DOI https://doi.org/10.30525/978-9934-26-110-7-30

ЧАСТОТНІ МОДЕЛІ РЕЧЕНЬ НА 4-6 СЛІВ ДЛЯ МОРФЕМИ «给»

Козоріз О. П.
кандидат філологічних наук,
асистент кафедри мов і літератур Далекого Сходу
та Південно-Східної Азії
Інституту філології
Київський національний університет імені Тараса Шевченка
м. Київ, Україна

Для досягнення мети дослідження слід виконати такі завдання: 1. Визначити критерії та відібрати матеріал для дослідження. 2. Запропонувати методику та виділити найчастотніші моделі речень на 4-6 слів, в тому числі з морфемою 给. 3. Створити програму автоматичного тегування, виокремлення моделей, створення списку частотних моделей речень. 4. Перевірити ефективність роботи програми на паралельних перекладах китайської та англійської мов.

Першим кроком шляхом завантаження з сайту QuWord [11] було створено китайсько-англійський паралельний корпус на 920 000 пар речень [1]. На його основі відібрані за англійською частиною речення з довжинами від 4 до 6 слів для морфеми «给». Відбір речень саме цієї довжини зумовлений їхньою частотністю у розмовній мові, складністю для перекладу системами машинного перекладу, теоретичною можливістю їхньої систематизації (через відносно невелику довжину). Морфема «给» була обрана через наявність у неї граматичної омонімії, аби перевірити якість роботи програм. Відбір проводився регулярними виразами: ^.*\t\w*\W\w*\W\w*\W\w*\W$ – 4 слова і т. д. Англійська мова використовувалися оскільки можна однозначно визначити межі слова. Всього було відібрано 1105 пар речень, 4 слова – 100 речень, 5 слів – 370 речень, 6 слів – 635 речень. Порівняльні графіки кількості слів у реченні для китайської та англійської мов можна бачити на **Рис. 1**. Китайські речення виявилися дещо довшими за англійські, це пов’язано в тому числі з функціями структурних елементів речення, які може виконувати зазначена морфема, утворювати пасив, прийменникові словосполучення тощо.

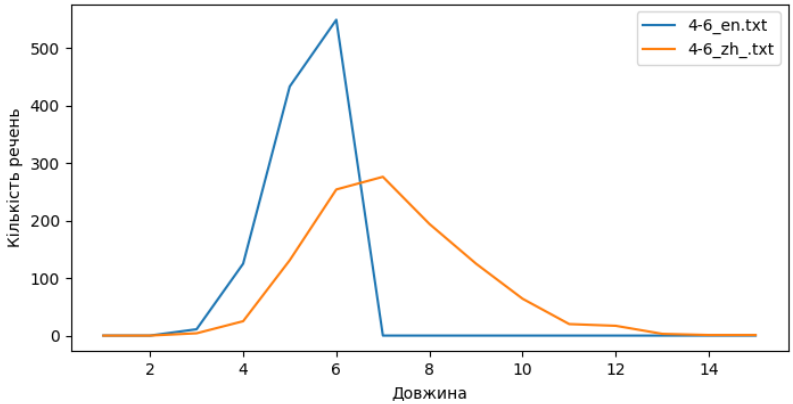

Рис. 1. Порівняльний графік кількості слів у реченні для китайської та англійської мов

Тегування англійської частини було проведене за допомогою Free CLAWS web tagger [5]. Тут мається на увазі POS тегування – (Part-of-speech) частиномовна розмітка, яка також називається граматичним тегуванням, є поширеною формою анотації корпусу, була першою формою анотації розробленою в UCREL Ланкастера, за різними даними досягає точності 93-97%.

У школах вивчають 9 частин мови, однак категорій та підкатегорій можна виділяти набагато більше. Ми користувалися UCREL CLAWS5 Tagset, що включає 62 категорії слів, для дієслів англійської мови тут пропонується 24 категорії, для іменників та займенників – 4 категорії на кожну частину мови [9].

