Can Linguistic Distance help Language Classification? Assessing Hawrami-Zaza and Kurmanji-Sorani

Hossein Hassani
University of Kurdistan Hewlêr
Kurdistan Region - Iraq
hosseinh@ukh.edu.krd

Abstract

To consider Hawrami and Zaza (Zazaki) standalone languages or dialects of a language have been discussed and debated for a while among linguists active in studying Iranian languages. The question of whether those languages/dialects belong to the Kurdish language or if they are independent descendants of Iranian languages was answered by MacKenzie (1961). However, a majority of people who speak the dialects are against that answer. Their disapproval mainly seems to be based on the sociological, cultural, and historical relationship among the speakers of the dialects. While the case of Hawrami and Zaza has remained unexplored and under-examined, an almost unanimous agreement exists about the classification of Kurmanji and Sorani as Kurdish dialects. The related studies to address the mentioned cases are primarily qualitative. However, computational linguistics could approach the question from a quantitative perspective. In this research, we look into three questions from a linguistic distance point of view. First, how similar or dissimilar Hawrami and Zaza are, considering no common geographical coexistence between the two. Second, what about Kurmanji and Sorani that have geographical overlap. Finally, what is the distance among all these dialects, pair by pair? We base our computation on phonetic presentations of these dialects (languages), and we calculate various linguistic distances among the pairs. We analyze the data and discuss the results to conclude[1].

1 Introduction

Kurdish is an Indo-European language that specifically belongs to the Iranian languages, and more particularly to Western-Iranian (Kreyenbroek, 2005), or North-Western-Iranian languages (Minorsky, 1945; Phillipson and Skutnabb-Kangas, 1996). The language is multi-dialect (Khalid, 2015), but there is no consensus about the dialects among the scholars. While according to MacKenzie (1961), Hawrami and Zaza (Zazaki) are standalone languages that he considers directly under Western-Iranian languages, (Hassanpour, 1992) addresses them as dialects of Kurdish. The question of which one of the *commonly* known dialects are should be recognized as Kurdish from a linguistic perspective has been an ongoing debate among the scholars who have studied the language for a long while. Importantly, a majority of people who speak the dialects believe that all of those dialects/languages are Kurdish. Their belief mainly seems to be based on the sociological, cultural, and historical relationship among the speakers of the dialects.

While the case of Hawrami and Zaza has remained unexplored and under-examined, an almost unanimous agreement exists about the classification of Kurmanji and Sorani as Kurdish dialects. The related studies to address the mentioned cases are primarily qualitative. However, computational linguistics could approach the question from a quantitative perspective.

In this research, we look into three questions from a linguistic distance point of view. First, how similar or dissimilar Hawrami and Zaza are, considering no common geographical coexistence between the two. Second, what about Kurmanji and Sorani that have geographical

[1] This paper is based on an abstract that was submitted to and presented under the same title at the 5^{th} *International Conference on Kurdish Linguistics* at the University of Graz, 25 September 2021.

overlap. Finally, what is the distance among all these dialects, pair by pair? We base our computation on phonetic presentations of these dialects (languages), and we calculate various linguistic distances among the pairs.

The rest of this paper is organized as follows. Section 2 reviews the related work. In Section 3, we present the method that we follow to compute the linguistic distance. We report and discuss the results in Section 4. Finally, Section 5 concludes the paper and suggests some areas for future work.

2 Related Work

MacKenzie (1961) set the basis of Kurdish language classification. Their work is still one of the essential references for Kurdish studies. Their qualitative approach was combined with historical studies of Kurdish development by other scholars such as Hassanpour (1992). While MacKenzie (1961) classified Kurmanji and Sorani under Kurdish stem, he considered Hawrami and Zaza as descendants of Iranian languages but not Kurdish.

