A Lingua Franca for Kurdish Populations

Sacha Bourgeois-Gironde, Victor Ginsburgh, Hossein Hassani, Shlomo Weber

DISCUSSION PAPER SERIES

DP16086

A Lingua Franca for Kurdish Populations

Sacha Bourgeois-Gironde, Victor Ginsburgh, Hossein Hassani and Shlomo Weber

PUBLIC ECONOMICS

CEPR

ISSN 0265-8003

A Lingua Franca for Kurdish Populations

Sacha Bourgeois-Gironde, Victor Ginsburgh, Hossein Hassani and Shlomo Weber

Discussion Paper DP16086
Published 28 April 2021
Submitted 28 April 2021

Centre for Economic Policy Research
33 Great Sutton Street, London EC1V 0DX, UK
Tel: +44 (0)20 7183 8801
www.cepr.org

A Lingua Franca for Kurdish Populations

Abstract

Kurdish languages and multiple dialects spread across several nation-states under various regimes varying from regional recognition (e. g. Iraq) to persistent attrition (e. g. Turkey). Kurdish linguistic faces a variety of challenges which can be attributed to different causes such as the historical background of the language, sociopolitical reasons, and forced compliance with national linguistic policies in some of the countries where Kurds live to name a few. In this paper we do not discuss the normative issue of linguistic rights entitlements of the speakers of different varieties of Kurdish. We consider their complex sociolinguistic situation from the point of view of communication efficiency in the face of the following dilemma: Either unification through the adoption of a lingua franca or standardized Kurdish, with the implication of disenfranchisement of some speakers, or the maintenance of multiple dialects, with the risk of fractionalization and its political and economic consequences. For reasons such as the multi-dialect feature of the language and its sociocultural attributes, the attempts to standardize Kurdish have not succeeded. To address this dilemma, we proceed to compute the lexical-linguistic distances between six dialects of Kurdish: three which are representative of Kurmanji and three of Sorani, i. e. the two main linguistic and regional varieties of Kurdish. Our selection of dialects, although incomplete, covers about 75% of the whole population of Kurdish speakers. Our study is the first one to propose an application of the Jaro-similarity index on a Swadesh-list of dialects of Kurdish. Our results reveal some significant distance within Sorani and Kurmanji dialects, and an expected more significant distance between Sorani and Kurmanji dialects. The latter distance is sufficiently important to favor a three-language policy rather than any other one: an international language, the national language (Turkish, Farsi or Arabic), and the local Kurdish variety. This policy maximizes efficiency, Kurdish identity as well as within and without group intercommunication. We compare it to similar linguistic policy attempts in India, Nigeria and Kazakhstan.

JEL Classification: D63, Z13

Keywords: Kurdish languages, Linguistic Distances, Three-language formula

Sacha Bourgeois-Gironde - sbgironde@gmail.com
Université Paris 2, Département d'Economie and Institut Nicoud

Victor Ginsburgh - vginsbur@ulb.ac.be
ECARES, Free University of Brussels

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

Shlomo Weber - sweber@smu.edu
Southern Methodist University and CEPR

A *Lingua Franca* for Kurdish Populations[*]

Sacha Bourgeois-Gironde
Université Paris 2, Département d'Economie and Institut Nicoud
Département d'études cognitives, ENS, EHESS, CNRS, PSL University

Victor Ginsburgh
ECARES, Université libre de Bruxelles, Brussels, Belgium
CORE, Université catholique de Louvain, Louvain-la-Neuve, Belgium

Hossein Hassani
Department of Computer Science and Engineering, University of Kurdistan Hewlêr,
Kurdistan Region, Iraq

Shlomo Weber
China, Russia, and Eurasia Research Center,
New Economic School, Moscow, Russia

March 11, 2021

[*] We are grateful to Salih Akin, Saman Idrees, Sîpan Monnet and Marwan Ali for their help, and David Laitin for his comments.

Abstract

Kurdish languages and multiple dialects spread across several nation-states under various regimes varying from regional recognition (e. g. Iraq) to persistent attrition (e. g. Turkey). Kurdish linguistic faces a variety of challenges which can be attributed to different causes such as the historical background of the language, sociopolitical reasons, and forced compliance with national linguistic policies in some of the countries where Kurds live to name a few. In this paper we do not discuss the normative issue of linguistic rights entitlements of the speakers of different varieties of Kurdish. We consider their complex sociolinguistic situation from the point of view of communication efficiency in the face of the following dilemma: Either unification through the adoption of a *lingua franca* or standardized Kurdish, with the implication of disenfranchisement of some speakers, or the maintenance of multiple dialects, with the risk of fractionalization and its political and economic consequences. For reasons such as the multi-dialect feature of the language and its sociocultural attributes, the attempts to standardize Kurdish have not succeeded. To address this dilemma, we proceed to compute the lexical-linguistic distances between six dialects of Kurdish: three which are representative of Kurmanji and three of Sorani, i. e. the two main linguistic and regional varieties of Kurdish. Our selection of dialects, although incomplete, covers about 75% of the whole population of Kurdish speakers. Our study is the first one to propose an application of the Jaro-similarity index on a Swadesh-list of dialects of Kurdish. Our results reveal some significant distance within Sorani and Kurmanji dialects, and an expected more significant distance between Sorani and Kurmanji dialects. The latter distance is sufficiently important to favor a three-language policy rather than any other one: an international language, the national language (Turkish, Farsi or Arabic), and the local Kurdish variety. This policy maximizes efficiency, Kurdish identity as well as within and without group intercommunication. We compare it to similar linguistic policy attempts in India, Nigeria and Kazakhstan.

Keywords: Kurdish languages; Linguistic distances; Three-language formula

JEL classification: D63, Z13

The width of a wall matters less than the will to climb over it.
Thucydides in *Peloponnesian War*.

1. Introduction

In this paper, we examine possible scenarios of choosing a consistent linguistic policy for the Kurdish population spread across various countries, some of which limit or at least do not support teaching and/or public usage of Kurdish languages.[1]

In 1925, the region populated by Kurds was essentially divided between four countries: Iran, Iraq, Syria and Turkey (for short, IIST countries). Though Mustafa Kemal had promised the creation of a Muslim state with a peaceful co-existence of Turks and Kurds, whereas the League of Nations and the British had offered a support for an autonomous Kurdish government, the promises remained promises and nothing really happened. Today, the majority of the Kurdish population still lives in IIST countries: Iran (12 million), Iraq (8.5 million), Syria (3.6 million), and Turkey (20 million).[2] A smaller number of Kurds live in other countries, some in the proximity of IIST countries—Armenia (37,000), Azerbaijan (38,000), Georgia (14,000), Israel (200,000), Lebanon (294,000), Russia (64,000), others in remote locations in Europe and America, especially in Germany (850,000), the United Kingdom (200,000), the Netherlands (100,000), Austria (80,000), Sweden (80,000), Belgium (70,000), France (23,000), and North America (58,000). The numbers are very difficult to assess and the essential sources, such as *Ethnologue* (2009), *Wikipedia*, and various scientific papers on the question differ quite widely.[3] We essentially focus our analysis on IIST countries with a large Kurdish population and assess their size by using various sources that we thought to be the most accurate. Each IIST country has its own national (or official) language(s), its own political views and rules, and one or even several Kurdish minority languages. Even though the total number of Kurds amounts to some 44 million people (according to the Kurdish Institute of Paris), we base our analysis and suggestions on the reduced number of 26.5 million people who speak Kumanji or Sorani varieties.

