PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

Escuela de Posgrado

Towards automatic detection of lexical borrowings in wordlists - with application to Latin American languages

Tesis para optar el grado académico de Doctor en Ingeniería que presenta:

John Edward Miller

Asesor:

PhD. César Armando Beltrán Castañón

Co-asesores:

PhD. Roberto Daniel Zariquiey Biondi

PhD. Johann-Mattis List

Lima, 2024

Informe de Similitud

Yo, **César Armando Beltrán Castañón**, docente de la Escuela de Posgrado de la Pontificia Universidad Católica del Perú, asesor de la tesis de investigación titulado "*Towards automatic detection of lexical borrowings in wordlists - with application to Latin American languages*", del autor John Edward Miller, dejo constancia de lo siguiente:

- El mencionado documento tiene un índice de puntuación de similitud de 12%. Así lo consigna el reporte de similitud emitido por el software Turnitin el 20/06/2023.
- He revisado con detalle dicho reporte y la Tesis o Trabajo de Suficiencia Profesional, y no se advierte indicios de plagio.
- Las citas y otros autores y sus respectivas referencias cumplen con las pautas académicas.

Lima, 19 de Setiembre del 2023

Beltrán Castañón, César Armando
Asesor
DNI: 20561260
ORCID: 0000-0002-0173-4140

"When a foreign word falls by accident into the fountain of a language, it will get driven around in there until it takes on that language's colour and resembles a native term in spite of its foreign nature."

(Jacob Grimm)

"You shall know a word by the company it keeps."

(J. R. Firth, 1957)

Acknowledgements

I thank my advisory committee for their encouragement, careful assessment, constructive criticism, and teaching throughout my classroom work, research, and writing of papers, and this doctoral thesis. Professor Beltrán for his advice in applying neural networks, connecting me with others from the PUCP, and for being open to cross-disciplinary studies in artificial intelligence and linguistics. Professor Zariquiey for involving me in linguistics projects related to my studies in borrowing detection, and for collaboration in annotating the Pano-Tacanan borrowing database. Professor List for giving so abundantly of his time and expertise in the computer assisted identification of lexical cognates and borrowings; and in the writing, editing, upgrading and publishing of research in computational linguistics. A heartfelt thank-you to you all.

I thank members of the PUCP Artificial Intelligence group for their mutual interest and promotion of all things AI related. Especially Arturo Oncevay, for welcoming me as a colleague, for his AI leadership in linguistics projects, and for his encouragement in my doctoral studies, even as he pursued his own doctorate in Edinburgh, UK. Also, Franco Emanuel Pariasca, who helped me implement various advanced neural networks for use in lexical borrowing borrowing detection.

I am very grateful to the PUCP Post Graduate School for the "Hiuiracocha Scholarship" during the years 2019-21. I have been delayed in completing my thesis versus the original first half of 2022 schedule, and I thank the Beca Hiuracocha for their patience with me in this regard.

I appreciate the encouragement, effort, and teaching of professors from the University of Delaware, where I received my Masters degree in computer science. Special thanks to Professors Kathleen McCoy and Vijay Shankar. Thanks also to the University of Delaware graduate program of computer science which provided some financial assistance.

Friends who encouraged me in this pursuit include Jean Jacques de Coster, French consul in Cusco Peru, and dear friend and colleague from my DuPont career, Gary C. Myers.

I also want to thank my family in the United States and here in Peru.

Karen Ann Fenlon Miller (*in memoriam*) who encouraged me to study computer science as I had always wanted. My sons Nathan and Kenneth and their families who have always been encouraging, respectful and admiring of my effort to do the work, do the research, and earn the doctorate.

My wife, Maria Elena Mendoza Altamirano de Miller, who had the good sense to realize and tell me that it was time for me to get back to my love of computer science and computational linguistics. So while Elena pursued her post graduate studies in art history, I have pursued my doctoral studies in computational linguistics. Thank you Elena for your support, advice, encouragement, and love. *Te quiero.*

Abstract

Towards automatic detection of lexical borrowings in wordlists - with application to Latin American languages

by John Edward Miller

Key words: computer assisted, historical linguistics, computational linguistics, lexical borrowing, lexical borrowing detection, language model, neural network, machine learning classifier, sequence comparison methods.

Knowing what words of a language are inherited from the ancestor language, which are borrowed from contact languages, which are recently created, and the timing of critical events in the culture, enables modeling of language history including language phylogeny, language contact, and other novel influences on the culture. However, determining which words or forms are borrowed and from whom is a difficult, time consuming, and often fascinating task, usually performed by historical linguists, which is limited by the time and expertise available. While there are semi-automated methods available to identify borrowed words and their word donors, there is still substantial opportunity for improvement.

We construct a new language model based monolingual method, competing cross-entropies, based on word source groupings within monolingual wordlists; improve existing multilingual sequence comparison methods, closest match on language pairs and cognate-based on multiple languages; and construct a classifier based meta-method, combining closest match and cross-entropy functions. We also define an alternative goal of borrowing detection for dominant donor languages, which allows determination of both borrowing and source. We apply monolingual methods to a global dataset of 41 languages, and multilingual and meta methods to a newly constituted dataset of seven Latin American languages. We also initiate work on a dataset of 21 Pano-Tacanan and regional languages with added Spanish, Portuguese, and Quechua donor languages for subsequent application of borrowing detection methods.

The competing cross-entropies method establishes a benchmark for automatic borrowing detection for the world online loan database, the dominant donor multiple sequence comparison method improves over the competing cross-entropies method, and the classifier meta-method with sequence comparison and cross-entropy functions performs substantially better overall.

Resumen

Hacia la detección automática de préstamos léxicos en listas de palabras - con aplicación a lenguas latinoamericanas

por John Edward Miller

Palabras clave: asistida por computadora, lingüística histórica, lingüística computacional, préstamo léxico, detección de préstamo léxico, modelo de lenguaje, red neuronal, clasificador de aprendizaje automático, métodos de comparación de secuencias.

Conocer qué palabras de una lengua son heredadas, cuáles son prestadas, cuáles son de reciente creación y el momento de los eventos culturales críticos permite modelar la historia de la lengua, incluyendo su filogenia, el contacto entre lenguas y otras influencias culturales novedosas. Sin embargo, determinar qué palabras o formas son prestadas y de qué lengua provienen es una tarea compleja y laboriosa, realizada generalmente por lingüistas históricos, que se ven limitados por el tiempo y la experiencia disponibles. Aunque existen métodos semiautomáticos para identificar préstamos y sus lenguas de origen, aún hay margen de mejora.

Construimos un nuevo modelo de lenguaje basado en un método monolingüe, entropías cruzadas competitivas, basado en agrupaciones de fuentes de palabras dentro de listas de palabras monolingües; mejoramos los métodos existentes de comparación de secuencias multilingües, la coincidencia más cercana en pares de idiomas y afines basados en múltiples idiomas; y construimos un meta-método basado en clasificadores, combinando funciones de coincidencia más cercana y de entropía cruzada. También definimos un objetivo alternativo de detección de préstamos para idiomas donantes dominantes, que permite determinar tanto el préstamo como la fuente. Aplicamos métodos monolingües a un conjunto de datos global de 41 idiomas (WOLD), y métodos multilingües y meta-métodos a un conjunto de datos recién constituido de siete idiomas latinoamericanos. También iniciamos el trabajo en un conjunto de datos de 21 idiomas pano-tacana y regionales con idiomas donantes agregados de español, portugués y quechua para la posterior aplicación de métodos de detección de préstamos.