Minorsky (1943) discussed Gurani from a historical point of view. Although he presented and discussed some literature written in Gurani, his view was primarily based on historical analysis. Minorsky (1943) considered Zaza and “Awrami” (Hawrami) as dialects of Gurani, which was claimed again by some other scholars. All those claims are mostly based on qualitative measures as well. However, several scholars argued against their proposition (see Hassanpour (1993)), 1993, for example). MacKenzie (1991) provided one of the earliest critiques on the matter. The issue was recently addressed again by Haig and Öpengin (2014) and Haig and Khan (2018). But, the classification of those dialects (languages) is still an open question. Regardless of their proposed classifications, so far, Kurdish linguists have followed a qualitative approach, and no quantitative study has been conducted about the subject.

However, research on language classification according to quantitative methods, for example, linguistic distance, has been emerging. To illustrate, Gamallo et al. (2017) classified European languages based on their similarities and divergence according to their linguistic distance. Also, Rama and Borin (2015) suggested a genetic classification of languages based on Swadesh-style core vocabulary. Using Swadesh’s list (Swadesh, 1955) as a basis for language similarity has been practiced for a long while (see, for instance, Holman et al. (2008)). In this study, we also use the Swadesh list as the fundamental instrument.

3 Method

Our method is similar to what is proposed by (Bourgeois-Gironde et al., 2021). We prepare a Swadesh list (a 207 entry version) for Zaza, Hawrami, Kurmanji, and Sorani. We calculate the Jaro similarity/distance between Zaza and Hawrami on the one hand and Kurmanji and Sorani on the other. We also calculate the linguistic distance between all mentioned dialects. We then discuss the finding against the previous qualitative work to investigate the conformance to or divergence of their propositions about the classification of the mentioned dialects (languages).

4 Result and Discussion

We prepare a Swadesh list (a 207 entry version) for each dialect under study[2]. Figure 1 shows the similarities and distances between each pair of the dialects. Figure 2 shows the number of words that were completely similar or were completely different between the Kurdish dialects. Figure 2 shows the number of words that were completely similar or were completely different between the Kurdish dialects.

We observe that Zazaki and Hawrami are less similar than Kurmanji and Sorani, while Zazaki and Hawrami share less number of completely similar words than Kurmanji and Sorani. Also, Hawrami and Sorani are less similar than Kurmanji and Sorani, but the difference is not significant. Zazaki and Sorani are as different as Zazaki and Hawrami, while Hawrami and Kurmanji

[2] The dataset is available at `https://github.com/KurdishBLARK/Dialect-Classification`

Table 1: Similarity/Distance between Kurdish Dialects (Za: Zaza; Hwa: Hawrami; Kur: Kurmanji; Sor: Sorani).

Jaro (Avg)	Za-Haw	Kur-Sor	Za-Kur	Za-Sor	Haw-Kur	Haw-Sor
Similarity	0.57	0.68	0.59	0.58	0.52	0.61
Distance	0.43	0.32	0.41	0.42	0.48	0.39

Table 2: Number of Completely Similar or Different between Kurdish Dialects (Za: Zaza; Hwa: Hawrami; Kur: Kurmanji; Sor: Sorani).

Similarity	Za-Haw	Kur-Sor	Za-Kur	Za-Sor	Haw-Kur	Haw-Sor
Completely Similar	10	55	22	19	7	23
Completely Different	20	23	29	32	37	21

Table 3: Percentages of Completely Similar or Different Words between Kurdish Dialects (Za: Zaza; Hwa: Hawrami; Kur: Kurmanji; Sor: Sorani).

Similarity	Za-Haw	Kur-Sor	Za-Kur	Za-Sor	Haw-Kur	Haw-Sor
Completely Similar	4.83%	26.57%	10.63%	9.18%	3.38%	11.11%
Completely Different	9.66%	11.11%	14.01%	15.48%	17.87%	10.14%

are as similar as they are different, though the similarity weighs more. Finally, the similarity of Zazaki – Kurmanji and Zazaki – Sorani is almost the same.