In some countries, Kurdish is taught in public schools, in some it is not. Sheyholislami (2017) provides a good overview of the treatment of Kurdish languages in IIST countries. Since 1991, in Iraq "Kurdish [is] the official languages of the Kurdistan Region, [but] the teaching of Arabic

[1] Kurds are not the only ethnic group dispersed across different countries. Many African countries were created by artificial constructions. In countries with large numbers of linguistic, ethnic, and religious varieties, their ethnic spreads often fail to coincide with the official borders.
[2] https://www.institutkurde.org/en/info/the-kurdish-population-123551004
[3] The estimated number of Kurds in Iran, for example, varies between 8 million according to *Wikipedia* and 'over 3 million' according to Eppler and Benedikt (2017).

is compulsory." In Iran "Kurdish has almost never enjoyed an official status, [and nowadays], it is restricted and under controlled tolerance." In Syria "since, 2014, the status of Kurdish has been elevated in many respects. For example, in the largest autonomous region of the Jazira canton, Kurdish, Arabic and Syriac-Aramaic are the official languages, while all other linguistic groups have the right to be taught in their native languages. In the three cantons of Syrian Kurdistan, the Kurmanji variety is now the language of education, media and public institutions." Finally, in Turkey, "Kurdish is largely restricted to private domains," [and] has no official status."

In addition, Kurdish is not a unique language, but a so-called "continuum of related northwest Iranian dialects" (Eppler and Benedikt, 2017, p. 110), sometimes written in different scripts: (a) the Latin script is used for Kurmanji in Syria and Turkey, (b) an adapted Persian-Arabic script where the vowels are written (which is not the case for Arabic itself), is in use for Sorani and Badini in Iran and Iraq; (c) Armenian and Cyrillic are still used in the former Armenian Soviet Socialist Republic. Moreover, the *Kurdish Academy of Language* 'constructed' Yekgirtu as a 'unified' script using the Latin alphabet. Recently, an augmented Latin alphabet was suggested that includes some missing Kurdish phonemes in the previous Latin-based scripts (Ahmadi, Hassani and McCrae, 2019).

The purpose of this paper is not to address the really deep and sensitive aspects of Kurdish linguistic rights. Our objective is, rather, to assess economic and sociologic arguments that could help finding a novel angle of examining the problem.

However, identifying an efficient and fair linguistic policy is not an easy task even within the framework of a single country. In some cases, such as Sri Lanka, it led to a protracted and bloody war whose human costs amounted to hundreds of thousands lives. The choice of linguistic policies has been a contentious issue in many African countries as well. For example, the Rwandese situation, led to large ethnic killings (500,000 to 800,000) of Tutsi (the higher class) by Hutus,[4] and introduced three language changes over the last twelve years which involved the switch from French to various combinations of English and the local language, Kinyarwanda. Painful and often dramatic switches of linguistic policies were implemented in Spain after the death of Franco and in the former Soviet republics after the break-up of the USSR. The list goes on and on and may even include the uncertainty regarding the challenges of the European Union which has to handle 24 official languages.[5]

[4] Though the Sri-Lankan problem was mainly ethnic, and started with an insurgency against the ruling government, in order to divide the island between Tamils who are from Indian origin and Sinhalese who are native to Sri Lanka.

[5] Moreover, the European Commission has to translate official documents into all 24 official EU's languages. It also has the task of interpreting languages in many meetings (if requested by attendants) and in the Parliament.

For the Kurdish people the situation is much more difficult, since the political situation in IIST countries is quite unstable, and none of them is prepared to address the problem of Kurdish minority languages.

Of course, trivially, those languages would become more important if the different enclaves in which they are spoken by a majority of individuals living there, and form a new, linguistically cohesive, political, and administrative unity. But since this is not a realistic scenario, at least in the short run, we proceed with the Kurdish linguistic challenges that we encounter in each IIST country.

Though there is the additional difficulty that in almost each Kurdish enclave, several dialects are spoken, should Kurds have a unified language in each IIST country? Should there be a unified language in all countries? Could an international language be taught in Kurdish schools, and liaise all Kurds? None of these options means that the existing 'dialects' or 'languages' should be banned, but that one of them (either Kurdish or international) should be chosen as *lingua franca* spoken or at least understood by all Kurds.

Or should one adopt the solution suggested by the *Kurdish Academy of Languages* which questions "whether one variety of Kurdish should be developed, into an international standard and if so, which one should it be."[6] Such a solution, called *standardization*, may reduce the number of (official) languages, but may also lead to *disenfranchisement,* depriving large segments of a society of their linguistic rights. In some cases, this sentiment turns out to be relatively soft, as it was in France after the 1992 change of the Constitution, which included an article imposing French as 'the language of the Republic.' By contrast, this may also lead to wars, as discussed earlier.

Still, too much fractionalization of languages across countries, and *a fortiori* in the same country, often has negative political as well as economic effects, simply because the various populations have difficulties to communicate and to engage in trade.[7] In many African countries, hundreds of languages are spoken. But one can argue that it is important to use one's native language as identifier of a cultural group. This importance is reinforced by Bretton (1976, p. 447) who suggests that, "language may be the most explosive issue universally and over time. This mainly because language alone, unlike all other concerns associated with nationalism and ethnocentrism, is so closely tied to the individual self." The importance of one's native languages is also underlined in a quotation attributed to Nelson Mandela: "If you

---

This costs over a billion Euros per year, though there exists a reasonable consensus between countries to cover this amount.

[6] Eppler and Benedikt (2017).

[7] See Easterly and Levine (1997) and Alesina and La Ferrara (2005).

talk to a man in *a* language he understands, that goes to his head. If you talk to him in *his* language, that goes to his heart."

To address this difficult dilemma, we propose a multilingual policy for areas populated by Kurds in IIST countries. More specifically, we develop a variant of the three-languages formulas applied with varying degrees of success in India, Nigeria, or Kazakhstan. Based on our estimation of linguistic distances between various Kurdish languages and the number of their speakers, our proposal focuses on three languages to be learnt by Kurdish populations: the official language in the country of residence (Arabic, Farsi, or Turkish) for the purpose of efficiency and necessity; a Kurdish language to preserve Kurdish identity (Kurmanji, Sorani or other) and a language for the worldwide communication and for inter-community exchange between different Kurdish groups, when communication in their own dialects is mutually unintelligible.