El método de entropías cruzadas competitivas establece un punto de referencia para la detección automática de préstamos en la base de datos mundial de préstamos en línea (WOLD). El método de comparación de secuencias múltiples del donante dominante mejora los resultados del método de entropías cruzadas competitivas. Finalmente, el meta-método clasificador, que combina la comparación de secuencias y las funciones de entropía cruzada, ofrece el mejor rendimiento general.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Research problem and objective

Lexical borrowing, the direct transfer of lexical material from donor to recipient languages, is one of the most pervasive processes of language evolution (Grant, 2014). We can almost see this process as it takes place, with the borrowing of Spanish words into several indigenous languages of Latin America, e.g., days of the week (*lunes, martes, ...*), cow (*vaca*), and borrowing from indigenous languages into Spanish as well, e.g., bean (*poroto*) and field (*chacra*) (see Tab. 1.1). While it took researchers much time to realize that languages were constantly changing (Campbell, 2013), there is evidence from ancient times that language communities were aware that they received lexical material from neighboring communities (Geisler and List, 2013). For example in Plato's *Kratylos* dialog (409d-10a) (Plato, 1921), Socrates addresses the difficulty in etymological studies when there are lexical borrowings. Estimates of lexical borrowing for languages, mostly from this epoch, from the World Loanword Database (http://wold.clld.org, Haspelmath and Tadmor, 2009) show a range of [1%, 62%] borrowing with a mean and standard deviation of 25% ± 13%.

Detection of *lexical borrowings* (borrowed words or *loanwords*) is an essential part of and crucial for the application of the *comparative method* in historical linguistics (Campbell, 2013). The *comparative method*, more appropriately meta-method, seeks to reconstruct ancestral languages, and describe language relationships (see Fig:1.1) and events. Detection of *borrowings* is also crucial for *phylogenetic reconstruction* which seeks to identify probable language phylogenies by which a family of languages evolved to their current state (Gray, Greenhill, and Atkinson, 2013).

Through synergy with other study areas in the natural, social, and historical sciences, we may be able to connect language events, e.g., language branching, language extinction, or exaggerated or accelerated language change, with the existence of events taking place in the human or ecological community at large. The innovation of words to describe new technologies, capabilities, behaviors, or religious or cultural customs, whether through lexical borrowing or other process, are signals of change. On a human dimension, the kinds of sound changes and word and language structures used within a language provide evidence for cognitive and even physical capabilities.

TABLE 1.1: Examples of lexical borrowing in recent time.

Language	Language family	Form	Donor	Donor form
concept: **Monday**				
Moseten	Mosetén-Chimané	roneš	Spanish	lunes
Cavinena	Pano-Tacanan	roneši	Spanish	lunes
Catuquina	Pano-Tacanan	segunda	Portuguese	segunda
ShipiboConibo	Pano-Tacanan	ronis-niti	Spanish	lunes
Yaminahua	Pano-Tacanan	ronĩs	Spanish	lunes
Itonama	Itonama	ulune	Spanish	lunes
Movima	Movima	lunes	Spanish	lunes
concept: **livestock**				
Moseten	Mosetén-Chimané	waka-ʔin	Spanish	vaca
concept: **cattle**				
Cavinena	Pano-Tacanan	waka	Spanish	vaca
EseEjja	Pano-Tacanan	waka	Spanish	vaca
Tacana	Pano-Tacanan	waka	Spanish	vaca
Catuquina	Pano-Tacanan	βoi	Portuguese	boi
concept: **cow**				
Aguaruna	Chicham	baka	Spanish	vaca
Yagua	Peba-Yagua	woka	Spanish	vaca
Aymara	Quechua-Aymaran	waka	Spanish	vaca
Catuquina	Pano-Tacanan	βoi_aĩ	Portuguese	boi
Yaminahua	Pano-Tacanan	βakka	Spanish	vaca
Cayuvava	Cayubaba	βaka	Spanish	vaca
Itonama	Itonama	u-waka	Spanish	vaca
ImbQuechua	Quechua-Aymaran	baka	Spanish	vaca
concept: **bean**				
Spanish	Indo-European	poroto	Quechua	purutu
Yagua	Peba-Yagua	purutu	Quechua	purutu
Tacana	Pano-Tacanan	poroto	Quechua	purutu
Catuquina	Pano-Tacanan	čičão	Portuguese	feijão
ImbQuechua	Quechua-Aymaran	purutu		

This table shows recipient language, family, and form along with corresponding donor language and form ordered by concept. The shear abundance of borrowing of days of the week (*Monday* shown here from Spanish 'lunes'), and for several concepts related to *cow* (from Spanish 'vaca'), creates a feeling of transfer of lexical material in real time. Lexical borrowing in these example cases, notably, spans concepts and several Peruvian language families.

Lexical borrowing is not directly comparable with other processes of language change. For example, sound change often proceeds in a regular manner that impacts most words in a particular language's lexicon where the sound occurs in a given phonetic context (Miller et al., 2020). In contrast, lexical borrowing depends much more on the initial language *contact situation* – the donor and recipient languages, the cultures involved, and the nature of the contact itself, e.g., mutually beneficial exchange, violent conquest, or gradual economic or cultural dominance. It has proven difficult to derive general rules or generalizations on

