International Journal of Social Science And Human Research

ISSN(print): 2644-0679, ISSN(online): 2644-0695

Volume 05 Issue 10 October 2022

DOI: 10.47191/ijsshr/v5-i10-09, Impact factor- 5.871

Page No: 4476-4483

Degree of Lexical Similarity between English and Kurdish

Mohammed Khalid Mahmood

Department of Health & Biomedical Sciences, Aix-Marseille University, France



ABSTRACT: Indo-European languages are the native languages of the habitants of south and west Eurasia. It is the largest spoken language family in the world with 3.5 billion speakers, corresponding to 46% of the world population. Kurdish and English are genetically related, belong to the same family branch of languages, and are believed to have evolved from a common proto-language. Lexical Similarity measures the similarity and/or difference between a set of words from any given two languages. Despite the abundance of lexical similarity coefficients between various world languages in the literature, there are no available data on Kurdish, even though many Kurdish natives learn English as L2. The objective of this paper is to estimate the lexical similarity between two remotely related Indo-European languages. Our results showed (8.75%) and (9.75%) lexical similarity between English and Kurdish according to the Leipzig-Jakarta and Swadesh lists respectively; giving the average of (9.25%). In conclusion, English and Kurdish languages are related Indo-European languages and the lists used in this research are proven to be reliable in the comparison between the studied languages. These cognates can be helpful for the learners and teachers of Kurdish and Kurdish in their language acquisition process.

KEYWORDS: English, Kurdish, Lexical Similarity, Indo-European, Indo-Iranian

INTRODUCTION

Indo-European languages are the native languages of the habitants of south and west Eurasia. It is the largest spoken language family in the world with 3.5 billion speakers, corresponding to 46% of the world population in a vast geography extending from Indian subcontinent to the Iranian plateau and Europe. Indo-European languages are believed to have evolved from a single common ancestor called proto-Indo-European (PIE). It has been hypothesized that PIE was spoken 4500-2500 years BC during the Neolithic and Bronze ages mainly somewhere around the Caspian region in Eastern Europe (Ramat & Ramat, 2015).

Kurdish language belongs to the western Iranian group of the Indo-Iranian branch of the Indo-European family. More than 40 million Kurds communicate in Kurdish in four major countries: Turkey, Iraq, Iran and Syria. The main dialects include Northern (also known as Kurmanji), Central (also known as Sorani) and Southern. On the other hand, English belongs to the West Germanic group of the Germanic branch of the Indo-European family of languages. Kurdish and English are genetically related, belong to the same family branch of languages, and are believed to have evolved from a common proto-language (Mallory & Adams, 2006).

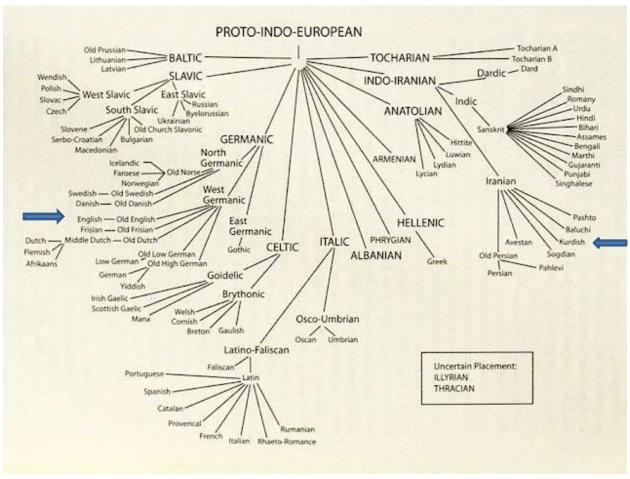


Figure 1: Relative positions of Kurdish and English languages

Lexical Similarity (LS) measures similarity and/or difference between a set of words from any given two languages. A lexical similarity of 1 (or 100%) means a total resemblance between vocabularies, whereas 0 means there are no common words; hence, ranging from mutual intelligibility to lack of relation. (Do et al., 2009).

There are several ways to estimate LS between two certain languages. Historically, various lists have been proposed by experts like (Swadesh, 1955), (Bender, 1975) and (Haspelmath, 2008).