The paper is organized as follows. In section 2 we describe the main features of the Kurdish 'linguistic problem.' Section 3 details our methodology. In particular, we compute linguistic distances between dialectal varieties of the Kurdish languages, three belonging to the main Kurmanji variety, and three to the Sorani variety. We explain why we had to limit ourselves to these varieties (and exclude Southern Kurdish varieties resorting to Gorani for instance) and yet, the extent to which the conclusions we can draw, can bear a general lesson. In section 4 we discuss our result and potential conventional linguistic policies for the Kurds. In section 5 we offer a brief historical overview of several cases that pave the way for our proposal outlined in section 6. Section 7 concludes.

2. Languages, dialects and their populations

It is hard to know how many Kurds live in the Middle East, and even more so, which language(s) or dialect(s) they speak. This situation should not be surprising since none of the four countries has any economic or political interest to separate Kurds from the majority of inhabitants and from the language of the majority, even if they are allowed to speak their own language (see above).[8] In addition, there probably exist no surveys about languages, and if such surveys existed "the set of speech forms that a person commands, is not, to be sure, all there is to language. Language, after all, is not only a means of communication, but it is also a marker of identity" (Laitin, 2000, p. 144), which indeed may induce some people to say 'I am a Kurd who lives in Iran, but I speak Farsi and English.'

[8] Sheyholislami (2012, p. 25), for example, writes that "The Iranian census does not provide information about the ethnicity or language of the population."

We also had to compromise between two types of data: the number of speakers in each language used in the four main countries of interest, and the availability of some two hundred words in each language (transposed into the International phonetic alphabet) which leads us to measure linguistic distances between languages.

Dialects and languages

A short discussion of the word *dialect* that is often used to describe Kurdish 'ways of speaking' is in order. Hassani (2018, p. 629) suggests that "if two dialects are mutually unintelligible, they are considered two different languages, otherwise, two dialects of the same language. However, there are dialects that are mutually intelligible which are considered languages and those which are mutually unintelligible yet considered a dialect of a language." This is all the most confirmed by the idea that the Kurdish speaking area is one of a dialectological continuum (Matras & Akin, 2012).

According to the *Oxford Living Dictionary*,[9] a dialect "is a particular form of a language peculiar to a specific region or social group." The literature on Kurdish language(s) often uses the word dialect (or even *continuum* of dialects) instead of language. As *Ethnologue* (2009, p. 9) notes, "every language is characterized by variation within the speech community that uses it. Those varieties are more or less divergent and are often referred to as dialects, which may be distinct enough to be considered separate languages or sufficiently similar." Moreover, "not all scholars share the same set of criteria for distinguishing a language from a dialect." The criteria used by *Ethnologue* (2009, p. 9) to arrive at their count "make it clear that the identification of a 'language' is not solely within the realm of linguistics."

This is so in Europe as well. Are Italian and French two dialects of Latin, or are they two distinct languages coming from Latin? One could consider French as one among the many dialects or varieties spoken in France, together with Gallo spoken in Brittany and Normandy, Lorrain, Picard, Provençal, to quote just a few languages that belong to the Romance branch, but are considered to be dialects today. It was only in 1539, that King François I ordered the language variety 'French' to become the only official language of the French administration (Ordinance of Villers-Cotterêts). So, one could assume that Latin was indeed the official language in France until 1539, and conclude that one of the dialects spoken in France was erected to become a language.[10]

A couple of examples are useful. *Ethnologue* (2009) lists 57 Zapotec languages in Mexico.

[9] https://www.lexico.com/definition/dialect [last consulted August 30, 2020].
[10] SeeWright (2016), pp. 450-455).

Some of those, such as Zapotec (San Augustin Mixtepec), count less than 100 first-language speakers. The largest, Zapotec (Isthmus), has 85,000 speakers. In contrast, *Ethnologue* (2009) also lists five *dialects* in the Flemish part of Belgium that are certainly all spoken by more than 100 people, and though they are considered variants of Dutch, can probably not be all that well understood by a Netherlander living in Amsterdam. Is the language spoken in Quebec close to French? This can be questioned since on French TV, it happens that French-Canadian series are subtitled in French. Canadian French spoken in Quebec is indeed closer to the French spoken in France during the 17th century—at the time French migrants landed in Canada — than today's French spoken in France. Words and accents can be very different.

This even happens under our eyes. In an interesting paper published by *Newsweek* on March 7, 2005, 'Not the Queen's English: Non-native speakers are transforming the global language,' the author distinguishes more than fifty varieties, including British English (BBC English, English English, Scottish English, Scots, Norn, Welsh English, Ulster Scots, Hiberno-English, Irish English), American English (Network Standard, Northern, Midland, Southern, Black English Vernacular, Gullah, Appalachian, Indian English) and Canadian English (Quebec English, Frenglish, Newfoundland English, Athabascan English, Inuit English), not to speak about Spanglish used by a growing community of recent immigrants in the United States.[11]

The line that we follow is not based on whether we should give more importance to a 'language' than to a 'dialect.' Our calculations are rather based on phonetic distances between the vocabularies of six couples of languages and dialects spoken in Iran, Iraq, Syria and Turkey. This will lead us to a quantitative answer, though vocabularies (that are similar or dissimilar) only may not be sufficient: Grammars contribute to heterogeneity as well. German has declensions, but only very few of them remain in English such as the so-called possessive genitive in the *moon's* last quarter. None of the declensions that existed in Latin were inherited by today's French, Italian or Spanish. Germanic languages usually have three genders, masculine, feminine and neutral. In English, there exists of course *she* and *he*, but most common names (place, thing, idea, action or quality) have no gender. Such cases happened in Kurdish languages as well.

The choice of Kurdish languages and populations

We chose languages that are spoken in four countries only, Iran, Iraq, Syria and Turkey, excluding Armenia, Azerbaijan, Georgia, Lebanon, where there exist large communities of speakers of Kurdish, and excluding countries that are geographically not part of what is usually called Kurdistan. We also dropped Gorani, called Zaza by some experts (spoken in Iran, Iraq

[11] See Ginsburgh and Weber (2016, p. 140).

and Turkey), which, according to many linguists are not Kurdish languages, though most Zaza speakers (some 2 to 3 million) define themselves as Kurds. Note that UNESCO includes Gorani and Zaza in their list of endangered languages.[12]

We decided to include three so-called Northern dialects of Kurmanji spoken in Northern Kurdistan: Badini (Iraq), Kurmanji-Rojavaei (Syria) Kurmanji-Bakur (Turkey), and three Central dialects linked to Sorani in Central Kurdistan: Hawleri (Iraq), Mahabadi (Iran) and Suleimani (Iraq). The choice of the six dialects are conditioned by two reasons. First, they are spoken by such a large community in IIST that allows one consider them as representatives of the dialects under study and the second is that we were unable to cover the entire Kurdish population.