Після тегування 1105 речень, **див. Табл. 1**, загалом отримали 872 моделі речень, що складає 79%: 4 слова – 79 моделей речень, 5 слів –285 моделей речень, 6 слів – 512 моделей речень. Можливо через занадто широку граматичну категоризацію англійської мови, моделей речень на 4-6 лише для одного слова виявилося майже 80% від початкової кількості речень.

Таблиця 1

Ілюстративна таблиця тегованого англійського тексту та відповідні моделі речень

Англійська мова	**Тегований англійський текст**	**Моделі речень**
Give me a kiss! Who'd you call?	Give_VVB me_PNP a_AT0 kiss_NN1 !_SENT	VVB PNP AT0 NN1 SENT

Give me a hug.	Who_PNQ ‘d_VM0 you_PNP call_VVI ?_SENT	PNQ VM0 PNP VVI SENT
Give me two cans.	Give_VVB me_PNP a_AT0 hug_NN1 ._SENT	VVB PNP AT0 NN1 SENT
Give me an int.	Give_VVB me_PNP two_CRD cans_NN2 ._SENT	VVB PNP CRD NN2 SENT
Doctor: For whom?	Give_VVB me_PNP an_AT0 int_NN1 ._SENT	VVB PNP AT0 NN1 SENT
I sent her tapes.	Doctor_NN1 :_PUN For_PRP whom_PNQ ?_SENT	NN1 PUN PRP PNQ SENT
Hand me a plate.	I_PNP sent_VVD her_DPS tapes_NN2 ._SENT	PNP VVD DPS NN2 SENT
I painted a mask.	Hand_VVB me_PNP a_AT0 plate_NN1 ._SENT	VVB PNP AT0 NN1 SENT
I pay the cashier.	I_PNP painted_VVD a_AT0 mask_NN1 ._SENT	PNP VVD AT0 NN1 SENT
Bill gave him one.	I_PNP pay_VVB the_AT0 cashier_NN1 ._SENT	PNP VVB AT0 NN1 SENT
The door blows open.	Bill_NP0 gave_VVD him_PNP one_PNI ._SENT	NP0 VVD PNP PNI SENT
Would you alter it?	The_AT0 door_NN1 blows_NN2 open_VVB ._SENT	AT0 NN1 NN2 VVB SENT
Good tidings to you.	Would_VM0 you_PNP alter_VVI it_PNP ?_SENT	VM0 PNP VVI PNP SENT
Stop that at once!	Good_AJ0 tidings_NN2 to_PRP you_PNP ._SENT	AJ0 NN2 PRP PNP SENT
Get John a drink.	Stop_VVB that_CJT at_AV0 once_AV0 !_SENT	VVB CJT AV0 AV0 SENT
Give me the invoice.	Get_VVB John_NP0 a_AT0 drink_NN1 ._SENT	VVB NP0 AT0 NN1 SENT
Give him the hats.	Give_VVB me_PNP the_AT0 invoice_NN1 ._SENT	VVB PNP AT0 NN1 SENT
He tuned the guitar.	Give_VVB him_PNP the_AT0 hats_NN2 ._SENT	VVB PNP AT0 NN2 SENT
Who gives you gifts?	He_PNP tuned_VVD the_AT0 guitar_NN1 ._SENT	PNP VVD AT0 NN1 SENT
	Who_PNQ gives_VVZ you_PNP gifts_NN2 ?_SENT	PNQ VVZ PNP NN2 SENT

Для китайської мови спочатку ми провели тегування за допомогою сайту 汉语分词和词性自动标注 [10], що містить 35 граматичних категорій слів: 11 – іменники, 4 – дієслова тощо. Після тегування 1105 речень, **див. Табл. 2**, загалом отримали 1027 моделі речень, що складає 93%.

Тут виникає багато запитань як щодо алгоритму розбивки тексту на слова, так і до безпосередньо тегування: ієрогліф 给 скрізь позначено як «/p» – тобто прийменник, хоча він може бути і дієсловом, **一个** – це одне слово, що позначено як «/r « – тобто займенник, у іншому випадку ієрогліф **一** – це числівник «/m», з чим ми згодні, а ієрогліф 张 – це прізвище «/nhf», хоча і є категорія «/q» – рахівні слова, що була б доречна у цьому випадку. 约/v **翰**/x – ім'я поділено на окремі частини.