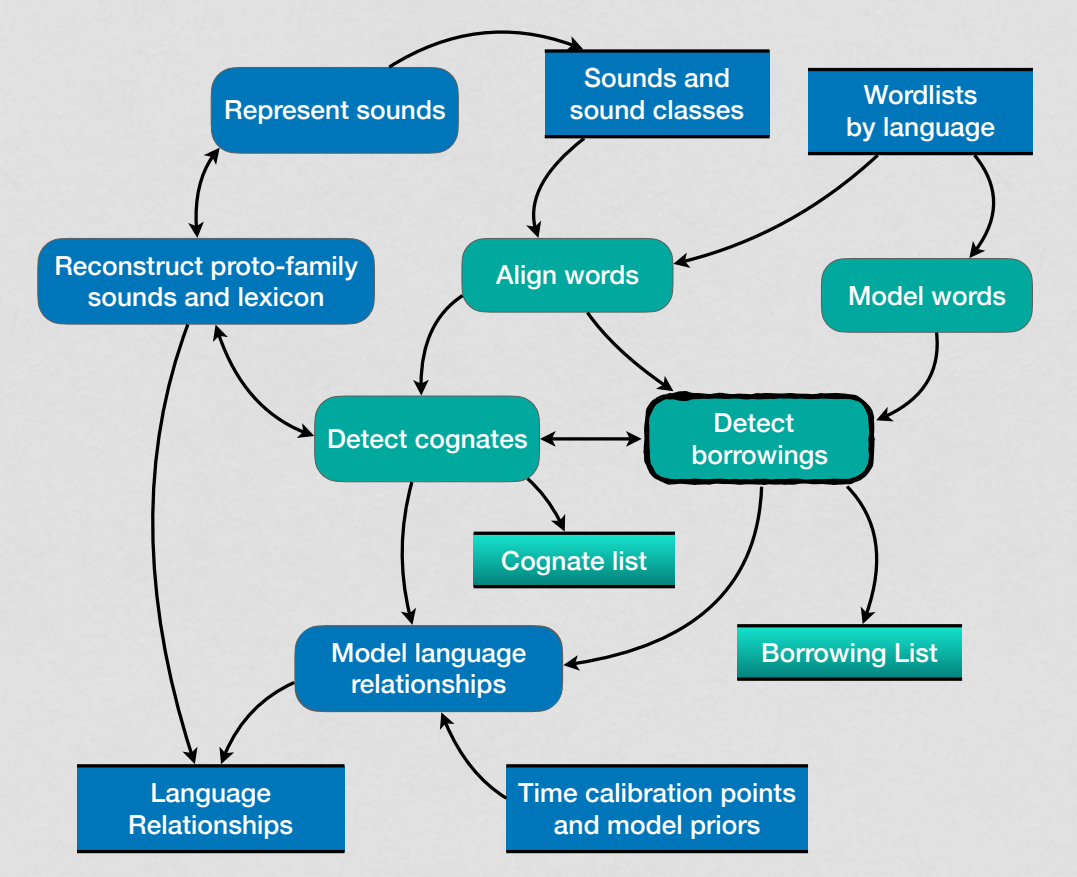


FIGURE 1.1: Comparative method - emphasis on lexical forms.
In this view of the comparative method, processes are rounded rectangles and data stores or reports are pairs of horizontal lines. Processes and stores in green (teal) are most related to this research, and are intermediate steps in service of the purpose of the comparative method of reconstructing ancestral families and modeling language relationships. Detect borrowings, our focus, uses wordlists and sounds and sound classes to develop language word models, align words, and detect cognates as inputs to detect and output borrowings.

lexical borrowing, although there is some agreement that some words are less likely to be borrowed depending on their semantics (Swadesh, 1952; Carling et al., 2019).

Detection of lexical borrowings is still one of the outstanding problems in historical linguistics, for computational approaches in particular (List, 2019b). With the availability of curated wordlists across languages and language families, and our ability to curate such wordlists as dictated by research direction, it becomes practical to develop or apply computer-assisted approaches in historical linguistics (Wu et al., 2020) to wordlists and discern lexical *borrowing* from lexical *inheritance*.

However, the issue of borrowed words is both subtle and profound. Words are received throughout the lifetime of language communities. What is obviously a borrowing from a recent epoch, becomes a subtle entity of unknown provenance when the loan comes from prior millennia (Heggarty, 2014). They become part of the language and have adapted more of the form of the language so that they

are hard to detect as borrowings (Kiparsky, 2014). Yet, indigenous language users often seem to know when a word doesn't seem to *fit* (Campbell, 2013) – where phonology, phonotactics, morphophonemics, or morphology, even syntax or semantics, is not consistent with common language patterns. Similarly, multilingual language users may recognize when words seem to *fit* across languages.

Objective: Our objective is to develop methods for automatic or semi-automatic detection of lexical borrowings from other lexical origins and apply these methods to wordlists organized by language, or by language and concept. Detection of lexical borrowings includes not only the decision of whether a form is *borrowed* or *inherited* (or other creative process), but also, where possible, the donor language, and likely donor form.

Benefits of such methods and applications include: 1. improved application of the *comparative method* and *phylogenetic reconstruction* for languages, 2. evidence of of language contact and impact, 3. increased borrowing detection productivity, 4. explicit, defined, and consistent detection processes, 5. benchmarking versus expert human performance,

1.2 Language wordlists

Nowadays there are abundant linguistic digital data resources available for use in research and as jumping off points for refining or developing new linguistic resources. We've worked exclusively with databases that follow the Cross-Linguistic Data Format (CLDF) standard (Forkel et al., 2018), which helps assure that data is accessible, shareable, and conforms to minimum quality standards where languages are linked to Glottolog (https://glottolog.org/, Hammarström, Forkel, and Haspelmath, 2021), concepts are linked to Concepticon (Concepticon, List et al., 2022a), and transcriptions may be optionally annotated to conform to broad international phonetic alphabet (BIPA or Broad IPA) conventions of the Cross-Linguistic Transcription Systems (https://clts.clld.org, List et al., 2021).

Wordlists for lexical borrowing minimally present data fields as shown in Tab. 1.2. Language names and concepts are accompanied by forms or values with language specific encoding (typically orthographic, but sometimes *phonemic*). Values are also translated uniformly into "sound segments" of Broad IPA tokens (List et al., 2021) providing consistency of representation across languages. The "Borrowed" field, either as a *boolean* or score, would be present in a training or test dataset, where it serves to teach or validate borrowing detection decisions, or in the predicted dataset as a result of borrowing detection. The donor language and even donor form (typically orthographic) more completely define the borrowing, where available.

Specific datasets we use in this investigation are:

1. World Online Loan Database (WOLD) (http://wold.clld.org, Tresoldi, Forkel, and Morozova, 2019; Haspelmath and Tadmor, 2009). This was

TABLE 1.2: Snippet of Imbabura Quechua wordlist.

Language	Concept	Value	Segments	Borrowed	Donor
Swahili	World	dunia	d u n i a	True	Arabic
TarifiytBerber	Valley	tizi	θ i z i	False	
English	Calm	calm	k ɑː m	True	French
Mapudungun	Foam	tronün	t s o n ɨ n	False	
Quechua	World	pacha mama	p a tʃ a + m a m a	False	
Quechua	Valley	yunga	j u ŋ g a	False	
Quechua	Foam	putsuju	p u ts u x u	False	
Quechua	Knife	kuchillo	k u tʃ i ʒ u	True	Spanish

From Miller, Pariasca, and Beltran Castañon (2021).

recently curated to add harmonized phonetic transcriptions as segmented Broad IPA tokens (Tab. 1.2). WOLD contains 41 wordlists with Concepticon glosses, in English, for 1,460 distinct concepts with wordlists varying in size from 956 to 2,558 word forms. Not all concepts are represented in all languages, and some languages have multiple words for the same concept. Phonetic transcriptions follow the unified *Broad IPA* transcription system (Miller et al., 2020) from the Cross-Linguistic Transcription Systems reference catalog (Anderson et al., 2019). WOLD annotates not only the likely borrowed status, but also the donor language and likely donated word form. To only consider clear-cut borrowings in our tests, we treated as borrowed, only the words labeled as *clearly borrowed* (Miller et al., 2020). The resulting database with phonetic transcriptions was curated using the CLDFBench toolkit (Forkel and List, 2020) and stored in a Cross-Linguistic Data Format (CLDF, (Forkel et al., 2018)). Besides using all of WOLD for development of monolingual methods in §2, a subset of Latin-American wordlists was used to develop our own SABor database in §3.

2. German Wordlist. This is used to construct artificial borrowings in simulation of a recent high intensity language contact event (see §2.2.1). This wordlist is taken from a German etymological dictionary (Kluge, 2002). Phonetic transcriptions were added with modifications from the CELEX database (CELEX).

3. Intercontinental Dictionary Series (IDS) (https://ids.clld.org/, Key and Comrie, 2015). This too is available as a CLDF database. An informal standard uses bracketing of the portion of the value that is thought to be borrowed. Encoding of word forms, annotation of borrowing, donor language and donor word form depends on individual wordlist authors. IDS and WOLD share most concepts in common. The Spanish wordlist from IDS is used in developing our SABor database in §3.1. Some 21 Peruvian and regional language wordlists from IDS, including Spanish, Portuguese, and Imbabura Quechua (from WOLD) are being used to develop our planned Pano-Tacanan Borrowing database.

4. Glottolog's (https://glottolog.org/, Hammarström, Forkel, and Haspelmath, 2021) goal is to be a "comprehensive reference for the world's languages, especially the lesser known languages". Workflow that constructs CLDF datasets, such as WOLD and IDS above, incorporate essential Glottolog data such as language names, identification, family, macro-area and geographic location. Similarly, in our workflows creating SABor and our planned Pano-Tacanan Borrowing databases, we incorporate essential Glottolog data.