[Insert Table1 approximately here]

Still, we have information about 26.5 million speakers out of 35.5, that is almost 75 percent of speakers, though this percentage varies across countries: In Iran, we cover only 15 percent of the population, while in the three other countries, coverage is much larger: 69 percent in Iraq, roughly100 percent in Syria and Turkey.

We need to recognize that other minority languages or dialects (Kurdish and others) are spoken in the areas that we are talking about, but could hardly take account of them. We certainly do not want any of them to die, or set aside the sociopolitical ramifications of ignoring them.

3. Linguistic distances

As discussed earlier, in our study we consider three variants of Kurmanji and Sorani dialects. Should we consider speakers of Kurmanji (or Sorani) as members of different groups of languages, or can they understand each other? And what about the distance between Kurmanji and Sorani? To address these questions, we have to recognize some degree of distinctiveness between languages. Is it vocabulary that makes them different, or pronunciation, grammar, phonetics, phonology, syntax, that is the "way in which linguistic elements (such as words) are put together to form constituents as phrases or clauses",[13] and this is even without going into the fundamental issue of whether languages have a common structure.

[12] See also Scalbert-Yücel (2006, p. 123).
[13] See *Merriam-Webster Dictionary*.

Here, we only use so-called *lexical linguistic distances*, supposedly based (or not) on common roots of words in the vocabularies of couples of languages. But we rely on phonetic distances between pairs languages, which are of course closer to spoken than to written languages. In our case, this also avoids the problems of comparing words in two alphabets (Latin and Arabic).

Since it would be a daunting task to compare the tens of thousands of words for each couple of languages, linguists suggest to rely on a small selection of carefully chosen words, a so-called 'list of meanings.' Morris Swadesh (1952) introduced some rigor into the choice of meanings (or words) that one can assume to be basic enough to exist in all languages and cultures, such as *I, you, he, we, animal, flower, blood, to bite*, etc., on which comparisons or distances between languages can be based.[14] The list we are interested in consists of 207 basic meanings. It is still in use today, though in his later research, Swadesh trimmed the number of words to one hundred.

Dyen, Kruskal and Black (1992), who used Swadesh's one hundred list of words for Indo-European languages, describe the *lexicostatistical* method as consisting of four stages:

Stage 1. Collecting for each of these standard words or meanings, those used in each speech variety under consideration.

Stage 2. Making cognate[15] decisions on each meaning in the lists for each pair of speech varieties, that is, deciding whether they have a common ancestral word or not, or whether no clear-cut decision can be made. This phase is performed by linguists who know the language family, in their case, Indo-European languages.

Stage 3. Calculating the lexicostatistical percentages, that is, the percentages of cognates shared by each pair of languages. These distances lie between 0, if all words are cognate and 1 if there is no cognate word in both languages.

A couple of comments on the use of the three stages in our framework is in order.

[14] See Kessler (2001, pp. 192-257) for the list of meanings chosen by Swadesh.
[15] In linguistics, cognates are words that have a common etymological origin. Cognates are often inherited from a shared parent language, but they may also involve borrowings from some other language. Here what matters is that a word in one language can be understood in the other language, and vice-versa.

Stage 1'. The 207 meanings suggested by Swadesh are collected for each of the six languages (or dialects), and transposed into phonetic script. In Table 2 we display, as an example, ten words taken from Swadesh's list of words. Their meaning in English is shown in column 2, while the following columns contain the words in each of the six Kurdish languages, in phonetic script. At first view, the languages are reasonably close.

[Insert Table 2 approximately here]

Stage 2'. While in the times of Dyen, Kruskal and Black (1992), the cognate decisions of Stage 2, were carefully examined by experts, in our case, the task is much easier and can be left to a computer, not only because computers are sometimes better than experts, but essentially because computers are faster than humans.[16]

Levenshtein (1966) seems to be the first scientist who suggested to use computers to calculate such distances. He understood that distances in words are insufficient if these words are pronounced in different ways in the languages that are being analyzed. Inter-comprehension is especially difficult due to vowels and diphthongs. This should make it quite obvious that the distances of words written in normal characters are insufficient, while Levenshtein's method can take into account phonetic distances, transcribing the words into their phonetic equivalents, using existing or especially tailored phonetic alphabets.

Levenshtein's idea is to convert the word of one language into the word of the other one, by inserting, deleting or substituting alphabetic or phonetic (see Table 2) characters. Roughly, the number of such transformations between the two words is the Levenshtein distance ($Lev(i,j)$ in what follows) also called *edit distance*. It uses the following equation for this calculation:

$$Lev(i,j) = min \begin{cases} Lev(i-1,j)+1 \\ Lev(i,j-1)+1 \\ Lev(i-1,j-1)+2; \begin{cases} if\, w_1(i) \neq w_2(j) \\ if\, w_1(i) = w_2(j) \end{cases} \\ 0; \end{cases} \quad (1)$$

[16] Note that Dyen, Kruskal and Black did their work in 1992. They could obviously use computers. The problem is however that their work was not so much about intercomprehension of languages, but going to the old roots to build linguistic trees. See McMahon and McMahon (2005) who compare at great length the various methods, including the heroic one used by Dyen, Kruskal and Black, and Levenshtein (1966) distances.

In Equation (1), W_1 and W_2 are two strings of words used to calculate the distance while i and j are corresponding indices of the elements (here, letters) of W_1 and W_2, while $w_1(i)$ and $w_2(j)$ are elements of the two strings.

In this method, 0 means the two words that are compared are identical while any number greater than 0 means they are different. The larger the number the wider the distance. For example, a distance of 1 means that with one insertion or deletion the two words become identical while a distance of 4 may require two substitutions or one substitution and two other operations (deletion/insertion) or four non-substitution operations (insertion/deletion).

To explain how this works, we discuss as example the word *to burn* which belongs to Swadesh's list. The two words have 6 characters in Kurmanji (Iraq) and in Kurmanji (Turkey), and 5 in Sorani (Iran). *To burn* is spelled *şivtîn* in Kurmanji (Iraq), *şewtîn in* Kurmanji (Turkey), and *sûtan* in Sorani (Iraq). A linguist would probably classify the words as cognate between all three dialects. The number of edits between Kurmanji (Iraq) and Kurmanji (Turkey) words is 4 since one needs two substitutions (*i* for *e* and *v* for *w*). For Kurmanji (Turkey) and Sorani (Iran), this number is 7 (substitute *s* for ş; insert *i;* substitute *û* for *v*; substitute *a* for *î*).