Таблиця 2

Ілюстративна таблиця тегованого китайського тексту та відповідні моделі речень від сайту 汉语分词和词性自动标注

№	Китайська мова	Тегований китайський текст	Моделі речень
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20	给我个吻 给谁打过电话？ 给我一个拥抱吧 给我来两罐 给我一个int**数** **医生**：给谁的 **我**给她寄磁带 递给我一个盘子 **我**给面具上了色 **我付**钱给收银 **比**尔给了他一张。 门给吹开了 **您能**给锁边吗？ **好消息**给你 马上给我停下来 给约翰拿杯饮料来 **把**发票给我。 **把**这些帽子送给他 。 **他**给吉他定弦。 谁给你的礼物?	给/p **我**/r **个**/q **吻**/v 。/w 给/p 谁/r **打**/v 过/vd 电话/n ？ 给/p **我**/r **一个**/r 拥抱/v **吧**/u 。/w 给/p **我**/r **来**/vd **两**/m **罐**/n 。/w 给/p **我**/r **一个**/r int/ws **数**/n 。 **医生**/n ：/w 给/p 谁/r **的**/u 。/w **我**/r 给/p **她**/r **寄**/v **磁**带/n 。/w 递给/v **我**/r **一个**/r 盘子/n 。/w **我**/r 给/p **面具**/n **上**/nd **了**/u **色**/n 。/w **我**/r **付**钱/v 给/p **收**银员/n 。/w **比**/p 尔/x 给/p **了**/u **他**/r **一**/m 张/nhf 。/w 门/n 给/p **吹开**/v **了**/u 。/w **您**/r **能**/vu 给/p 锁/n 边/n 吗/u ？/w **好消息**/n 给/p **你**/r 。/w 马上/d 给/p **我**/r **停下来**/v ！/w 给/p 约/v **翰**/x **拿**/v **杯**/n 饮料/n **来**/vd 。/w **把**/p 发票/n 给/p **我**/r 。/w **把**/p 这些/r **帽子**/n **送**给/v **他**/r 。/w **他**/r 给/p **吉他**/n **定**/v **弦**/n 。/w 谁/r 给/p **你**/r **的**/u **礼物**/n ？/w	p r q v w p r v vd n w p r r v u w p r vd m n w p r r ws n w n w p r u w r p r v n w v r r n w r p n nd u n w r v p n w p x p u r m nhf w n p v u w r vu p n n u w n p r w d p r v w p v x v n n vd w p n p r w p r n v r w r p n v n w r p r u n w

Інша програма тегування китайського тексту SegmentAnt [12], містить 24 граматичних маркування слів. Після тегування 1105 речень, див. **Табл. 3**, загалом отримали 1015 моделей речень, що складає 91%. Тут також виникають питання до 给_p, яке скрізь тегується прийменником, дивні поєднання словосполучень **我个**_r , **一个**_m, **一**张_m, натомість розрізняє імена – 约翰_nrt. Тобто довіряти такому тегуванню можна досить відносно, а отже отримані моделі речень будуть мати великий відсоток помилок (біля 10%). Визначення частин мови для таких морфем як 给, **在**, **到** потребує досить серйозного алгоритму.

Таблиця 3

Ілюстративна таблиця тегованого китайського тексту виконаного програмою SegmentAnt та відповідні моделі речень