Similar words (cognates) may appear among different languages mainly due to:

- 1. Etymological roots from a common parent language.
- 2. Simple borrowing between languages.
- 3. Biological inclination and preferences: Such as the similarity of the words for 'mother' and 'father' in many languages of the world. Since certain sounds are easier to pronounce by babies, they are often used to give the name to the two lovely creatures who are always there for the child.
- 4. Possible connection between sound and word meanings. Interestingly enough, a recent study shows a positive correlation between these two, although classical linguistics believed that there is no apparent relation between sound and meaning. Researchers have studied a large data of 6000 languages including two thirds of the world's languages. It appears that different and unrelated languages tend to use the same sets of sounds for certain words more than others (Blasi et al., 2016).

The lists prepared for the purpose of measuring lexical similarity depend on lexical cognates. Scientists have different opinions on the nature and definition of cognates; some believe that an orthographic, phonetic and semantic similarity is enough for two pairs to be defined as cognates; however, others seek also for an etymological condition from a parent language (Otwinowska, 2015).

For this reason, the main characteristic of the lists proposed for lexical similarity is that they contain words that are most resistant to change over time and least probable that they have been borrowed from other languages. This explains why name of the body parts, name of animals, name of natural elements of the environment and the basic daily life verbs comprise the majority of these lists. Because

these names were necessary for the first speakers from the very start of any language and they had plenty of time to deepen their roots in the culture; therefore, they are less prone to changes (Holman et al., 2011).

There are plenty of papers in the literature on the lexical similarity between languages, especially concerning English. Table 1 shows the lexical comparison between several Indo-European languages. The objective of this paper is to estimate the lexical similarity between two remotely related Indo-European languages: Kurdish and English; a task which is never done before. Hence, the subject of this research is necessary and useful in one hand, original and unprecedented on the other hand.

Table 1: Lexical similarity between selected Indo-European languages

					Lex	ical similarity	coefficient	s			
	Catalan	English	French	German	Italian	Portuguese	Romanian	Romansh	Russian	Sardinian	Spanish
Catalan	1	-	0.85	(523)	0.87	0.85	0.73	0.76		0.75	0.85
English	-	1	0.27	0.60	*			-	0.24	-	(341)
French	0.85	0.27	1	0.29	0.89	0.75	0.75	0.78	-	0.80	0.75
German		0.60	0.29	1	*		*				3#3
Italian	0.87		0.89	3(5)	1		0.77	0.78		0.85	0.82
Portuguese	0.85		0.75	R#	•	1	0.72	0.74	-	**	0.89
Romanian	0.73	-	0.75	0.50	0.77	0.72	1	0.72	-	0.74	0.71
Romansh	0.76	3	0.78		0.78	0.74	0.72	1		0.74	0.74
Russian	V.S	0.24	:	7/20					1	520	714
Sardinian	0.75		0.80	85 2 0	0.85	-	0.74	0.74	-	1	0.76
Spanish	0.85		0.75	848	0.82	0.89	0.71	0.74		0.76	1

(**Source:** *Ethnologue.com*)

METHOD

Several methods have been used to compute lexical similarity between languages by (Holmes & Ramos, 1993), (Kessler, 1995) and (List, 2012). The main idea is presented as the following: "Cognate identification is usually based on a similarity or distance score calculated from the number of matches and mismatches in the alignment" (List, 2012). The scoring system of this research is mainly based on the Levenshtein algorithm by the (Kessler, 1995).

In this survey, the 100 words of Leipzig-Jakarta list (Haspelmath, 2008) by the Max Planck Digital Library and the 100 words of the famous Swadesh list (Swadesh, 1955) are used as references. There is a 38% difference between the two lists and Only 62 items on the Leipzig–Jakarta list are also present in the Swadesh list.

For Kurdish, two main dialects of North and Central Kurdish which are spoken by the majority of the native Kurds are used in the comparison. The North dialect (also known as Kurmanji) already uses an adapted Latin alphabet; however, the Central dialect (also known as Sorani) uses an adapted Arabic alphabet. For the purpose of easier visual comparison, words of the Sorani dialect are also converted into the Latin alphabet. Table 2 shows information on the adapted Latin alphabet of Kurdish language.