Applying (1) for Kurmanji (Iraq) and Kurmanji (Turkey) words is 2/6, since there are only two *edits*, and each word has 6 characters. The distance between Kurmanji (Iraq) and Sorani (Iran) words is 4/6; again, 4 is the number of *edits* and 6 is the largest number of characters between the two languages (6 in Kurmanji and 5 in Sorani). We can assert that the distances in this example are is in accordance with our intuition: Kurmanji (Iraq) and Kurmanji (Turkey) are both Kurmanji languages and are closer to each other than Kurmanji (Iraq) and Sorani (Iran) that belong to two different dialects, though all three are Kurdish languages. Comparing one word only between two languages or dialects is obviously not sufficient and the computation is made for all words (here 207) between all couples of languages.

Several methods based on *string metrics* are available. Some are modifications of the Levenshtein method. For example, one of these variants assumes that the *cost* is equal to 1 for all types of operations. Others, sometimes called *string similarity*, are not based on Levenshtein. They consider other parameters in addition to the operations, such as the string's length. Jaro (1989) is one of the earliest and most well-known string similarity method. A

modification to Jaro, Jaro-Winkler (Winkler, 1990), considers the prefix matching among the strings as well. These methods might provide different and more accurate results, depending on the context and the purpose they are used for, but their outcomes on distances do not drastically differ from each other. Both Jaro and Jaro-Winkler map the similarity score to a number between 0 and 1. A practical outcome of this normalization is to obtain the distance between two strings by subtracting 1 from their similarity measure. Jaro and Jaro-Winkler methods are shown to outperform or perform better than many existing methods in a variety of tasks (see, for example, Cohen, Ravikumar and Fienberg, 2003; Pradhan, Gyanchandani and Wadhvani, 2015). The normalized outcome of Jaro is more meaningful in our case that focuses on similarity (or dissimilarity) of pairs of languages (dialects) than the magnitude that the Levenshtein edit distance offers. Furthermore, Jaro-based methods are more precise when applied to short strings (Piskorski and Sydow, 2007).

Jaro uses the following equation to measure similarity:

$$Jaro(a,b) = min\begin{cases} 0 \; if \; m = 0 \\ \frac{1}{3} \times \left(\frac{m}{|a|} + \frac{m}{|b|} + \frac{m-t}{m}\right) \; otherwise \end{cases} \quad (2)$$

In Equation (2), a and b are two strings for which we calculate the Jaro similarity, m is the number of matching characters (here, letters), and t is half of the transpositions required to make the two strings identical. This equation calculates similarity, but for our purpose we prefer to deal with dissimilarity or distance between a pair of languages/dialects. The Jaro distance then is:

$$Jaro_{distance}(a,b) = 1 - Jaro(a,b) \quad (3)$$

Step 3'. The distance between two languages is now simply the average distance over all the meanings that are compared. Computing d string similarity/distance is reasonably easy to program on a computer and allows doing this for a large number of languages meanings. Table 3 displays the distances between six important Kurdish languages or dialects though some are spoken by a large number of people, others by a few people. Bakur, in Turkey is spoken by roughly half of the Kurdish population (18 out of 35.5 million, while Badini in Iraq and Mahabadi in Iran are both spoken by less than a million people.

[Insert Table 3 approximately here]

Average distances within Sorani ((0.168+0.109+0.182)/3=0.153) are smaller than those within Kurmanji (0.236), while average distances between Kurmanji and Sorani dialects are the largest (0.332). Sorani dialects are thus closer to each other than Kumanji dialects. But the distance between Kurmanji and Sorani languages is larger. On average, inter-comprehension is thus easier within Kurmanji or within Sorani dialects, than between any Kumanji and any Sorani dialect.

Since average distances could be considered as a rough measure, we also calculated the percentage of fully matched words ($Jaro_{\text{distance}} = 0$) and completely different ones ($Jaro_{\text{distance}} = 1$) among the dialects. Table 4 summarizes the results. The columns that show these percentages are headed *NM%* for completely different words and *FM%* for fully matched ones.

[Insert Table 4 approximately here]

4. In which direction could we go to increase inter-comprehension?

The perfect (but unfeasible solution) would be to get back to the pre-1925 situation, the year in which the Kurdish population found itself divided between four countries and though promises had been made, namely to set up an autonomous Kurdish government, they were not hold. In this ideal situation, schools and universities should teach the local dialect as well as one common language, probably the one that is spoken by the largest number of people, or a foreign language, which means that two languages would have been sufficient.

In what follows, we consider various feasible possibilities.

(a) Unify both Kurmanji *and* Sorani dialects? This idea has been suggested during the previous century (Haig and Matras 2002; Hassanpour, 2012; Saeid, 2014). It has been discussed from different perspectives, some have emphasized the linguistic and sociocultural barriers (Hassanpour, 2012; Khalid, 2015), and some others have underlined political factors (Saeed and Jukil, 2018; Ibrahim, Jukil, Beckett, 2020). In our approach, the distances discussed in Section 3 and Table 3 are all larger than 0.30. Therefore, unifying would probably be more

difficult to achieve than in (b) and (c), but the latter is not the best either. That corroborates with the scholars that have studied this issue from the linguistic point of view.[17]

(b) Unify Kurmanji dialects *only*? The largest one (Bakur) is spoken by 18 million people who live in Turkey. One could suggest that the 3.4 million who do not speak Bakur, but Badini (in Irak) or Rojavaei (in Syria) be trained to speak Bakur. Bakur and Rojavaei are not very distant (0.280), and they are probably able to understand each other quite easily; they count for 2.5 million people who live in Syria. The problem is a little more difficult between the Badini spoken in Iraq, and Turkey's Bakur, but the population is rather small (0.9 million).

(c) Unify Sorani dialects *only*? Probably not, since no individual language is far from any other: the average distance between the dialects is equal to 0.153: inter-comprehension is probably relatively easy.

Those are more 'conventional' proposals that, as we argued above, are not very satisfactory. In the next section we utilize the experience of other countries which may be useful for laying out our multi-lingual proposal. In what follows, we suggest a so-called three-language formula, meant to be spoken by all Kurds.

5. A three-language formula for Kurds

This-three language 'Kurdish formula' should include: the language of the country in which they live (Farsi, Arabic or Turkish), their local Kurdish language or dialect, *and* L, the third should be a fully common language. Here is to what it would lead:

Iran: Farsi, Mahabadi *or* Suleimani *and* L,
Iraq: Arabic, Badini, Hawleri *or* Suleimani *and* L,
Syria: Arabic, Rojavaei *and* L,
Turkey: Turkish, Bakur *and* L.