№	Китайська мова	Тегований китайський текст	Моделі речень
1	给我个吻。	给_p **我个**_r **吻**_v 。_x	p r v x
2	给谁打过电话？	给_p 谁_r **打**_v 过_ug 电话_n ？_x	p r v ug n x
3	给我一个拥抱吧	给_p **我**_r **一个**_m 拥抱_v **吧**_y 。_x	p r m v y x
4	给我来两罐。	给_p **我**_r **来**_v **两罐**_m 。_x	p r v m x
5	给我一个int**数**。	给_p **我**_r **一个**_m int_eng **数**_n 。_x	p r m eng n x
6	**医生**：给谁的。	**医生**_n ：_x 给_p 谁_r **的**_uj 。_x	n x p r uj x
7	**我**给她寄磁带。	**我**_r 给_p **她**_r **寄**_v **磁**带_n 。_x	r p r v n x
8	递给我一个盘子	递给_v **我**_r **一个**_m 盘子_n 。_x	v r m n x
9	**我**给面具上了色	**我**_r 给_p **面具**_n **上**_f **了**_ul **色**_n 。	r p n f ul n x
10	**我付**钱给收银员	**我**_r **付**钱_v 给_p **收**银员_n 。_x	r v p n x
11	**比**尔给了他一张	**比**尔_nrt 给_p **了**_ul **他**_r **一**张_m 。	nrt p ul r m x
12	门给吹开了。	门_n 给_p **吹开**_v **了**_ul 。_x	n p v ul x
13	**您能**给锁边吗？	**您**_r **能**_v 给_p 锁边_n 吗_y ？_x	r v p n y x
14	**好消息**给你。	**好消息**_n 给_p **你**_r 。_x	n p r x
15	马上给我停下来	马上_d 给_p **我**_r **停下来**_v ！_x	d p r v x
16	给约翰拿杯饮料	给_p 约翰_nrt **拿**_v **杯**_q 饮料_n **来**_v 。_x	p nrt v q n v x
17	来。		
18	**把**发票给我。	**把**_p 发票_n 给_p **我**_r 。_x	p n p r x
19	**把**这些帽子送给	**把**_p 这些_r **帽子**_n **送**给_v **他**_r 。	p r n v r x
20	他。	**他**_r 给_p **吉他**_ns **定弦**_n 。_x	r p ns n x
	他给吉他定弦。		r p r uj n x
	谁给你的礼物？	谁_r 给_p **你**_r **的**_uj **礼物**_n ？_x	

Після алфавітного сортування списку моделей речень, регулярним виразом типу «(^.*$)\n(\1)\n(\1)\n…» вручну було виділено 15 найчастотніших моделей китайських речень з морфемою «给» на 4-6 слів, див. **Табл. 4**, які розташовані у порядку спадання частотності. Для пошуку моделей речень серед тегованого тексту можна користуватися регулярним виразом типу: [^_]*_r[^_]*_p[^_]*_n [^_]*_v[^_]*_ul[^_]*_r[^_]*_x. Через недоліки роботи програми тегування китайського тексту, хоча ми і змогли виявити найчастотніші моделі речень, моделі 7-10 та 14 виглядають сумнівно з точки зору граматичної приналежності «给» до прийменника, але через невелику їх кількість піддаються коригуванню та можуть бути використані у навчальному процесі.

Таблиця 4

Частотні моделі китайських речень з морфемою 给

	Модель речення	**Приклади вживання**
	r v r m n x	他_r 借给_v 我_r 一辆_m 自行车_n 。_x 他_r 送给_v 她_r 许多_m 礼物_n 。_x
	r p r v ul m n x	她_r 给_p 我_r 烤_v 了_ul 一块_m 牛排_n。 他_r 给_p 我_r 带来_v 了_ul 一些_m 糖果_n。
	n p v ul x	门_n 给_p 吹开_v 了_ul 。_x 下水道_n 给_p 堵住_v 了_ul 。 肿瘤_n 给_p 切除_v 了_ul。_x
	r p ul r m n x	他_r 给_p 了_ul 我_r 一点_m 墨水_ns 。_x 他_r 给_p 了_ul 我_r 一些_m 水果_n 。_x 他们_r 给_p 了_ul 我_r 许多_m 礼物_n 。_x
	r p n v r x	他_r 把_p 小刀_n 借给_v 我_r 。_x 他_r 把_p 球_n 扔给_v 我_r 。_x
	r p n v ul r x	她_r 把_p 钱_n 借给_v 了_ul 我_r 。_x 我_r 把_p 钥匙_n 扔给_v 了_ul 他_r 。_x
	r p r uj n v r ul x	他_r 把_p 他_r 的_uj 被子_n 借给_v 我_r 了_ul 。她_r 把_p 她_r 的_uj 邮票_n 送给_v 我_r 了_ul。_x
	n p n n x	蜜蜂_n 给_p 果树_n 传粉_n 。_x 大人_n 给_p 小孩_n 红包_n 。_x
	p r m n x	给_p 我_r 一把_m 螺丝刀_n 。_x 给_p 我_r 一些_m 自主权_n 。_x
0	v p r m n x	请_v 给_p 我_r 一些_m 餐巾纸_n 。_x 请_v 给_p 他_r 一些_m 瓶子_n 。_x
	r d p n v n x	他_r 必须_d 给_p 锅炉_n 加_v 燃料_n 。_x 她_r