5. Concepticon (Concepticon, List et al., 2022a) provides lists of concepts, tools for accessing and linking concept lists, and a backbone comprehensive concept list. Individual concept lists may emphasize different semantic fields, ontological relations, concept relations, gloss languages, or other. Workflow that constructs CLDF datasets, incorporates essential concepticon data, concepticon id and concepticon gloss at a minimum. Our SABor and planned Pano-Tacanan Borrowing databases incorporate essential Concepticon data.

1.3 Previous work

Historical linguists detect lexical borrowings using a toolkit of different techniques aimed at detecting conflicts or similarities in the data for individual words (List, 2019a). Techniques can be categorized between monolingual, multilingual, and (phylogenetic) model discrepancy. We consider each in turn providing both problem context and state of the art.

1.3.1 Monolingual methods:

For many language contact situations, when borrowed words enter into a language, they still retain certain, sometimes even most, of their donor language properties. These may include specific phonological properties ("foreign sounds") or *phonotactic* properties ("foreign sound patterns"), which may disappear over time through *loanword nativization* (Trask, 2000, p. 200).

Although erased or replaced over time, borrowing *language-internal evidence* is observable over many languages and language families, as seen in examples from Miller et al. (2020):

> In many Hmong-Mien languages, for example, some Chinese words are borrowed with a very specific tone that only occurs in Chinese words (Qiguang, 2013). Similarly, it is easy for German speakers to identify job as a loan from English, since only in borrowed words the grapheme *j* is pronounced as [dZ] in German. In the same line, but in a radically different context, speakers of Iskonawa, an obsolescent Panoan language spoken in Central Peruvian Amazonia can easily identify loanwords from Shipibo-Konibo, the dominant language in the area, due to straightforward phonological features. For instance, Iskonawa has dropped word-initial [h], thus forms like

> [hana] 'tongue' or [huni] 'man' are easily detected as loanwords from Shipibo-Konibo. (Miller et al., 2020)

Language-internal evidence may include particular donor language constructions, phonotactic elements such as particular consonant clusters or vowel combination, or foreign stress patterns (Maddieson, 1986; Grossman et al., 2020).

While we expect language-internal evidence for lexical borrowing to dissipate over time as words adapt to their recipient language, we want to test how well such evidence would aid in detection of lexical borrowings. Assuming that phonology and phonotactics provide the strongest evidence for lexical borrowing, "all that we need to do in a computational approach to monolingual borrowing detection is to derive computational models of phonology and phonotactics from annotated wordlists of a given language and then calculate to which degree a word resembles a typically inherited or a typically borrowed word" (Miller et al., 2020).

We use different lexical *language models*, where a *language model* refers to "any system trained only on the task of string prediction, whether it operates over characters, words or sentences, and sequentially or not" (Bender and Koller, 2020). Our language models are based on lexical data provided in the form of wordlists, with words represented by segmented IPA phonetic transcriptions. Models are trained on a *training* part of the wordlist and then applied to a *test* part of the wordlist to verify detection of lexical borrowings.

While much of the evidence linguists employ to detect borrowed words is based on the comparison of *several* languages, conflicts in phonology and phonotactics, detected by monolingual methods, are also used for borrowing detection (Miller et al., 2020). Such ought to be particularly effective when dealing with recent borrowing events.

State of the Art

Mi et al. (2016), Mi et al. (2018), and Mi, Xie, and Zhang (2020) have enjoyed some success for the specific case of predicting lexical borrowings to Uyghur language from Chinese, Russian and Arabic languages. Their initial approach (Mi et al., 2016) is monolingual using a recurrent encoder-decoder neural network for borrowing detection. F1 validation scores on cross-domain data are in the 0.79 to 0.80 range. Subsequent approaches add a significant multilingual component and are discussed below §1.3.2.

Miller et al. (2020) develop language models for inherited and borrowed words for individual languages from the World Loanword Database (WOLD) (Haspelmath and Tadmor, 2009) using Markov chain and recurrent neural network models, trained on inherited and borrowed word datasets. They compare word entropies for inherited and borrowed language models in order to identify borrowings based on monolingual information alone. F1 scores on borrowing detection averaged 0.604 over the entire WOLD database. This effort forms part this thesis.

Cristea et al. (2021) use support vector machine (SVM) and recurrent neural network (RNN) direct borrowing classification to discriminate Latin borrowings in Romance languages from inherited words (which come largely from a more ancient Latin). The feature rich support vector machine outperforms direct classification by recurrent neural network. F1 scores for detection of Latin borrowings over several Romance languages sampled from Wikitionary ranged from 0.86 to 0.92.

1.3.2 Multilingual methods:

Whereas monolingual methods assess whether borrowed words differ from inherited words in the recipient (target) language, and use such discrepancies to classify words as inherited or borrowed in test data, multilingual methods assess whether donor language words are similar to recipient language words for similar word meanings. Monolingual and multilingual methods are highly complementary, and we shall discover in §3.2.3 how both can work well together for improved lexical borrowing detection.

Multilingual methods can be grouped into sequence comparison methods versus feature based classifier methods. In sequence comparison methods string sequences over multiple languages are compared, usually after some alignment process, and decisions made as to whether words are related across languages or language families based on some similarity or distance measure.

Normalize edit distance (NED) directly calculates distances between phonetic sequences with costs for additions, deletions, replacements, or interchanges of sequence elements with normalization by the longer sequence. Words *adobe* and *fazofe* seem far apart (Fig. 1.2) based on the NED method.

Language	Word	Segments	Alignment	Edit Costs
Spanish	adobe	a ð o β e	- a ð o β e	1 0 1 0 1 0
Mapudungun	fazofe	f a z o f e	f a z o f e	

FIGURE 1.2: Normalized edit distance - example.

Sound class based methods cluster phonetic segments into sound classes and then compute distances between sound class sequences (List, 2012). The sound-class based alignment method (SCA) provides sound class categories, similarity measures between sound class categories, scoring functions for distance measures, and modifiable gap scores based on prosodic context (List, 2012; List et al., 2018). Words *adobe* and *fazofe* seem close together, while *adobe* and *alulis* seem far apart (Fig. 1.3) based on SCA.

When words from more than two languages are compared at the same time, a pairwise method such as normalized edit or sound-class alignment distance methods is incorporated into a multiple alignment and comparison method, which clusters similar word forms. Words for the concept adobe in Fig. 1.4 are clustered into words derived from *adobe* where *alulis* is left out, and words similar to *saami* from Mexican indigenous languages.

Language	Word	Segments	Sonority	Prosody	Sound Cl	Alignment
Spanish	adob	a ð o β e	7 3 7 3 7	XBYBZ	ADUBE	- a ð o β e
Mapudungun	fazofe	f a z o f e	3 7 3 7 3 7	AXBYBZ	BASUBE	f a z o f e

Language	Word	Segments	Sonority	Prosody	Sound Cl	Alignment
Spanish	adobe	a ð o β e	7 3 7 3 7	XBYBZ	ADUBE	a ð o β e -
Wichi	alulis	a l u l i s	7 5 7 5 7 3	XBYBYN	ALYLIS	a l u l i s

FIGURE 1.3: Sound-class alignment method - example.

