-делать-PST-FIN Сказал, что хочет есть.
		JX		jxt	топик		

			выделительная частица	jxc	другие выделительные частицы		
		EF	финитное окончание	jxf	завершающая выделительная частица		
				ef	финитное окончание		
		EC	связующее окончание	ecc	соединительное окончание		
				ecs	подчинительное окончание		
				ecx	вспомогательное окончание		
		EP	суффикс перед окончанием	ep	суффикс перед окончанием		суффиксы прошедшего и будущего времени, гонорифик
		ETN	номинализатор	etn	номинализатор		
		ETM	адъективизатор	etm	адъективизатор		
		XPN	именной префикс	xp	префикс	связанные атрибутивы и наречия	비- pi- ‘не-’
		XSN	именной суффикс	xsn	именной суффикс		суффикс множественного числа
		XSV	глагольный суффикс	xsv	глагольный суффикс	послеименной глагол	공부하다 kongpu-**ha**-ta учёба-**делать**-FIN ‘учиться’
		XSA	суффикс прилагательного	xsm	суффикс прилагательного	послеименное прилагательное	가능하다 kanung-**ha**-ta возможность-**делать**-FIN ‘возможный’
				xsa	суффикс наречия		
		XR	корень				