

Էզդուարդ Սեկյան

ԲՈՒԼՂԱՐԵՐԵՆԻ ԵՎ ՀԱՅԵՐԵՆԻ ԱՆՈՐՈՇՈՒԹՅԱՆ ԶԱՓԸ
(ԷՆՏՐՈՊԻԱՆ) ԵՎ ՄՈՏԻԿՈՒԹՅԱՆ ԱՍՏԻՃԱՆԸ (ԿՈՌԵԼՅԱՑԻԱՆ)

Մաթեմատիկական լեզվաբանության կարևորագույն խնդիրներից մեկն է հանդիսանում թարգմանության մեքենայացման հարցը: Խնդրի բարդությունը բխում է տարրեր զույգ լեզուների միջև փոխմիարժեք համապատասխանության բացակայությունից: Մի լեզվից մյուսը թարգմանելու ժամանակ ծագում է երկխոսություն (դուալիզմ), կամ անորոշություն, որը անհնարին է դարձնում թարգմանության ավտոմատացման հարցը: Այդ տեսակետից օգտակար է քանակապես դնահատել տարրեր զույգ լեզուների մոտիվությունը, կամ փոխհարաբերակցությունը (կոռելյացիան): Այդպիսի չափի հիմքում ընկած է հնարավոր հասկացողությունը: Եթե որևէ փորձ, կամ զեպք ունի մեկից ավելի հնարավոր ելքեր (վախճաններ), բնդ որում յուրաքանչյուր ելք ունի իրականացման իր (p) հավանականությունը, ապա ասում են, որ փորձը (զեպքը) ունի անորոշություն:

Էնտրոպիան հանդիսանում է անորոշության չափ և արտահայտվում է իրեն հավանականությունների ֆունկցիա կամ $f(p)$:

Ինչքան ավելի մեծ է անորոշությունը, այնքան ավելի մեծ է էնտրոպիան:

Պահանջվող ֆունկցիայի $f(p)$ ավելի լրիվ որոշակիության համար հարկավոր է ներկայացնել նրան լրացուցիչ պահանջներ՝ քննարկել երկու անկախ փորձեր:

Կատարված են երկու փորձ. α (բուլղարերեն լեզվի հետ) և β (հայերեն լեզվի հետ):

Փորձի համար հիմք է ծառայել հետևյալ հիմնական (1) աղյուսակը՝

Աղյուսակ № 1 կազմելու համար կատարված է հատուկ մանրազնին աշխատանք, որովհետև յուրաքանչյուր բուլղարական բառի համար՝ կազմված մեկ կամ ավելի վանկերից, հարկավոր է եղել հայերեն տեքստում գտնել շատ թե քիչ ճիշտ արտահայտված բառ կամ բառեր՝ կազմված մեկ կամ ավելի վանկերից: Աղյուսակը կազմելու համար ուսումնասիրված է բուլղարերեն տեքստի 1040 բառ, որոնք համապատասխանում են հայերեն նույն տեքստում 681 բառի: α և β փորձերի համար օգտագործված է բուլղարացի հեղինակ Է. Մանուկի «Ինլիշանների վախճանը» վեպը, որը բնագրից հայերենի թարգմանել է Ավագյանը: Համեմատության համար ընտրված են վեպի տարրեր գլուխներից հատվածներ՝ երկու լեզուներով էլ:

Այժմ ցույց տանք, թե ինչպես է կազմված № 1 աղյուսակը:

1. Цялото село притихна = Ամբողջ գյուղը լռեց

Ця-ло-то се-ло при-тих-на

Ամ-բողջ գյու-ղը լը-ռեց

		1											
		α											
		1	2	3	4	5	6	7	8	9	10	11	
β	1	$P_{(1,1)}$ 68	32	3	1								104
	2	35	$P_{(2,2)}$ 75	$P_{(3,2)}$ 40	$P_{(4,2)}$ 18	4	1	1	1				175
	3	18	48	68	24	9	6		3	1	1		182
	4	5	22	39	27	12	4	2					111
	5		10	20	18	11	3	1					63
	6		2	6	8	8				$P_{(9,6)}$ 1		1	26
	7		1	1	3	1	2	1		2			11
	8		1		1							1	3
	9						1					1	2
	10				1	2			1				4
	11												0
		126	191	177	105	47	17	5	5	4	1	3	681

2

		α	
		1	2...
β	1	0,08	
	2		

3

		α	
		1	2...
β	1	-1,1	
	2	\	

Цялото նշանակում է ամբողջ: Այս բառը բուլղարերենում կազմված է 3 վանկից, իսկ հայերենում՝ 2 վանկից:

Աղյուսակի վերի հորիզոնական 1—11 թվային շարքը ցույց է տալիս բուլղարերեն բառերի մեջ եղած վանկերի քանակը, իսկ աղյուսակի կողքի ուղղահայաց 1—11 թվային շարքը ցույց է տալիս հայերեն համապատասխան բառերի մեջ եղած վանկերի քանակը:

Ընդգծված բուլղարերեն և հայերեն նույն բառի վանկերի քանակային համապատասխանությունը 3/2 է: Այս բառի տեղը աղյուսակում $P_{(3,2)}$ քառա-

կուսիում է: Ինչպես տեսնում ենք այդ քառակուսիում նշված է 40 թիվը: Իսկ նշանակում է, որ մեր ուսումնասիրած բառերի քանակի մեջ բացի դրանից կան նաև 39 նման դեպք:

¹ ²
Բուլղ. се-ло
¹ ²
Հայ. գյուղը

Համապատասխանությունը 2/2 է: Աղյուսակում այդ բառի տեղը $P_{(2,2)}$ քառակուսիում է: Ինչպես տեսնում ենք այդ քառակուսիում նշված է 75 թիվը: Իսկ նշանակում է, որ մեր ուսումնասիրած բառերի քանակի մեջ բացի դրանից կան նաև 74 նման դեպք:

¹ ² ³
Բուլղ. при-тих-на
¹ ²
Հայ. լը-ոհեց

Վանկերի համապատասխանությունը դարձյալ 3/2 է: Աղյուսակում սրա տեղը ընկնում է $P_{(3,2)}$ քառակուսին: Վերևում մենք նշեցինք, որ բացի վերևի օրինակից կան նաև 39 այլ օրինակներ: Այս վերջինս մտնում է այդ 39-ի մեջ:

2. Какво си се окупил такъв? = Էդ ո՞րտեղից բար

¹ ² ³ ⁴ ⁵ ⁶ ⁷ ⁸ ⁹
Как-во си се о-ку-мил та-къв
¹ ² ³ ⁴ ⁵ ⁶
Էդ որ-ան-դից բը-սար

Այստեղ մի այլ երևույթ. բառացիորեն բուլղարերենից հայերեն թարգմանել ճիշտ չի լինի: Այս պատճառով բուլղարերեն բառերի կապակցությունը թարգմանելիս համապատասխանում է հայերեն բառերի մի այլ կապակցության: Վերևի օրինակում բուլղարերեն 5 բառին համապատասխանում են հայերեն 3 բառ: Դրանց վանկային համապատասխանությունը 9/6 է, որը աղյուսակում գրավում է $P_{(9,6)}$ քառակուսին: Այնտեղ նշված թիվը ցույց է տալիս, որ նման երևույթ (դեպք) ուսումնասիրված տեքստի մեջ 1 հատ է:

3. Лицето му застина в сурова размисъл =
= Ինքրուջացու դեմքը միջնեց ու սառեց

Այստեղ բուլղարերեն երկու բառին (¹ ² ³ ⁴ $ли-це-то му$) համապատասխանում է հայերեն 1 բառ (¹ ² $դեմ-քը$): Տեղը աղյուսակում կլինի $P_{(4,2)}$ քառակուսին: Բառային համապատասխանության մեջ անտեսվել է Ինքրուջացու բառը, նկատի ունենալով նրա հատուկ անվան նշանակությունը:

Վերևում ասացինք, որ ուսումնասիրության համար հենվել ենք կատարված 2 փորձերի տվյալների վրա. այդ փորձերն էին ա և զ:

Փորձի անորոշության չափը ընդունենք $\lg p$, որտեղ p -ն ա փորձի հավասարահավանական էլքն է:

p_K -ն արտահայտված է հետևյալ կերպ.

(1) $p_K = \frac{n_K}{N}$, որտեղ K -ն աղյուսակի հորիզոնական և ուղղահայաց շարքերում շարժվում է 1—11 ինտերվալում:

ա փորձում n_K -ն համապատասխանորեն լինելու է. $n_1=126$, $n_2=191$, $n_3=177$, $n_4=105$, $n_5=47$, $n_6=17$, $n_7=5$, $n_8=5$, $n_9=4$, $n_{10}=1$, $n_{11}=3$:

α փորձի համար $n_1 + n_2 + \dots + n_{11} = 681$:

β փորձի n_k -ն լինելու է.