Language	Word	Segments	Sonority	Prosody	Sound Cl	Alignment	BorId
Spanish	adobe	a ð o β e	7 3 7 3 7	XBYBZ	ADUBE	- a ð o β e	45
Imbabura Quechua	adubi	a d u b i	7 1 7 1 7	XBYBZ	ATYPI	- a d u b i	45
Wichi	alulis	a l u l i s	7 5 7 5 7 3	XBYBYN	ALYLIS	- a l u l i	FN
Mapudungun	fazofe	f a z o f e	3 7 3 7 3 7	AXBYBZ	BASUBE	f a z o f e	45
Yaqui	saami	s aː m i	3 7 4 7	AXBZ	SAMI	s aː m i -	***
ZinacantanTzotzil	shamit	ʃ a m i tʰ	3 7 4 7 1	AXBYN	SAMIT	ʃ a m i tʰ	***
Qeqchi	xan	ʃ a n	3 7 4	AXN	SAN	ʃ a n - -	***

FIGURE 1.4: Cognate method - SCA multiple alignment - example.

In feature based classifier methods, multiple features from recipient and donor languages are selected for use by the classifier. Features may include phonological and phonotactic elements, e.g., presence of particular sound segments or sound segment sequences, morphological elements, sentence elements if available, and even punctuation and capitalization in the case of text. Classifiers themselves can be any of logistic regression, support vector machines, advanced neural networks, log-linear model, or other.

State of the art - sequence comparison methods

The state of the art for sequence comparison methods is essentially the same as we reported in Miller and List (2023):

> Early studies by van der Ark et al. (2007) and later Mennecier et al. (2016) compute edit distances between words from genetically unrelated languages and compare distances to thresholds, in order to detect borrowed words in multilingual wordlists.
>
> Zhang et al. (2021) compare borrowing detection performance for edit distance versus SCA distance (List, 2012; List et al., 2018), finding that SCA outperforms edit distance in borrowing detection accuracy. Hantgan, Babiker, and List (2022) build on this work, using dedicated methods for automated cognate detection applied to languages from different language families in order to identify clusters of related words resulting from lexical borrowing. List and Forkel (2022) expand this work further, by applying a two-stage workflow in which they first identify language-family-internal cognates, using a method specifically apt for the detection of deep cognates, and then compute

SCA distances between cognate sets from genetically unrelated languages in order to infer sets of words related by lexical transfer.

Kaiping and Klamer (2022) use automated methods for cognate detection (List, Greenhill, and Gray, 2017) on a target set of Timor-Alor-Pantar languages. In order to infer borrowings from Indonesian and Tetun (not in the target set), they include both languages in their sample and treat all cognate sets that involves words from either of the two languages as borrowings. Moro, Sulistyono, and Kaiping (2023) apply a similar approach to investigate borrowings in Alorese. (Miller and List, 2023)

State of the art - classifier methods

Mi et al. (2018)'s multilingual approach is based on cross-lingual embeddings constructed using word embeddings from monolingual models with the help of bilingual Uyghur and donor dictionaries. The approach is still focused entirely on the specific case of Uyghur as the recipient language now with Chinese, Russian, Turkish and Arabic donor languages. Part of speech (POS) and sentence level features were added to the previous work based on cross-lingual embeddings with borrowing prediction now using a log-linear model (Mi, Xie, and Zhang, 2020). This is a complex borrowing detection work-flow focused on and developed for just the Uyghur language. Resulting test F1 scores on borrowing detection range from 0.72 to 0.74.

Nath et al. (2022) trains binary classifiers, mainly advanced neural network based, on large wordlists to predict borrowed words. They achieve F1 scores in the 0.75 to 0.85 range. Corpora were largely majority languages with some lower resource languages represented as well. Wordlists were scraped from Wiktionary resulting in 16 different recipient-donor wordlists with 15 unique recipient languages and 10 unique donor languages. Their workflow seems cumbersome, compute intensive, and not minimalist. The large constructed wordlists are not comparable to standard wordlists such as WOLD or IDS, or Swadesh, so it is difficult to know how the seemingly promising results compare to our own work.

Note that for both classifier methods discussed here, recipient and donor languages were processed pairwise, i.e., for each recipient language, just Uyghur in the case of (Mi, Xie, and Zhang, 2020), borrowing detection and evaluation was versus a single donor language. This is similar to the *dominant donor* concept that we introduce in §2 and define more explicitly in §3.

1.3.3 Discrepancies from phylogenetic models:

In this model discrepancy approach to borrowed word detection, one constructs phylogenetic models (trees) of language families based on wordlists, optionally including intruder languages that are typically not part of the language family, and then looks for phylogenetic conflicts (Minett, Wang, and Kong, 2003; Nakhleh, Ringe, and Warnow, 2005; Nelson-Sathi et al., 2011; List et al., 2014a;

List et al., 2014b; List, 2015; Willems et al., 2016). Observed discrepancies in the model, in particular lexical items that detract from hierarchical family relations and contribute instead to lateral transfers, are likely due to borrowed words (List et al., 2014a; Delz, 2014). Recently, Neureiter et al. (2022) introduced Bayesian phylogenetic language models with horizontal transfer, called `contacTrees`, to better handle language events where there is language contact and borrowing in addition to inheritance. While we don't take advantage of discrepancies from Phylogenetic models in our research, such advances as this (Neureiter et al., 2022), offer promise.

1.4 Research directions

All methods developed in this work are based on sequences of lexical forms represented as segmented IPA tokens.

Monolingual borrowing detection (§2): Lexical language models are trained and then used to compute cross-entropies for lexical forms. In cross-entropy and competing cross-entropies approaches, cross-entropy results are used in decision procedures to classify lexical forms as borrowed or inherited. Both Markov chain and recurrent neural network (RNN) language models are used to compute cross-entropies, with decision results compared to a baseline support vector machine (SVM) using a simpler *bag of sounds* classifier approach.

A lexical language model on inherited forms serves as a dramatically simplified model of a proficient indigenous language speaker. Decisions based on this model, where cross-entropies are compared to a critical value, are analogous to categorizations by an indigenous speaker of words as from or not from their language.

A significant innovation we made to this approach, competing cross-entropies, employs lexical language models each for inherited and borrowed lexical forms. Cross-entropies are calculated by each language model and the lesser cross-entropy model determines whether the lexical form is inherited or borrowed. This innovation produces improved borrowing detection over the simpler single inherited lexical form model and the baseline *bag of sounds* approach. Between Markov chain and recurrent neural network language models, the neural network model produces marginally better results, but at a cost of substantially increased complexity and increased training times. Subsequent experiments with neural network Transformer language models reduce training times, with little improvement in borrowing detection. Experiments using multiple donor models, instead the single *borrowed* category, also show little improvement in borrowing detection. Further experiments with data augmentation for borrowed lexical form models result in changes in the distribution of precision and recall with little net change in borrowing detection.

We also explored using direct neural network models of lexical forms, where neural models are trained and produce an output state that is used to predict

borrowed versus inherited lexical forms. Using only the output state of transformer based sequence models, borrowing detection results are inferior to that produce by the competing entropies approach.

Multilingual borrowing detection (§3): Sequences of lexical forms are aligned and compared across languages for the same or similar concepts. Forms which are sufficiently similar to one another are *cognate*, either inherited from the same proto form or borrowed.[1] We include likely contact languages in the sample of multiple languages, and attribute forms as borrowed from the intruder when found to be cognate with the intruder form. Both *Closest Match*, pairwise, and *Cognate-Based*, multiple alignment, methods are explored, in each case running trials with normalized edit distance (NED) and sound class phonetic alignment (SCA) methods. Input to the methods in each case are the same segmented IPA forms as for monolingual methods, but the approach aligns and compares sequences across languages. NED counts the number of insertions, deletions, and replacements of IPA segments needed to match, while SCA transforms IPA segments into classes of similar sounds and performs weighted scoring and alignment taking into account sequence position and prosody to match (NED and SCA shown above in Fig. 1.2).