Table 2: Main characteristics of the adapted Kurdish alphabet

Kurdish	Correspondent in English
A	Long a like in art
b, d, f, g, h, k, l, m, n, o, p, r, s, t, u, v, w, y, z	Same as English
С	like /dj/ in June
Ç	/ch/ like in chain
Е	short a like /a/ in any
Ê	/e/ like in leg
i	like the short unwritten vowel between s and t in star
î	/i/ like in list
j	/zh/ like in the French j
q	[q] like Arabic (ق)

ş	/sh/ like in shame
X	[kh] like Arabic (さ)

After enlisting the words, the sound and the orthography are compared. Five categories of scorings are applied. (0) denotes no resemblance between the two pairs (Ex: Tongue/ziman). (0.25) is the score for slight similarity between the words, accounting for one or two letters of the word according to the word size (Ex: Root/reg). (0.5) is the score for a similarity that exist in half of the word or the sound (Ex: You/tû). (0.75) is for an apparent resemblance between the sound of the pairs with a slight difference (Ex: Navel/navik). Score of (1) denotes a total similarity between the words; which is not observed in our study.

In some cases, the resemblance in one of the dialects was clearer and closer as compared to its counterpart dialect in relation to English. In this case, the one with the higher level of similitude is used as reference in the comparison and taken into parenthesis.

RESULTS

20 words in the Leipzig-Jakarta list and 25 words in the Swadesh list revealed some kind of similarity. Table 2 and 3 show those words that have shown a positive signal according to the Leipzig-Jakarta list and the Swadesh list respectively. The complete lists are present in appendix A and B.

Table 3: Words with a positive score according to Leipzig-Jakarta list (Total score: 8.75)

No	English	Kurdish	Score	No.	English	Kurdish	Score
		North/Central				North/Central	
1.	Fire	agir	0.25	11	Navel	(Navik)/nawik	0.75
2.	Nose	(poz)/lwt	0.5	12	To stand	(Standin)/westan	0.75
3.	You	(tû)/to	0.5	13	new	Nşh/(nwê)	0.5
	(singular)						
4.	Root	reh/reg	0.25	14	not	Na, ne	0.25
5.	Rain	Baran	0.5	15	liver	Ceger/ciger	0.5
6.	I/me	Ez/(min)	0.5	16	To run	(Revîn)/rakirdin	0.5
7.	Name	Nav/naw	0.5	17	Tail	(Terrî)/kilk	0.25
8.	Far	dûr	0.25	18	Dog	Seh/(seg)	0.5
9.	Leg/foot	(Ling)/qaç, pê	0.5	19	star	Stêrk/estêre	0.5
10.	egg	(Hêk)/hêlke	0.25	20	To chew	Cûtin/(ciwîn)	0.25

Table 4: Words with a positive score according to Swadesh list (Total score: 9.75)

No	English	Kurdish	Score	No.	English	Kurdish	Score
		North/Central				North/Central	
1.	I/me	Ez/(min)	0.5	14	knee	Çog/(ejno)	0.25
2.	You	(tû)/to	0.5	15	liver	Ceger/ciger	0.5
	(singular)						
3.	not	Na, ne	0.25	16	To kill	kuştin	0.25
4.	Two	du	0.5	17	to fly	firîn	0.25
5.	Woman	jin	0.25	18	To stand	(Standin)/westan	0.75
6.	man	(Mêr)/piyaw	0.25	19	moon	Hêv/(mang)	0.5
7.	Dog	Seh/(seg)	0.5	20	star	Stêrk/estêre	0.5
8.	Root	reh/reg	0.25	21	Rain	Baran	0.5
9.	egg	(Hêk)/hêlke	0.25	22	Fire	agir	0.25
10.	Tail	(Terrî)/kilk	0.25	23	cold	Sarma/(sard)	0.25
11	Nose	(poz)/lwt	0.5	24	new	Nşh/(nwê)	0.5
12	Nail (of	Nenûk/nînok	0.25	25	Name	Nav/naw	0.5
	finger)						
13	Leg/foot	(Ling)/qaç, pê	0.5				

Regarding the scoring calculation technique, previous researches and the data from Ethnologue.com have calculated the number of words with positive lexical similarity (regardless their degree) divided by the whole number of words studied in the survey. Hence,

according to this approach, the degree of lexical similarity between Kurdish and English is (20%) and (25%) according to the reference lists.

However, this is an exaggerated result, especially when we compare it with the previous results. For example, (Maldonado García & Borges de Souza, 2014) have computed the degree of lexical similarity between Portuguese and English to be (20%) according to the previously described calculation technique. This may be true for such languages with Latin roots and strong cultural and historical interactions.