The results do not suggest that Zazaki and Hawrami are any closer than Kurmanji and Sorani. Therefore, some suggestion that considers them as a single language/dialect cannot be attested. Furthermore, the results show that Hawrami and Kurmanji are quite similar to Kurmanji and Sorani that could open a discussion among the scholars who consider these dialects not to be Kurdish to reconsider the way that they have classified those dialects.

5 Conclusion

In this research, we looked into the similarity and distances of Kurdish dialects. We prepared Swadesh lists (the 207 entry version) for each one of the four Kurdish dialects, namely Kurmanji, Sorani, Hawrami, and Zazaki. To calculate the similarity and difference between each pair of those dialects, we applied the Jaro measure. The results did not suggest that Zazaki and Hawrami are any closer than Kurmanji and Sorani. That suggests a need to investigate and perhaps reconsider the way that those dialects have been classified.

In the future, we would like to expand the dataset beyond the Swadesh list to a larger common vocabulary that could help to obtain a more thorough view of the similarity/distance between Kurdish dialects. Also, we are interested in considering other lists such as Leipzig–Jakarta and comparing the results with the current ones to see whether they are reciprocal or not.

Acknowledgments

My utmost appreciations to Şadê Agon for her generous assistance in preparing Zazaki Swadesh list, Dr. Srwa Rostaminezhad for her generous assistance in preparing Hawrami Swadesh list, Sîpan Monnet, Saman Idrees, and Marwan Ali for their generous assistance in preparing Kurmanji Swadesh list, and Dr. Ergin Öpengin and Dr. Nava Hassani for their help in coordination with the Zazaki and Hawrami informants

References

Bourgeois-Gironde, S., Ginsburgh, V., Hassani, H., and Weber, S. (2021). A Lingua Franca for Kurdish Populations. London, Centre for Economic Policy Research https://cepr.org/active/publications/discussion_papers/dp.php?dpno=16086.

Gamallo, P., Pichel, J. R., and Alegria, I. (2017). From language identification to language distance. *Physica A: Statistical Mechanics and its Applications*, 484:152–162.

Haig, G. and Khan, G. (2018). *The Languages and Linguistics of Western Asia: An Areal Perspective*, volume 6. Walter de Gruyter GmbH & Co KG.

Haig, G. and Öpengin, E. (2014). Introduction to special issue-kurdish: A critical research overview. *Kurdish studies*, 2(2):99–122.

Hassanpour, A. (1992). *Nationalism and language in Kurdistan, 1918-1985*. San Francisco: Mellen Research University Press.

Hassanpour, A. (1993). Kurdish studies: Orientalist, positivist, and critical approaches: Review essay.

Holman, E. W., Wichmann, S., Brown, C. H., Velupillai, V., Müller, A., Bakker, D., et al. (2008). Advances in automated language classification. *Quantitative Investigations In Theoretical Linguistics (QITL3)*, page 40.

Khalid, H. S. (2015). Kurdish dialect continuum, as a standardization solution. *International Journal of Kurdish Studies*, 1(1):27–39.

Kreyenbroek, P. G. (2005). On the kurdish language. In *The Kurds*, pages 62–73. Routledge.

MacKenzie, D. N. (1961). The Origins Of Kurdish. *Transactions of the Philological Society*, 60(1):68–86.

MacKenzie, D. (1991). Compendium Linguarum Iranicarum.

Minorsky, V. (1943). The Guran. *Bulletin of the School of Oriental and African Studies*, 11(Part I):75–103.

Minorsky, V. (1945). The tribes of western Iran. *The Journal of the Royal Anthropological Institute of Great Britain and Ireland*, 75(1/2):73–80.

Phillipson, R. and Skutnabb-Kangas, T. (1996). Colonial language legacies: the prospects for Kurdish. In *Self-determination: international perspectives*, pages 200–213. Palgrave Macmillan.

Rama, T. and Borin, L. (2015). Comparative evaluation of string similarity measures for automatic language classification. In *Sequences in language and text*, pages 171–200. De Gruyter Mouton.

Swadesh, M. (1955). Towards greater accuracy in lexicostatistic dating. *International journal of American linguistics*, 21(2):121–137.