An innovation we made to this approach is to characterize the likely contact language as a dominant donor or intruder. We focus on the dominant donor for borrowing detection and assert the direction of borrowing and the borrowed lexical form based on whether the dominant donor lexical form is cognate with lexical forms from other languages in the pairwise or multiple alignment.

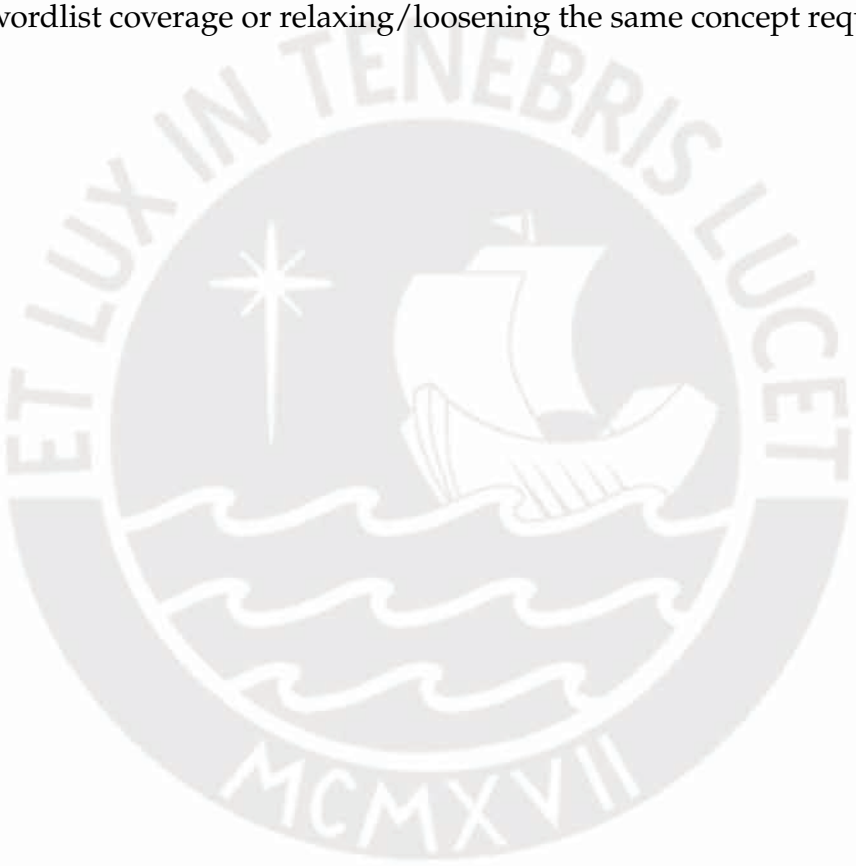
Another innovation we made to this approach is to combine pairwise and multiple alignment methods via a *Classifier* initially implemented as a support vector machine (SVM). Individual multilingual methods, NED and SCA, are somewhat complementary, and so the use of a *Classifier* meta-method which combines individual methods produces better overall borrowing detection than any single method.

Based on a error analysis of *Classifier* results, we learned that non-algorithmic factors have a greater impact on borrowing detection than individual sequence alignment methods. Adequate representation of donor source lexical forms for individual concepts, i.e., coverage of the donor wordlist, opens up the possibility for a 5% to 10% improvement in recall. Another problem is that borrowings are not always from the same concept. We experimented with relaxing the same concept requirement in the pairwise case, and without any concept restriction that the numerous false cognates detected makes the approach unusable, while restriction to the same central concept, provides only a minimal improvement with false cognates offsetting increased borrowings detected. Perhaps a more complex option which considers both some measure of semantic distance along with NED or SCA distance would result in some improvement.

[1] Historical linguistics' use takes cognates as inherited from the same form. Here we follow the computational linguistic practice of either inherited or borrowed forms where borrowing in this work is across languages or sub-groups.

An advantage of the monolingual cross-entropy methods is that they do not depend on the wordlists of other language lexical forms, specifically not those of possible donor forms. Nor do they depend on the concept represented by the word, so a same or similar concept restriction does not apply. Monolingual cross-entropy methods are complementary to multilingual sequence comparison methods.

A significant innovation we made is to form a *Classifier* meta-model combining monolingual *Least Cross-Entropy* (LCE) and multilingual *Closest Match* (CM) methods. This *Classifier* combines complementary Markov chain competing cross-entropies models (LCE method) with pairwise sequence matching multilingual models (CM method). There is a substantial resulting improvement in borrowing detection for the combined model, even without making the improvements in wordlist coverage or relaxing/loosening the same concept requirement.



Chapter 2

Monolingual borrowing detection

Historical linguists, in order to detect lexical borrowings (see §1.1), make use of various strategies, combining evidence from multiple sources. Even with increased popularity of computational linguistics, automated approaches to detect lexical borrowing are still early on, sometimes simplifying the problem and disregarding evidence that would be routinely considered by human experts. An example for this kind of language-internal evidence are phonological and phonotactic clues that are especially useful to detect borrowings that have not been completely adapted or assimilated into their recipient languages.

In this chapter, we test how such clues can be exploited in automated frameworks to detect borrowings. By modeling phonology and phonotactics with a simple support vector machine model (set of sounds), and Markov chain and recurrent neural network language models (sequences of sounds), we develop a framework for the detection of borrowings via supervised learning in monolingual wordlists. Using a substantially revised dataset in which lexical borrowings have been thoroughly annotated for 41 different languages from different families, featuring a large typological diversity (Haspelmath and Tadmor, 2009; Tresoldi, Forkel, and Morozova, 2019), we apply these models in series of experiments to investigate their performance in monolingual borrowing detection (Miller et al., 2020).

In a significant innovation, language models (Markov chain or neural network) are constructed and trained from *inherited* and *borrowed* words separately, and then used to compute cross-entropies on test (held-out) words. In this competing cross-entropies approach, models which calculate the least cross-entropy for a word, *win* that word and so categorize the word according to the winning least cross-entropy model.

Results appear unsatisfying at a first glance, but further tests show that method performance improves with increasing amounts of attested borrowings, especially in those cases where most borrowings originated from a single dominant language (Miller et al., 2020). A preliminary conclusion is that phonological and phonotactic clues derived from monolingual language data alone appear insufficient to detect borrowings when used in isolation. Based on our detailed findings, however, we are hopeful that (1) monolingual methods would be useful in integrated approaches that also take multilingual information into account, and

(2) more powerful or refined neural network models to detect borrowings might improve performance.

A Transformer based lexical model is developed to improve detection performance and experiment further. Transformer experiments performed are: (1) a direct model alternative to the competing entropies approach, (2) lexical *donor* models replacing the single lexical *borrowed* model with multiple competing cross-entropies, and (3) data augmentation via a supplementary dominant donor language wordlist.

Results are still not convincing, and so we explore multilingual borrowing detection methods in §3. Subsequently a monolingual method is incorporated into multilingual methods in order to reap the benefit of both approaches in.

Parts of this chapter were previously reported in (Miller et al., 2020; Miller, Pariasca, and Beltran Castañon, 2021) and will be cited as appropriate.

2.1 Methods

Our monolingual methods encompass: 1. Language models of sound sequences which produce cross-entropy estimates of lexical forms and are used to discriminate between inherited and borrowed words, 2. Direct borrowing models of lexical forms, some of which include language model components, which produce inherited or borrowed word decisions, 3. Data augmentation methods with the intent of improving results for cross-entropy based language models (above). Standard precision, recall, and F score measures used to evaluate experimental results are also explained.