For this reason, the author of this paper has chosen a rather different methodology for the calculation, since this is the first published paper on the lexical comparison concerning an Indo-Iranian language. In this approach, the degree of similarity is taken into consideration and the summations of the positive scores are taken as the reference. Eventually, a total degree of (8.75%) and (9.75%) similarity is observed between Kurdish and English according to the Leipzig-Jakarta list and the Swadesh list respectively; giving the average of (9.25%).

CONCLUSION AND PRACTICAL IMPLICATIONS

In this study, for the first time in the literature, Kurdish and English languages are studied in relation to lexical similarity. Result of this paper (9.25%) show traces of common genetic roots between these two languages. In addition, this result proves for one more time the reliability of the Leipzig-Jakarta and Swadesh lists in lexical comparison. Since (~10%) appears logical and high enough to assume a common ancestor for the two languages but it is smaller than the results that are obtained from comparing English to closer languages like Germanic, Romance and even Slavic.

Being aware of this kind of similarities by learners and teachers of English and Kurdish languages as L2, may help them in their task, make them understand better the logic between the two languages and motivate their curiosity along the path of learning and teaching.

LIMITATIONS

The used reference lists do not account for synonyms. In our study, we have chosen the item with the highest level of resemblance in the comparison. Therefore, the result should be seen as maximum and the actual outcome is expected to be lower if all the synonyms are included.

REFERENCES

- 1) Bender, M. L. (1975). Toward a Lexicostatistic Classification of Ethiopian Languages. Bynon and Bynon, eds, 377-388.
- 2) Blasi, D. E., Wichmann, S., Hammarström, H., Stadler, P. F., & Christiansen, M. H. (2016). Sound–meaning association biases evidenced across thousands of languages. *Proceedings of the National Academy of Sciences*, 113(39), 10818-10823.
- 3) Do, Q., Roth, D., Sammons, M., Tu, Y., & Vydiswaran, V. (2009). Robust, light-weight approaches to compute lexical similarity. *Computer Science Research and Technical Reports, University of Illinois*, 9.
- 4) Haspelmath, M. (2008). Loanword typology: Steps toward a systematic cross-linguistic study of lexical borrowability. *Empirical Approaches to Language Typology*, 35, 43.
- 5) Holman, E. W., Brown, C. H., Wichmann, S., Müller, A., Velupillai, V., Hammarström, H., Sauppe, S., Jung, H., Bakker, D., & Brown, P. (2011). Automated dating of the world's language families based on lexical similarity. *Current Anthropology*, 52(6), 841-875.
- 6) Holmes, J., & Ramos, R. G. (1993). False friends and reckless guessers: Observing cognate recognition strategies. *Second language reading and vocabulary learning*, 86-108.
- 7) Kessler, B. (1995). Computational dialectology in irish gaelic. arXiv preprint cmp-lg/9503002.
- 8) List, J.-M. (2012). LexStat: Automatic detection of cognates in multilingual wordlists. Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH,
- 9) Maldonado García, M. I., & Borges de Souza, A. M. (2014). Lexical similarity level between English and Portuguese. *ELIA*, 14, 145-163.
- 10) Mallory, J. P., & Adams, D. Q. (2006). *The Oxford introduction to proto-Indo-European and the proto-Indo-European world*. Oxford University Press on Demand.
- 11) Otwinowska, A. (2015). Cognate vocabulary in language acquisition and use. Multilingual Matters.
- 12) Ramat, A. G., & Ramat, P. (2015). The Indo-European Languages. Routledge.
- 13) Swadesh, M. (1955). Towards greater accuracy in lexicostatistic dating. *International journal of American linguistics*, 21(2), 121-137.

Appendix A

Leipzig-Jakarta List

No	English	Kurdish	Score	No.	English	Kurdish	Score	No.	English	Kurdish	Score
	_	North/Central				North/Central			_		
1.	Fire	agir	0.25	11	Bone	hestî/êsk	0	21	Night	Şev/şew	0
2.	Nose	(poz)/lwt	0.5	12	Breast	pêsîr/Memk	0	22	Ear	Gûh/gwê	0
3.	To go	çûn	0	13	Rain	Baran	0.5	23	Neck	Hûstû/mil	0
4.	Water	av/aw	0	14	I/me	Ez/(min)	0.25	24	Far	dûr	0.25
5.	Mouth	dev/dem	0	15	Name	Nav/naw	0.5	25	To do	Kirin/kirdin	0
6.	Tongue	ziman	0	16	Louse	Spî/espê	0	26	House	Xanî/xanu	0
7.	Blood	xwîn/xwên	0	17	Wing	bask/bal	0	27	Stone	Kevir/berd	0
8.	You	(Tû)/to	0.5	18	Flesh/meat	goșt	0	28	Bitter	tal	0
	(singular)										
9.	Root	reh/reg	0.25	19	hand	dest	0	29	To say	Gotin/witin	0
10.	To come	hatin	0	20	fly	mêş	0	30	tooth	Diran/dan	0