1		曾_d 把_p 房间_n 租给_v 大学生_n 。_x
2	r p n v r ul x	他_r 把_p 流感_n 传给_v 我_r 了_ul 。_x 我_r 把_p 自行车_n 借给_v 他_r 了_ul 。_x
3	r p n v ug n y x	你_r 给_p 狗_n 打_v 过_ug 虫_n 吗_y ?_x 你_r 给_p 奶牛_n 挤_v 过_ug 奶_n 吗_y ？_x
4	r p r m n x	他_r 给_p 她_r 一个_m 苹果_n 。_x 我_r 给_p 你_r 四分_m 钱_n 。_x
5	r p r uj n v r x	他_r 把_p 他_r 的_uj 财产_n 让给_v 我_r 。_x 他_r 把_p 他_r 的_uj 钱_n 借给_v 她_r 。_x

Щоб автоматизувати зазначений процес, завдяки бібліотекам Jieba [6] ми розробили програму на мові Python, що створює список усіх моделей речень корпусу китайського тексту:

```
import jieba
import jieba.posseg as pseg
import re

with open('4-6_zh.txt', 'rt', encoding='utf-8') as f0:
    s = f0.read()
words = pseg.cut(s)
with open('teg_text.txt', 'wt', encoding='utf-8') as f1:
    for word, flag in words:
        s2 = ('%s_%s' % (word, flag))
        print(s2, file=f1, end=' ')
with open('teg_text.txt', 'rt', encoding='utf-8') as f2:
    with open('teg_text_models.txt', 'wt', encoding='utf-8') as f3:
        s2 = f2.read()
        s2 = re.sub(r'\_x ', '', s2)
        s2 = re.sub(r'^[^\_]+_', '', s2, flags=re.MULTILINE)
        s2 = re.sub(r' [^\_\n]+?\_', '_', s2, flags=re.MULTILINE)
        frequency = {}
        match_pattern = re.findall(r'\b.+\b', s2,
flags=re.MULTILINE)
        for word in match_pattern:
         count = frequency.get(word,0)
         frequency[word] = count + 1
        list_d = list(frequency.items())#сортировка по значению
        list_d.sort(key=lambda i: i[1], reverse=True)
        for i in list_d:
            print(i[0], '\t', i[1], file=f3)
```

Результати роботи програми можна переглянути у **Табл. 5**., де зазначені моделі китайських речень з морфемою «给» та частотність їх вживання в автоматичному режимі.

Таблиця 5

Частотні моделі китайських речень з морфемою «给» отримані автоматично

	Модель речення	**частотність**		**Модель речення**	**частотність**
	r_v_r_m_n	9	1	r_p_r_m_n	3
	r_p_r_v_ul_m_n	7	2	r_p_n_v_ug_n_y	3
	n_p_v_ul	4	3	n_p_n_n	3
	r_p_ul_r_m_n	4	4	r_p_r_uj_n_v_r	3
	r_p_n_v_r	4	5	r_d_p_n_v_n	3
	r_p_n_v_ul_r	4	6	r_p_r_v_n	2
	r_p_r_uj_n_v_r_ul	4	7	r_p_n_f_ul_n	2
	v_p_r_m_n	3	8	p_r_n_v_r	2
	p_r_m_n	3	9	p_n_v_r	2
0	r_p_n_v_r_ul	3	0	v_m_n_p_r_v	2

Аби перевірити висновки щодо кількісних характеристик моделей речень не для конкретного слова, було проведено ще один експеримент. Із зазначеного вище китайсько-англійського паралельного корпусу на 920 000 пар речень за англійською частиною відібрано речення лише із 4 слів і побудовано відповідні моделі. Усього виявилося 13 000 речень із 4 слів, після тегування отримано близько 4200 моделей речень, що складає 32%. Тобто практично можна скоротити кількість моделей на дві третини відносно початкової кількості речень.