2.1.1 Lexical language models and cross-entropy

Lexical language models used here which estimate cross-entropy are:

1. Markov chain models (Markov, 2006; Shannon, 2001; Jurafsky and Martin, 2009),
2. Recurrent neural network models (Bengio et al., 2003), and
3. Light-weight transformer models (Bahdanau, Cho, and Bengio, 2015; Vaswani et al., 2017).

Markov chain models represent "words by their sound *n-grams*, and *neural network* models ... in the form of sequences of learned vector representations of sounds" (Miller et al., 2020). Markov and neural models take into account the sequence of sounds modeling both the phonology and phonotactics of words. Light-weight transformers provide a more current alternative to recurrent neural networks with more power and flexibility at a reduced cost in parameters and computation. More recent advances such as Bi-directional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) and the alphabet mash of ElMo, RoBERT, and GPT@ estimate millions, billions, and even trillions of parameters and cost days, weeks, and even months of training time. While these

are dramatic overkill for our problem, the idea of bi-directional modeling with a masked language model such as in mBERT, a light-weight mBERT, merits future consideration.

In the subsection on direct borrowing models (§2.1.2), a *bag of sounds* model is presented that only takes into account the presence of sounds in a word. Such simple language model considers only the phonology of words and serves as a point of reference in contrast with the sequence models covered here.

Markov chain model. The methods section of Miller et al. (2020) provides a detail explanation of the Markov chain language model:

> An $n-1$ order Markov chain model, emits a sound segment with probability dependent on the $n-1$ previous sound segments (an n-gram model). The product of sound segment probabilities estimated by the Markov model are transformed into per sound segment word entropies which are then used in borrowing detection.
>
> We use a second order Markov model, a 3-gram model, from the Natural Language Toolkit (NLTK) (Steven Bird and Loper, 2019). In the second order model, the emission probability, $P(c_k|c_{k-2}^{k-1})$, is conditioned on the previous 2 sound segments. The second order Markov model is local with longer range effects resulting from the second order probabilistic process.
>
> We can approximate the probability of a sequence of n sound segments that make up a word, $P(c_1^n)$, by the product of the n second order conditional probabilities:
>
> $$P(c_1^n) \approx \prod_{k=1}^{n} P(c_k|c_{k-2}^{k-1}).$$
>
> (Miller et al., 2020)

We transform word probability estimates to length normalized cross-entropies given the word model estimates,

$$H(w, m) = -(1/n) \log P(c_1^n).$$

Miller et al. (2020) discusses characteristics of the cross-entropy distribution, and challenges of parameter estimation. For the Markov chain word model case here

> cross-entropy typically exhibits a smooth distribution with moderate right skew for wordlists when the model fits well. The second order model with a sound segment vocabulary size V requires V^3 probability parameters for sound segment emission probabilities conditioned on the previous two sound segments.

> With wordlists of just 1,000 to 2,500 word forms and a typical sound segment vocabulary size of $V \approx 50$, estimating $50^3 = 125,000$ parameters by maximum likelihood would cause sparse parameter estimation with problems of both undefined conditional probabilities and overfitting. We use interpolated Kneser-Ney smoothing to accommodate unseen tri-grams, reduce overfitting, and reduce the number of estimated parameters to less than the V^3 required under maximum-likelihood. (Miller et al., 2020)

A tally of non-zero n-gram counts used in estimating Markov chain emission probabilities, reveals $\approx 12,500$ non-zero n-gram counts for a typical second order Markov chain, 3-gram, language model.

Recurrent neural network. The methods section of Miller et al. (2020) also provides a detail explanation of the recurrent neural network language model:

> Recurrent neural networks provide word length order conditioning via the recurrent layer with memory. Word probabilities are expected to be better estimated, i.e., better approximating human performance, than for the Markov chain model, as we can infer from early work of language modeling by (Bengio et al., 2003) and more recent work with transformer language models (Vaswani et al., 2017).
>
> Conditional sound segment emission probabilities are dependent on and estimated from all earlier sound segments of the current word:
>
> $$P(c_k|c_1^{k-1}) = f(c_{k-1}, ..., c_1).$$
>
> We can approximate the probability of a sequence of n sound segments that make up a word, $P(c_1^n)$, by the product of the n corresponding conditional probabilities:
>
> $$P(c_1^n) \approx \prod_{k=1}^{n} P(c_k|c_1^{k-1}).$$
>
> (Miller et al., 2020)

Word probability estimates are again transformed to length normalized cross-entropies given the (word, model) estimates,

$$H(w, m) = -(1/n) \log P(c_1^n).$$

Miller et al. (2020) delves into the details of recurrent neural model architecture (Fig 2.2):

> The challenge and advantage of the recurrent neural network method is in the estimation of the conditional sound segment probabilities, with the function $f(c_{k-1}, ..., c_1)$, using a more complex architecture but with fewer parameters (Tab. 2.1) than the second order Markov model. Sparse indicator vectors, c_k, representing sound segments

are transformed into dense real input vectors, x_k. In the recurrent layer, input vectors, x_k, and prior hidden state vectors, h_{k-1}, are linearly transformed and passed through a *tanh* activation function to produce current hidden state, h_k, and output, o_k, vectors. Resulting output vectors are linearly transformed in a dense output layer of logits, y, representing possible output segments. The $softmax$ activation function transforms logit values y_k into sound segment probability estimates,

$$\widehat{P}(c_n|c_{n-1},...,c_1) = e^{y_{cn}} / \sum_k e^{y_k}.$$

(Miller et al., 2020)

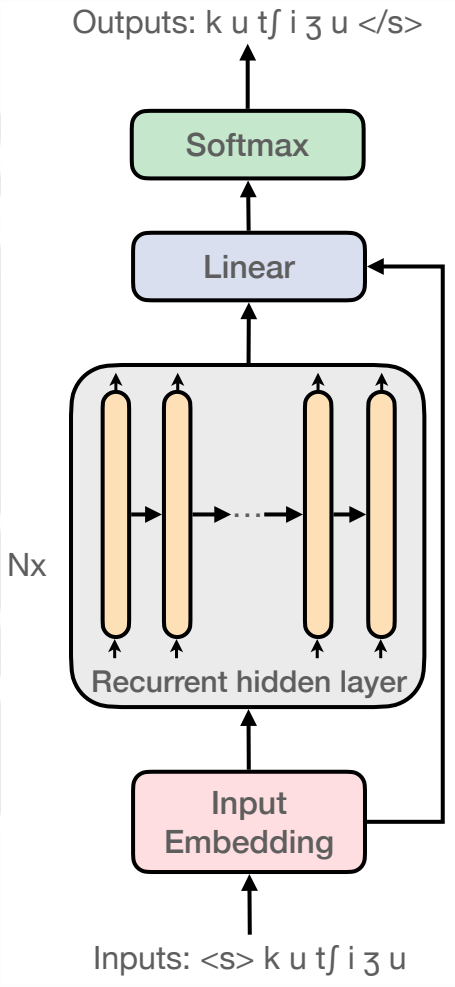


FIGURE 2.1: Recurrent neural network lexical model.
From Miller, Pariasca, and Beltran Castañon (2021).

Model parameter counts, implementation details, and model fitting challenges are also discussed by Miller et al. (2020):

> While the recurrent neural network model requires a high baseline number of parameters given its embedding length and recurrent layer length, the growth in number of parameters is just linear with the vocabulary size. As a result, the number of parameters in the neural

network is on the order of 10,000, which does not change much with the vocabulary size. Furthermore, the number of parameters does not increase as a power of the word length in sound segments even though the conditioning is on all previous sound segments.