N o	Englis h	Kurdish North/Central	Scor e	No ·	English	Kurdish North/Central	Scor e	No ·	Englis h	Kurdish	Scor e
31	Hair	Por/qij, mu	0	41	yesterd ay	Do/dwênê	0	51	Child	Zarok/minal	0
32	Big	Mezin/gewre	0	42	To drink	Vexwarin/xwardin ewe	0	52	egg	(Hêk)/hêlke	0.25
33	One	yek	0	43	Black	reș	0	53	To give	Bidin/pêdan	0
34	Who?	Kî/kê	0	44	Navel	(Navik)/nawik	0.75	54	new	Nşh/ nwê	0.5
35	He/she/ it	ew	0	45	To stand	(Standin)/westan	0.75	55	To burn	Şewitandin/sutan din	0
36	To beat/hit	lêdan	0	46	To bite	Devlêkirin/gazgirti n	0	56	not	Na, ne	0.25
37	Leg/fo ot	(Ling)/qaç, pê	0.5	47	Back	Paş/dwa	0	57	good	baş	0
38	Horn	Qiloç/şax	0	48	Wind	ba	0	58	to know	zanîn	0
39	This	ev/em	0	49	Smoke	Dixan/dukel	0	59	knee	Çog/(ejno)	0.25
40	fish	masî	0	50	What?	çi	0	60	sand	Xiz/lim	0
N o	Englis h	Kurdish North/Central	Scor e	No ·	English	Kurdish North/Central	Scor e	No ·	Englis h	Kurdish	Scor e
61	To laugh	Kenîn/pêkenîn	0	71	ant	Gêrîk/mêrule	0	81	To run	(Revîn)/rakirdin	0.5
62	To hear	bîstin	0	72	heavy	Giran/qurs	0	82	To fall	Ketin/kewtin	0
63	soil	Erd/xak, zewî	0	73	wood	Text/texte	0	83	Eye	Çav/ çaw	0
64	leaf	Pel/gela	0	74	To take	Bigirin/birdin	0	84	Ash	Xwelî/xolemêş	0
65	red	Sor/sur	0	75	old	Kevn/kon	0	85	Tail	(Terrî)/kilk	0.25
66	liver	Ceger/(ciger)	0.5	76	To eat	Xwarin/xwardin	0	86	Dog	Seh/(seg)	0.5
67	To hide	Veşartin/şardine we	0	77	thigh	Çîp/ran	0	87	To cry	Girîn/giryan	0
68	skin	Çerm/pêst	0	78	Thick	estur	0	88	To tie	girêdan	0
69	To suck	Mêtin/mijîn	0	79	Long	dirêj	0	89	To see	Dîtin/bînîn	0

70	To	helgirtin	0	80	To	Nepixandin/futêkir	0	90	sweet	şîrîn	0	l
	carry				blow	din						l

No	English	Kurdish	Score
		North/Central	
91	rope	Werîs/guris	0
92	shadow	Sî/sêber	0
93	bird	Teyr/balinde	0
94	Salt	xwê	0
95	Small	biçûk	0
96	wide	pan	0
97	star	Stêrk/estêre	0.5
98	in	Li/le	0
99	hard	req	0
100	to chew	Cûtin/(ciwîn)	0.25