We think making a good choice for L is more efficient than the simple alphabet converter Yekgirtu, proposed by the Kurdish Academy. Kurds who live in one of those countries would

[17] Hassanpour (1992) suggested to follow a bi-standard approach, that is to have one standard for Kurmanji and one for Sorani.

be able to communicate much better not only with their neighbors, but also with Kurds living in other countries than IIST, who do not speak the same language. If, in addition, L were a reasonably international language, Kurds could communicate with a larger part of the rest of the world.

Choosing linguistic policies has been faced by multilingual societies for very long. Many countries and empires had to make difficult choices to select a unique language for official use an purposes, taught in schools, and used for prayers in churches, mosques, and synagogues. This type of linguistic rationalization or standardization has often been achieved by imposing a unique language. The most frequent case was choosing the language of the majority group (French in France, Han Chinese in China, Kuotsugo Japanese in Japan). In the absence of a majority group, some societies have turned to a *lingua franca*, spoken by the majority of the population but that fails to be the mother tongue of sometimes large population groups (Swahili in Tanzania, and, to some extent, everywhere in East Africa, Bahasa in Indonesia, or even English in the United States). In some cases, historical and political considerations led to linguistic policies based on recognizing the language of a ruling minority group. This was the case of Amharic in Ethiopia and Afrikaans in South Africa.

However, the choice of a unique official language has been reconsidered over time, and has accelerated during the second part of the last century. Countries and communities have invested considerable resources in protecting their regional and native languages in North and South America, Africa and Europe. In 1992, for example, the Council of Europe adopted the European Charter for Regional or Minority languages, a treaty designed to protect and promote historical, regional and minority languages in Europe. The growing recognition of the will of the people, as well as the fact that oppressing or even suppressing community languages leads to disenfranchising various groups and possibly generate conflicts or wars, as was mentioned earlier. One should point out that the effects of globalization "proceeds in English" (De Swaan, 2013, p. 186), but, as Dorfman (2002, p. 92) puts it, "the ascendency of English, like so many phenomena associated with globalization, leaves too many invisible losers, too many people silenced." In short, the choice of linguistic policies should take into account the multilingual nature of the population, and in fact, many countries have several official languages: South Africa has eleven, while Belgium and Luxembourg, that are both very small, have three. In these countries, the choice of official languages, was rooted in regional considerations and led

to accepting several regional languages as official languages. Let us briefly examine what happened in several countries (India, Nigeria, Kazakhstan and Rwanda) which tried to implement a third language, supposed to be spoken by the whole population of the country, but that also has an international flavor.

India. In defining its own linguistic policy, India took a step further by extending the principle of regional inclusion. Its three-language policy (the local language, Hindi and English) was suggested in 1965-66 and adopted by the Parliament in 1968. The policy was initiated as a national response to bitter complaints from Tamil Nadu and other Southern states which claimed that the use of Hindi in government services imposed formidable barriers since they were required to become proficient in two non-native languages, English and Hindi, whereas speakers of Hindi had to learn English only. Therefore, the new formula, that varied across states, required children in Hindi-speaking states to study Hindi, their own language(s), but also English and one of the Southern languages, whereas children in non-Hindi speaking states were supposed to learn their own regional language, Hindi and English. This apparently well-crafted formula was aimed at developing:

(a) Group identity, preservation of mother tongues and traditions through the study of regional languages (in areas with non-Hindi native languages),
(b) National pride and unity by acquiring Hindi (as native or non-native language), and
(c) Administrative efficiency through standardization (by learning and speaking English).

The implementation of the formula turned out to be of mixed success. In addition to the reluctance and lukewarm support of the regional administrations, the formula failed to generate wide public support both in the North and the South. In Hindi regions, relatively little effort or resources were spent on studying English and even less so on learning other languages. There was more interest in studying English and Tamil in Tamil Nadu, but almost no interest in acquiring Hindi at the same time. In short, the lack of public commitment and of required resources killed the smooth and wide-ranged implementation of the formula.

Nigeria. Some elements of the Indian formula design have been experimented in Nigeria, a country with over 500 languages and a population of some 200 million inhabitants. Three major regional languages Hausa (72 million speakers), Igbo (27 million speakers), and Yoruba

(between 45 and 55 million speakers), were suggested as the cornerstone of the nation-unifying device (Laitin and Watkins IV, 1986). Again, the idea did not go very far and was slowed down due to the lack of public funds, and the absence of commitment of students, their families and regional authorities.

Kazakhstan. The former Soviet republic that declared its independence from the Soviet Union in 1991 is an interesting example of the recent implementation of the three-language formula. There are two large ethnic groups in the country, Kazakhs and Russians. 74 percent of the 12 million-wide population speak, or at least understand, Kazakh, while 94 percent speak, or at least understand, Russian (Smailov, 2011). Under the Constitution adopted in 1993, Kazakh became the state language, whereas Russian was declared an official language, used routinely in business, government, and inter-ethnic communication. Despite the official support of the Kazakh language after the independence, scholastic achievements were substantially lower for pupils taught in Kazakh, a probable consequence of the comparatively poor quality of schools that teach in Kazakh.

In 2011, Kazakhstan introduced a three-language formula that, in addition to Kazakh and Russian, included English as a world language. Unlike India and Nigeria, Kazakhstan spent considerable resources in improving infrastructure, and incentives to learn English, developing the culture of language use, raising demand for all three languages in government programs, and preserving linguistic diversity. An important factor that contributed to the success of the formula was the public willingness to buy into a program that strengthened the role of the national Kazakh language, preserved the current and future role of Russian as a vehicle of regional communication as well as support for social stability in the country. It also highlighted the role of English as an indispensable factor of international communication and business cooperation.

6. A proposal for a three-language formula

As already said the three-language Kurdish formula should include: Farsi, Arabic or Turkish, depending in which country they live, the local Kurdish language, *and* English, which has become the language of globalization around the world. Here is to what this would lead:

Iran: Farsi, Mahabadi *or* Suleimani *and* English,
Iraq: Arabic, Badini, Hawleri *or* Suleimani *and* English,
Syria: Arabic, Rojavaei *and* English,
Turkey: Turkish, Bakur *and* English.

We think that this would be more efficient than the simple alphabet converter Yekgirtu, proposed by the Kurdish Academy. The 44 million Kurds who live in IIST countries would not be able to communicate much better not only with their neighbors, but also with Kurds living in other countries (especially former USSR countries, America, Europe, other West Asian countries) who do not speak the same language (some 2 million people).