Після сортування списку моделей речень, регулярним виразом вручну за попередньо описаною технологію було виділено 20 найчастотніших моделей речень англійської мови на 4 слова, див. **Табл. 6**, розташовані у порядку спадання частотності.

Таблиця 6

Частотні моделі англійських речень на 4 слова

№	Кіль-кість	Модель речення	Приклади вживання	
1	185	PNP VVD AT0 NN1	I_PNP got_VVD the_AT0 flu_NN1 .	I_PNP dreamed_VVD a_AT0 crocodile_NN1 .
2	160	PNP VVD DPS NN1	I_PNP bumped_VVD my_DPS head_NN1 .	I_PNP blared_VVD my_DPS horn_NN1 .
3	124	PNP VVD PRP NN1	They_PNP went_VVD by_PRP ship_NN1 .	We_PNP woke_VVD at_PRP dawn_NN1 .
4	121	AT0 NN1 VVD AV0	The_AT0 tide_NN1 ebbed_VVD away_AV0 .	The_AT0 bullet_NN1 hit_VVD home_AV0 .
5	85	AT0 NN1 VBZ AJ0	A_AT0 drum_NN1 is_VBZ noisy_AJ0 .	The_AT0 tyre_NN1 is_VBZ flat_AJ0 .
6	84	PNP VBZ AV0 AJ0	He_PNP ‘s_VBZ very_AV0 dim_AJ0 .	It_PNP ‘s_VBZ too_AV0 salty_AJ0 .
7	80	PNP VVD DPS NN2	I_PNP sent_VVD her_DPS tapes_NN2 .	He_PNP wet_VVD my_DPS pants_NN2 .
8	78	PNP VVB DPS NN1	I_PNP like_VVB my_DPS job_NN1 .	I_PNP miss_VVB my_DPS dad_NN1 .
9	66	VVB AT0 NN1 NN1	Put_VVB the_AT0 book_NN1 back_NN1 .	Take_VVB a_AT0 heap_NN1 dump_NN1 .
10	64	AT0 NN1 VVZ AV0	The_AT0 wind_NN1 blows_VVZ hard_AV0 .	The_AT0 road_NN1 ends_VVZ here_AV0 .
11	59	AT0 NN1 VVZ NN1	A_AT0 tree_NN1 needs_VVZ bark_NN1 .	The_AT0 sun_NN1 emits_VVZ light_NN1 .
12	57	AT0 NN1 VBD AJ0	The_AT0 bin_NN1 was_VBD full_AJ0 .	The_AT0 sea_NN1 was_VBD wild_AJ0 .
13	55	DT0 VBZ DPS NN1	This_DT0 is_VBZ my_DPS pen_NN1 .	This_DT0 is_VBZ my_DPS toy_NN1 .
14	54	DT0 NN2 VBB AJ0	All_DT0 men_NN2 are_VBB mortal_AJ0 .	These_DT0 shirts_NN2 are_VBB new_AJ0 .

15	54	NP0 VVD AT0 NN1	Eddie_NP0 felt_VVD a_AT0 shiver_NN1 .	Hugh_NP0 left_VVD the_AT0 room_NN1 .
16	53	DPS NN1 VBZ AJ0	His_DPS fur_NN1 is_VBZ white_AJ0 .	Its_DPS trunk_NN1 is_VBZ big_AJ0 .
17	53	VVB AT0 AJ0 NN1	Surf_VVB a_AT0 Hawaiian_AJ0 wave_NN1 .	Click_VVB the_AT0 General_AJ0 TAB_NN1 .
18	52	AT0 NN1 VVD AVP	The_AT0 gamble_NN1 paid_VVD off_AVP .	The_AT0 ground_NN1 caved_VVD in_AVP .
19	52	AT0 NN2 VBB AJ0	The_AT0 rides_NN2 are_VBB free_AJ0 .	The_AT0 ants_NN2 are_VBB busy_AJ0 .
20	52	PNP VVZ PRP NN1	He_PNP suffers_VVZ from_PRP asthma_NN1 .	She_PNP sighs_VVZ in_PRP agreement_NN1 .