$n_1 = 104, n_2 = 173, n_3 = 181, n_4 = 112, n_5 = 64, n_6 = 27, n_7 = 10, n_8 = 4, n_9 = 2, n_{10} = 3, n_{11} = 1$:

β փորձի համար $n_1 + n_2 + \dots + n_{11} = 681$:

N-ը երկու փորձում էլ (α և β) 681 է, որը բուլղարերեն տեքստի 1040 բառերի համապատասխանությունն է հայերենում:

Անորոշության չափը հավասար է՝

(2) $H_{(\alpha)} = - \sum_k p_k \lg p_k$, որտեղ p_k մեղ հայտնի է (1) բանաձևից:

էնտրոպիան հավասար է p_k և $\lg p_k$ արտադրյալների գումարին:

α փորձի համար էնտրոպիան հավասար է $\approx 0,7543$

β փորձի համար էնտրոպիան հավասար է $\approx 0,432$

Կախյալ էնտրոպիան երկու լեզուների համար հավասար է՝

(3) $H_{(\alpha, \beta)} = H_{(\alpha)} + H_{(\beta)}$

Մեր փորձում

$$H_{(\alpha, \beta)} \approx 1,142,$$

որտեղ հաշվված է № 2 և 3 աղյուսակների հիման վրա:

$$\text{№ 2 աղյուսակում } P_{(1,1)} = \frac{68}{681} = 0,08$$

Նույն ձևով համապատասխանաբար հաշվված են նաև $P_{(1,2)}, \dots$ ընդամենը 5-ը տարբեր $P_{(1,1)}$ հանդիսանում է միավանկ բառերի համատեղելի հավանականությունը:

$P_{(1,2)}$ հանդիսանում է միավանկ բուլղարական և երկվանկ հայկական բառերի հավանականությունը:

Այնուհետև որոշված են դրանց լոգարիթմները: Այսպես, օրինակ՝ $P_{(1,1)}$ (2-րդ աղյուսակում) հավասար է 0,08, իսկ սրա լոգարիթմը 3-րդ աղյուսակում հավասար է 1,1: Սրանից հետո 2-րդ բանաձևով որոշված է $H_{(\alpha, \beta)}$:

Հայտնի է, որ լեզուների ցեղակցության չափը (4) հավասար է $R = H_{(\alpha)} + H_{(\beta)} - H_{(\alpha, \beta)}$, որտեղ R -ը կոչված է ավելցուկություն (избыточность) և քանակապես հավասարվում է ինֆորմացիայի միջին թվին, հաղորդված թարգմանության միջոցով: Ավելցուկությունը այնքան շատ է, որքան շատ է տվյալ զույգ լեզուների մոտիկությունը (նրանց կոռելյացիայի առտիճանը): Համեմատելու համար՝

Ռուս-անգլերեն զույգի համար $R \approx 0,17$ ինֆորմացիայի կրկնակի միավոր:

Բուլղար-հայերեն զույգի համար $R \approx 0,7543 + 0,432 - 1,142$

$R \approx 0,0443$ ինֆորմացիայի տասնապատիկ միավոր $= 0,1328$ ինֆորմացիայի կրկնակի միավոր:

Պետք է նշել, որ համեմատելով հայկական թարգմանությունը բուլղարերենի հետ, պարզվեց, որ թարգմանողը թույլ է տվել սխալներ, ինչպես օրինակ՝

погледна — նստեց

պետք է լինի նայեց

կամ чисто — սուր

պետք է լինի մաքուր

Ի՞ր ոչ ճիշտ թարգմանությունը Ավագյանը փոխել է թարգմանված նախադասությունների միտքը:

Այս գործում լեզվաբանական որոշ տերմիններ ստացել են նոր իմաստ: Իհարկե, կարելի է ստեղծել դրանց փոխարեն նոր տերմիններ:

Այդ ցույց կոտա ապագան:

Эдуард Селян

Мера неопределенности (энтропия) и степень близости (коррелированность) болгарского и армянского языков

(Р е з ю м е)

Машинный перевод является одним из самых важных вопросов математической лингвистики. При переводе с одного языка на другой появляется двусмыслие, неопределенность. Делается невозможным вопрос автоматизации перевода. В этом отношении полезно количественно оценивать близость различных пар языков или взаимоотношение (коррелированность). Энтропия является мерой неопределенности и выражается как функция вероятностей или $f(p)$.

Сделаны два опыта: α с болгарским языком и β (с армянским языком). Для сравнения (таблица 1) был использован роман болгарского писателя Манова Э. „Краят на Делиите“, переведенный с оригинала на армянский язык Авакяном. Общее количество переведенных болгарских слов—1040, которым соответствует 681 армянское слово.

Получено, что энтропия опыта α ($H_{(\alpha)} = - \sum_k p_k \lg p_k$) равна $\approx 0,7543$ десят. ед. инф.

Опыта $\beta \approx 0,432$ десят. ед. инф. Зависимая энтропия двух языков равна:

$$H_{(\alpha, \beta)} = H_{(\alpha)} + H_{\alpha(\beta)}$$

в нашем опыте $\approx 1,142$ десят. ед. инф.

Откуда коррелированность будет: $R = H_{(\alpha)} + H_{\alpha(\beta)} - H_{(\alpha, \beta)}$ или $\approx 0,0443$ десят. ед. инф., которые равняются 0,1328 двоич. ед. инф.