While our proposal contains several elements of the Indian and Kazakh formulas, they differ in one very important dimension. Namely, the formulas in India, Nigeria and Kazakhstan tried to deal with the challenge of developing state identity, while in Kurdistan this would be replaced by ethnic identity, which means that the cross-learning of different Kurdish languages and dialects is a financial, rather than an ethnic issue. One is faced with the costs of learning only, rather than the pain and damage of linguistic disenfranchisement. The issue is the distances between various Kurdish languages, which are quite large for dialects of different languages (Kurmanji and Sorani) and quite small for dialects of the same language (dialects of Kumanji and dialects of Sorani). We thus suggest trying to implement the following three-language formula based on three principles:

(a) Efficiency and necessity of official language in the country of residence: Arabic in Iraq and Syria, Farsi in Iran, and Turkish in Turkey,
(b) Preservation of Kurdish identity: local Kurdish local language,
(c) Inter-community between different Kurdish groups, and access to other countries in the world: English.

The first two languages are taught in most schools and universities, and English is also taught is some countries (Iraq for example). The cost of implementing this formula should therefore be sustainable especially if non-IIST countries would be ready to contribute. It may also well be that the gains of trade obtained by this common language would cover part of the cost. We realize that our 3-language formula imposes unequal burdens on various individuals, as some speak all three recommended languages, and some do not. It is similar to the Indian policy

which Laitin (1989) called 3+/-1 formula, indicating that some people have to study less, while others do more. However, the goal of linguistic standardization and efficient communication should outweigh the described challenges.

7. Conclusion

By investigating inter-dialectal lexico-statistical distances of two main varieties of Kurdish – Kurmanji and Sorani – we could not identify an optimal way to adopt one of those dialectal varieties as the common linguistic vehicle adopted by all speakers of a Kurdish dialect in order both to maximize inter-comprehension and to minimize learning costs. Our main conclusion therefore would plea in favor of the generalized use of another language. Instead of trying to reduce the numbers of dialects or to advise to suppress any of the two languages that each Kurd speaks in any country (i.e. a variety of Kurdish and an official national language: Arabic, Turkish, Farsi), we suggest adding a third language at school and universities. This would not represent any form of linguistic alienation to the extent that speakers continue, officially and in practice, to speak the two other languages. As a matter of fact, English is the second spoken language in Iran, for instance, where the younger generation has relatively high English language abilities. Most tourists who visit Iran are surprised by the number of people who understand and speak English. Until the early 1950s the second official language of Iran was French. Many French words remain in the Persian everyday language. In Iran the months of the calendar are called with their French pronunciation. But for the last 50 years English is the second language of the country. Moreover, English is understood and (sometime only roughly) spoken by 1.5 billion people all over the world, while French lags behind with 300 million, mainly in France and North and West Africa.

In Iraq too, English is taught in many Kurdish schools, while in Syria, many educated Syrian also speak English or French, but English is the most widely understood.[18]

The issue is not especially with English. French, which was a colonial language in the region, and is still spoken in Syria, could have performed the same role we assign to English. But it could seem a paradoxical result that avoidance of disenfranchisement and loss of dialectal varieties – which is a desirable aim which most Kurds advocate – needs resorting to a foreign

[18] See https://www.cs.mcgill.ca/~rwest/wikispeedia/wpcd/wp/d/Demographics_of_Syria.htm

language. But we do not see any contradiction, in this particular context, to follow two objectives: Integration, that Kurds aim at, in a global economy and global politics, by means of an international language, and maintenance of dialectological variety and linguistic identity.

Regardless of the chosen linguistic policy, Kurmanji and Sorani continue their mutual influence to enrich each other. That is the nature of languages. We do not expect this process to develop a "standard" or unified Kurdish. Perhaps focusing on each dialect to overcome their issues would be a better decision regarding the achievement of a "standard version" for each dialect.

References

Ahmadi, Sina, Hossein Hassani, and John McCra (2019), Towards electronic lexicography for the Kurdish language, In *eLex Proceedings of the Sixth Biennial Conference on Electronic Lexicography*, Sintra, Portugal, pp. 881-906.

Alesina, Alberto and Eliana La Ferrara (2005), Ethnic diversity and economic performance, *Journal of Economic Literature* 43, 762-800.

Bretton, Henry (1976), Political science, language, and politics, In William O'Barr and Jean O'Barr, Eds., *Language and Politics*, The Hague: Mouton.

Cohen, William, Pradeep Ravikumar and Stephen Fienberg (2003), A comparison of string distance metrics for name-matching tasks, In *American Association for Artificial Intelligence* (www.aaai.org).

De Swaan, Abram (2013), *Words of the World: The Global Language System*, John Wiley & Sons.

Dorfman, Ariel (2002), The nomads of language, *The American Scholar* 71, 89-94.

Dyen, Isidore, Joseph B. Kruskal and Paul Black (1992), An Indo-European classification: A lexicostatistical experiment, *Transactions of the American Philosophical Society* 82, Philadelphia: American Philosophical Society.

Easterly, William and Ross Levine (1997), Africa's growth tragedy: Policies and ethnic divisions, *Quarterly Journal of Economics* 112, 1203-1250.

Eppler, Eva and Joseph Benedikt (2017), A perceptual dialectological approach to linguistic variation and spatial analysis of Kurdish varieties, *Journal of Linguistic Geography* 5, 109-130.

Ethnologue (2009), *Languages of the World*, M. Paul Lewis, Ed., Dallas, TX: SIL International.

Ginsburgh, Victor and Shlomo Weber (2016), Linguistic distances and ethno-linguistic

fractionalisation and disenfranchisement, In V. Ginsburgh and S. Weber (eds.), *The Palgrave Handbook of Economics and Language*, Basingstoke, UK: Palgrave-Mac Millan.

Haig, Geoffrey and Yaron Matras (2002), Kurdish linguistics: a brief overview. *STUF-Language Typology and Universals* 55,3-14.

Hassani, Hossein (2018), BLARK, for multi-dialect languages: towards the Kurdish BLARK, *Language Resources & Evaluation* 52, 625-644.

Hassanpour, Amir (1992*), Nationalism and Language in Kurdistan*, San Francisco: Mellen Research Press.

Hassanpour, Amir (2012), The indivisibility of the nation and its linguistic divisions, *International Journal of the Sociology of Language* 217, 49-73.

Ibrahim, Sangar, Ali Mahmoud Jukil and Gulbahar Beckett (2020). The impact of the Kurdish language on the components of nation building and the impact of the components of the nation building on the Kurdish language *Zanco Journal of Humanity Sciences* 24, 258-272.

Jaro, Matthew (1989), Advances in record linkage methodology as applied to the 1985 census of Tampa Florida, *Journal of the American Statistical Association* 84, 414-420.

Kessler, Brett (2001), *The Significance of Words Lists*, Stanford, CA: Center for the Study of Languages and Information.

Khalid, Hewa , 2015. Kurdish dialect continuum, as a standardization solution, *International Journal of Kurdish Studies* 1, 27-39.

Laitin, David (1989), Language policy and political strategy in India, *Policy Studies* 22, 415-436.