Якщо зменшити кількість категорій, нагадаю, тут для англійської мови було використано 62 граматичні категорії, то теоретично кількість моделей речень можна скоротити ще.

Для англійського тексту щоб автоматизувати зазначений процес, ми розробили аналогічну програму на мові Python, що створює список усіх моделей речень корпусу завдяки бібліотекам NLTK [7]. Відрізняється лише початкова частина коду:

```
with open('quword_en_4.txt', 'rt', encoding='utf-8') as f0:
    with open('teg_text_en.txt', 'wt', encoding='utf-8') as f1:
        while True:
            sentence = f0.readline()
            tokens = nltk.word_tokenize(sentence)
            tagged = nltk.pos_tag(tokens)
            for word, flag in tagged:
                s2 = ('%s_%s' % (word, flag))
                print(s2, file=f1, end=' ')
            print('', file=f1)
            if not sentence:
                break
```

Повторивши зазначений експеримент, ми отримали дещо інші данні через різницю в кількості граматичних категорій тегування, тут їх лише 36 [8]. Початкова кількість речень – 13 000, після тегування

отримано 3536 моделей речень, що складає уже 27%, найчастотніші 20 див.**Табл. 7**.

Історично найвпливовішими наборами тегів були ті, що використовувались для позначення американського корпусу Брауна та серії тегів, розроблених в Університеті Ланкастера, а потім Британського національного корпусу CLAWS5; CLAWS5 також називається набором тегів c5 Penn Treebank Tag. Набір тегів Penn Treebank – це спрощена версія набору тегів Brown, широко використовується в обчислювальних роботах для англійської мови. Загалом, набори тегів включають морфологічні відмінності певної мови, і тому безпосередньо не застосовуються до інших мов [2, 139-141].

Таблиця 7

Частотні моделі англійських речень на 4 слова виділені в автоматичному режимі

№	Кіль-кість	Модель речення	Приклади вживання	
1	220	PRP_VBD_DT_NN_	I_PRP got_VBD the_DT axe_NN ._.	It_PRP was_VBD a_DT nova_NN !_.
2	145	PRP_VBP_DT_NN_	I_PRP need_VBP a_DT vase_NN ._.	I_PRP am_VBP the_DT winner_NN !_.
3	143	PRP_VBD_PRP$_NN_	He_PRP wet_VBD my_PRP$ pants_NNS .	I_PRP dropped_VBD my_PRP$ spoon_NN .
4	136	DT_NN_VBZ_JJ_	A_DT drum_NN is_VBZ noisy_JJ ._.	The_DT tyre_NN is_VBZ flat_JJ ._.
5	126	PRP_VBD_IN_NN_	They_PRP went_VBD by_IN ship_NN ._.	We_PRP woke_VBD at_IN dawn_NN ._.
6	124	DT_NN_VBD_RB_	The_DT table_NN was_VBD messy_RB ._.	The_DT baby_NN walks_VBD fast_RB ._
7	119	DT_NN_VBZ_RB_	The_DT water_NN is_VBZ off_RB	The_DT job_NN is_VBZ yours_RB ._.
8	100	PRP_VBZ_DT_NN_	It_PRP ‘s_VBZ a_DT duck_NN ._.	He_PRP is_VBZ a_DT parson_NN ._.