Appendix B Swadesh List

N o	English	Kurdish North/Centra	Scor e	No ·	Englis h	Kurdish North/Centra l	Scor e	No ·	English	Kurdish North/Centra	Scor e
1.	I/me	Ez/(min)	0.25	11	One	yek	0	21	Dog	Seh/(seg)	0.5
2.	You (singular	(Tû)/to	0.5	12	Two	du	0.5	22	Louse	Spî/espê	0
3.	We	Em/ême	0	13	Big	Mezin/gewre	0	23	tree	dar	0
4.	This	ev/êm	0	14	Long	dirêj	0	24	seed	Toxim/tow	0
5.	That	Va/ew	0	15	Small	biçûk	0	25	leaf	Pel/gela	0
6.	Who?	Kî/kê	0	16	Woma n	jin	0.25	26	Root	reh/reg	0.25
7.	What?	Çi	0	17	man	(Mêr)/piyaw	0.25	27	bark	Ewtînî/werre	0
8.	Not	Na, ne	0.25	18	person	kes	0	28	skin	Çerm/pêst	0
9.	All	Gişt/hemu	0	19	fish	masî	0	29	Flesh/mea t	goșt	0
10	Many	Gelek/zor	0	20	bird	Teyr/balinde	0	30	Blood	xwîn/xwên	0

N o	Englis h	Kurdish North/Centr al	Scor e	No ·	English	Kurdish North/Centr al	Scor e	No ·	Englis h	Kurdish North/Central	Scor e
31	Bone	Hestî/êsk	0	41	Nose	(poz)/lwt	0.5	51	Breast	pêsîr/Memk	0
32	Grease (fat)	Rûn/ron	0	42	Mouth	dem	0	52	heart	dil	0
33	Egg	(Hêk)/hêlke	0.25	43	tooth	Diran/dan	0	53	liver	Ceger/(ciger)	0.5
34	Horn	Qiloç/şax	0	44	Tongue	ziman	0	54	To drink	Vexwarin/xwardine we	0
35	Tail	(Terrî)/kilk	0.25	45	Nail (of finger)	Nenûk/nînok	0.25	55	To eat	Xwarin	0
36	feather	Perrîk/perr	0	46	Leg/foo t	(Ling)/qaç, pê	0.5	56	To bite	Devlêkirin/gazgirtin	0
37	Hair	Por/qij, mu	0	47	knee	Çog/(ejno)	0.25	57	To see	Dîtin/bînîn	0

38	Head	ser	0	48	hand	dest	0	58	То	bîstin	0
									hear		
39	Ear	Gûh/gwê	0	49	belly	Zik/sik	0	59	to	zanîn	0
		_			-				know		
40	Eye	Çav/ çaw	0	50	Neck	Hûstû/mil	0	60	То	Xewîn/xewtin	0
									sleep		

N o	Englis h	Kurdish North/Central	Scor e	No ·	English	Kurdish North/Cent ral	Scor e	No ·	English	Kurdish North/Central	Scor e
61	to die	Mirin/mirdin	0	71	To say	Gotin/witin	0	81	Smoke	Dixan/dukel	0
62	To kill	kuştin	0.25	72	sun	xor	0	82	Fire	agir	0.25
63	To swim	(Soberîkirin)/meleki rdin	0	73	moon	Hêv/(mang)	0.5	83	Ash	Xwelî/xolemêş	0
64	to fly	firîn	0.25	74	star	Stêrk/estêre	0.5	84	To burn (intransiti ve)	Şewitandin/suta ndin	0
65	Walk	Çûyin/roiştin	0	75	Water	av/aw	0	85	path	rê	0
66	To come	hatin	0	76	Rain	Baran	0.5	86	mountain	Çîya/Şax	0
67	To lie (on side, reclin e)	Razekirin/rakşan	0	77	Stone	Kevir/berd	0	87	red	Sor/sur	0
68	To sit	Rûniştin/daniştin	0	78	sand	Xiz/lim	0	88	green	Kesk/sewz	0
69	To stand	(Standin)/westan	0.75	79	Earth/s oil	Erd/xak, zewî	0	89	Yellow	zerd	0
70	To give	Bidin/pêdan	0	80	cloud	Ewr/hewr	0	90	white	spî	0

No	English	Kurdish	Score
		North/Central	
91	Black	reș	0
92	Night	Şev/şew	0
93	hot	germ	0
94	cold	Sarma/(sard)	0.25
95	full	Tije/pir	0
96	new	Nşh/(nwê)	0.5
97	good	baş	0
98	round	xir	0
99	dry	Zûha/wişk	0
100	Name	Nav/naw	0.5



There is an Open Access article, distributed under the term of the Creative Commons Attribution–Non Commercial 4.0 International (CC BY-NC 4.0)

(https://creativecommons.org/licenses/by-nc/4.0/), which permits remixing, adapting and building upon the work for non-commercial use, provided the original work is properly cited.