Laitin, David (2000), What is a language community, *American Journal of Political Science* 44, 142-155.

Laitin, David and James Watkins IV (1986), *Hegemony and Culture: Politics and Change Among the Yoruba*, Chicago, IL: University of Chicago Press.

Levenshtein, Vladimir (1966), Binary codes capable of correcting deletions, insertions, and reversals, *Cybernetics and Control Theory* 10, 707-710.

Matras, Yaron and Salih Akin (2012), A survey of the Kurdish dialect continuum, In *Proceedings of the 2nd International Conference on Kurdish Studies*.

McMahon, April and Robert McMahon (2005), *Language Classification by Numbers*, Oxford: Oxford University Press.

Piskorski, Jakub and Marcin Sydow (2007), String distance metrics for reference matching and search query correction, In *Proceeding of the International Conference on Business Information Systems*, Berlin and Heidelberg: Springer.

Pradhan, Nitesh, Manasi Gyanchandani and Rajesh Wadhvani (2015), A review on text similarity technique used in IR and its application, *International Journal of Computer Applications* 120, 29-34.

Saeed, M.Q. and Ali Mahmoud Jukil, (2018) The language policy in Iraqi Kurdistan region from the perspective of Spolsky's theories, 9th International Visible Conference on Educational Studies & Applied Linguistics 2018, doi: 10.23918/vesal2018.a14 I could not find the full first name for Saeed

Saeid, Moslih Aowni (2014), Two varieties of Kurdish in competition. Dissertation, University of Vienna. Faculty of Philological and Cultural Studies.

Scalbert-Yücel, Clémence (2006), Les langues des Kurdes de Turquie : la nécessité de repenser l'expression 'langue kurde', *Langage et Société* 117, 116-140.

Sheyholislami, Jaffer (2012), Kurdish in Iran: A case of restricted and controlled tolerance, *International Journal of the Sociology of Language* 217, 19-47.

Sheyholislami, Jaffer (2017), Language status and party politics in Kurdistan-Iraq: The case of Badini and Hawrami varieties*, Today and Tomorrow*, Graz, Austria: Grazer Linguistische Monographien.

Smailov, Arman (2011), Results of the 2009 national population census of the Republic of Kazakhstan: Analytical report.

Swadesh, Morris (1952), Lexico-statistic dating of prehistoric ethnic contacts, *Proceedings of the American Philosophical Society* 96, 121-137.

Winkler, William (1990), String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage, In *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 354–359.

Wright, Sue (2016), Language choices: Political and economic factors in three European States, In Victor Ginsburgh and Shlomo Weber, eds, *The Palgrave Handbook of Economics and Language*, Palgrave Macmillan.

Table 1. Kurdish Main Dialects, and Populations

	Iran	Iraq	Syria	Turkey	Total
Kurmanji	0	0.9	2.5	18	21.4
Badini (Iraq)	0	0.9	0	0	0.9
Rojavaei (Syria)	0	0	2.5	0	2.5
Bakuri (Turkey)	0	0	0	18	18
Sorani	1.2	3.9	0	0	5.1
Mahabadi (Iran)	0.8	0	0	0	0.8
Hawleri (Iraq)	0	2	0	0	2
Suleimani (Iraq)	0.4	1.9	0	0	2.3
Kurmanji+Sorani (Total)	1.2	4.8	2.5	18	26.5
Kurdish population*	8	7	2.5	18	35.5
Kurdish population (KIP)**	12	8.5	3.6	20	44.1

* Population that speaks this dialect and covered by our Swadesh lists of words.
** Kurdish Institute of Paris https://www.institutkurde.org/en/info/the-kurdish-population-1232551004

Table 2. An Example of Ten Words of Swadesh's List for Six Kurdish Languages

Swadesh list (no.)	English	Kurmanji			Sorani		
		Badini Iraq	Bakuri Syria	Rojavaei Turkey	Mahabadi Iran	Hawleri Iraq	Suleimani Iraq
1	I	ez	ez	ez	min	min	min
2	you (singular)	tu	te	te	to	to	to
3	he	ev	ew	ew	ew	ew	ew
4	we	em	em	em	eme	ême	ême
5	you (plural)	hîn	hûn	wun	engo	êwe	êwe
44	animal	ẖeywan	ẖeywan	ẖeywan	ẖeywan	ẖeywan	ajel
59	flower	gol	ç̭i̭çek	gul	goł	goł	goł
64	blood	xîn	xûn	xûn	xwên	xwên	xwên
94	to bite	leqdan	gez kirin	dev kirin	gezîn	gezîn	gezîn
196	correct	dirist	rast	rast	durust	rast	rast

Table 3. Jaro Distances Between Dialects

	Kurmanji			Sorani		
	Badini (Iraq)	Rojavaei (Syria)	Bakuri (Turkey)	Mahabadi (Iran)	Hawleri (Iraq)	Soleimani (Iraq)
Kurmanji						
Badini (Iraq)	0.000	0.280	0.282	0.350	0.302	0.311
Rojavaei (Syria)	0.280	0.000	0.147	0.355	0.332	0.324
Bakuri (Turkey)	0.282	0.147	0.000	0.356	0.334	0.323
Sorani						
Mahabadi (Iran)	0.350	0.355	0.356	0.000	0.168	0.182
Hawleri (Iraq)	0.302	0.332	0.334	0.168	0.000	0.109
Suli (Iraq)	0.311	0.324	0.323	0.182	0.109	0.000

Table 4. Jaro Distances Between Dialects

	Kurmanji						**Sorani**					
	Badini (Iraq)		Rojavaei (Syria)		Bakuri (Turkey)		Mahabadi (Iran)		Hawleri (Iraq)		Soleimani (Iraq)	
	NM %	FM %	NM %	FM %	NM %	FM %	NM %	FM %	NM %	FM %	NM %	FM %
Kurmanji												
Badini (Iraq)	0.0	100.0	10.6	36.2	9.7	32.9	14.0	19.8	9.7	23.2	8.7	22.2
Rojavaei (Syria)	10.6	36.2	0.0	100.0	5.3	61.8	14.5	23.2	14.5	26.6	12.1	25.1
Bakuri (Turkey)	9.7	32.9	5.3	61.8	0.0	100.0	13.5	23.2	10.1	25.1	10.1	24.6
Sorani												
Mahabadi (Iran)	14.0	19.8	14.5	23.2	13.5	23.2	0.0	100.0	7.3	63.8	7.3	58.5
Hawleri (Iraq)	9.7	23.2	14.5	26.6	10.1	25.1	7.3	63.8	0.0	100.0	4.4	69.1
Suli (Iraq)	8.7	22.2	12.1	25.1	10.1	24.6	0.2	7.3	58.5	69.1	100.0	0.0

The % are headed NM% for the completely different words and FM% for the fully matched ones.