9	100	DT_NNS_VBP_JJ_	The_DT rides_NNS are_VBP free_JJ ._.	All_DT men_NNS are_VBP mortal_JJ ._.
10	100	NNP_VBD_DT_NN_	Tommy_NNP was_VBD a_DT leaf_NN ._.	Pony_NNP had_VBD no_DT idea_NN ._
11	98	PRP_VBZ_RB_JJ_	It_PRP is_VBZ very_RB cute_JJ ._.	It_PRP ‘s_VBZ so_RB unfair_JJ !_.
12	95	PRP_VBP_PRP$_NN_	I_PRP let_VBP her_PRP$ go_NN ._.	I_PRP like_VBP my_PRP$ job_NN ._.
13	94	PRP_VBP_JJ_NNS_	I_PRP see_VBP many_JJ cows_NNS ._.	I_PRP like_VBP fat_JJ chicks_NNS ._.
14	85	NNP_VBZ_DT_NN_	Sophie_NNP knows_VBZ a_DT lot_NN ._.	Kate_NNP has_VBZ a_DT cousin_NN ._.
15	83	DT_NN_VBZ_VBN_	The_DT boat_NN has_VBZ sunk_VBN ._.	The_DT sky_NN is_VBZ cloudy_VBN ._.
16	74	DT_NN_VBZ_NN_	The_DT knife_NN is_VBZ blunt_NN ._.	The_DT soup_NN needs_VBZ salt_NN .
17	71	DT_NN_VBZ_VBG_	A_DT wolf_NN is_VBZ coming_VBG	The_DT dog_NN is_VBZ barking_VBG ._
18	69	DT_NN_IN_NN_	A_DT type_NN of_IN muffin_NN ._.	A_DT plaything_NN of_IN fate_NN ._.
19	67	DT_NN_VBD_JJ_	The_DT pig_NN went_VBD wild_JJ ._.	The_DT weather_NN was_VBD bad_JJ .
20	65	PRP_VBD_PRP_NN_	He_PRP taught_VBD us_PRP physics_NNS ._.	It_PRP made_VBD me_PRP speechless_NN ._.

Порівнявши таблиці **Табл. 6**. та **Табл. 7**., бачимо, що хоча є відмінність у кількості категорій, але загалом частотні моделі речень лишаються ті ж самі.

Враховуючи кількість моделей навіть для коротких речень у 4 слова, якісний машинний переклад на основі правил (Rule-based), що враховуватиме абсолютно всі лінгвістичні варіації, сьогодні досить складне завдання. Методика автоматичного виділення частотних моделей речень може бути застосована для створення якісних підручників іноземних мов.

Література:

1. Козоріз О. Порівняльний аналіз різнотематичних лінгвістичних корпусів. Актуальні питання гуманітарних наук. 2021. Вип. 35. Т. 3. С. 117–125. URL: http://www.aphn-journal.in.ua/archive/35_2021/part_3/18.pdf (accessed 1 July 2021).

2. Christopher D. Manning, Hinrich Schütze, 1999, Foundations of Statistical Natural Language Processing, 680 p.

3. AntConc Homepage. URL: http://www.laurenceanthony.net/software/antconc/ (accessed 1 July 2021).

4. Corpus software and related tools. URL: http://ucrel.lancs.ac.uk/tools.html (accessed 1 July 2021).

5. Free CLAWS web tagger. URL: http://ucrel-api.lancaster.ac.uk/claws/free.html (accessed 1 July 2021).

6. Jieba. URL: https://github.com/fxsjy/jieba (accessed 1 July 2021).

7. NLTK. URL: http://www.nltk.org/ (accessed 1 July 2021).

8. Penn Treebank II Tags. URL: https://web.archive.org/web/ 20130517134339/http://bulba.sdsu.edu/jeanette/thesis/PennTags.html (accessed 1 July 2021).

9. UCREL CLAWS5 Tagset. URL: http://ucrel.lancs.ac.uk/ claws5tags.html (accessed 1 July 2021).

10.汉语分词和词性自动标注. URL: http://corpus.zhonghuayuwen.org/CpsWParser.aspx (accessed 1 July 2021).

11. QuWord. URL: https://www.quword.com/ (accessed 1 July 2021).

12. SegmentAnt. URL: https://www.laurenceanthony.net/software/segmentant/ (accessed 1 July 2021).