We implement our recurrent Neural Network in Tensor-Flow 2.2 (Abadi et al., 2015) and parameterize the model to permit ready changes in architecture, regulation, and fitting parameters during experimentation. The [architectural] configuration used in this study is shown in Fig 2.1 [with corresponding parameter settings in Tab. 2.1]. Neural network models, even with just thousands of parameters, may suffer from substantial variance between training and test due to overfitting, especially when the amount of training data is comparatively small as in this case. We apply methods of dropout and l2 regulation to reduce overfitting. (Miller et al., 2020)

TABLE 2.1: Recurrent and Transformer Model Parameters

Parameter	Recurrent	Transformer
# Architetural		
embedding_len	32	32
hidden_layer_len	32	32
hidden_layer_cell_type	GRU	LSTM
n_layers	1	1
# Regulation		
embedding_dropout	0.0	0.3
recurrent_l2	0.001	
recurrent_output_dropout	0.2	
merge_embedding_dropout	0.2	
attention_dropout		0.2
transformer_dropout		0.1
# Model fitting		
epochs	45	80
learning_rate	0.01	0.0055
learning_rate_schedule	Decay	Transformer
learning_rate_decay_factor	0.95	
val_split	0.0	0.0
number of parameters	$\approx 13,000$	$\approx 10,000$

Adapted from Miller, Pariasca, and Beltran Castañon (2021).

Light-weight transformer neural network. In more recent neural network language modeling, (Bahdanau, Cho, and Bengio, 2015) incorporated *attention* into recurrent models to better retain important information over long sequences through a learned *attention* mechanism. Subsequently, (Vaswani et al., 2017) discovered that "Attention is all you need" and replaced the recurrent and attention layers with a reformulated attention layer, also adding *addition & normalization* layers to makeup the current *transformer* architecture. In language modeling of

text at the sentence level or greater for complex applications such as machine translation, many transformers may be stacked together to compose the resulting language model.

The model used here from (Miller, Pariasca, and Beltran Castañon, 2021) replaces the recurrent neural model with a "light-weight Transformer module (Vaswani et al., 2017) which includes Attention (Bahdanau, Cho, and Bengio, 2015) and Transformer features of an adding and normalization layer, and a feed forward layer (Fig 2.2). The model uses a forward only (left-to-right) causal model — this reduces model complexity and avoids unintended dependencies between inputs and outputs." Conditional sound segment emission probabilities, word emission probabilities, cross-entropies given (word, model), and *softmax* estimation of sound segment probabilities is the same as for the recurrent neural network.

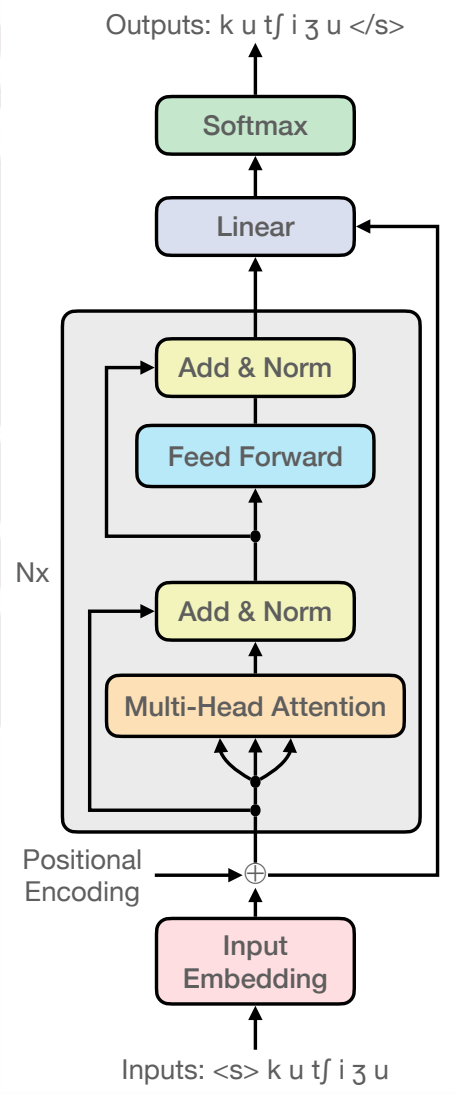


FIGURE 2.2: Light-weight transformer lexical model.
From Miller, Pariasca, and Beltran Castañon (2021).

Parameters settings for the transformer model are shown in Tab. 2.1. Regulation is applied appropriate to the Transformer module, learning rate is reduced almost by half and training epochs increased almost by half.

The expectation is that this model will execute much more rapidly than recurrent neural network model, since the attention mechanism operates in parallel and not sequentially as recurrent neural network, and that model prediction will also improve since the attention mechanism should focus on more important data over the entire sound sequence.

Cross-entropy based decision procedures

We take a theoretical linear model based look at the inherited word only, and borrowed versus inherited word decision procedures. In both cases cross-entropy is used as the decision metric.

Inherited word cross-entropy. We can think of our lexical language models for words and corresponding cross-entropy estimates for words given the model, as offering some measure of lexical expectation versus surprise, that might approximate some aspects of human lexical processing. Entropy has been used to characterize lexical predictability in natural reading (Lowder et al., 2018). Some human sensor processes are *log* based just as the cross-entropy. [1]

It is claimed that a native language users often "know when something is *off* about a word" Campbell, 2013, and so can often distinguish word source (e.g., inherited versus recently borrowed). We use cross-entropy as a computational surrogate for this human sense of a word being *off*.

As a thought experiment, let's think of cross-entropy in terms of a mixed effects linear model. This is not meant to be definitive, but it does help explain why competing cross-entropies is a superior approach to inherited only cross-entropy.

A useful mixed effects linear model of cross-entropy is:

$$H(m, s, w) = \mu_m + \mu_s + \mu_{m,s} + \mu_{w(s)} + \mu_{m,w(s)} + \epsilon,$$

where μ_m is the language model main effect (fixed), μ_s is the word source main effect (fixed), $\mu_{m,s}$ is the model-source interaction effect (fixed), $\mu_{w(s)}$ is word effect nested within source (random), $\mu_{m,w(s)}$ is the model-word interaction nested within source (random), and ϵ is model lack of fit and error. The distinction of random versus fixed effects recognizes individual words and errors as variable and better characterized by distributions than fixed effects.

In the case of a model of a native language user, who has an abundant repertoire of inherited words and a lexical language model learned largely from inherited words, the language model is inherited and the effects model, where I signifies

[1] Sound power and volume measurement, and musical scales are log based (Decibels is a $log10$ based measure, and an octave or equivalent in other musical tradition corresponds to frequency doubling.

inherited, reduces to :

$$H(s, w|m = I) = \mu'_{s|m=I} + \mu'_{w(s)|m=I} + \epsilon.$$

When a native language user hears or speaks an inherited word, it should sound ordinary or common. This commonness corresponds to a lower valued cross-entropy for inherited words and in particular negative $\mu_{s=I|m=I}$ and positive $\mu_{s=B|m=I}$ fixed effects. Other terms remain in the effects model as random effects.

One can think of the distribution of calculated cross-entropies as having been generated by the linear mixed model, and composed of inherited and borrowed source distributions. To discriminate between inherited and borrowed words, one of checks the cross-entropy calculated by the inherited model against a critical value, per this boolean decision function:

$$s = borrowed \leftarrow H(s, w|m = I) > H_{crit}.$$

This decision function is not particularly powerful. Except for the fixed difference between inherited and borrowed sources, $\Delta = \mu_{s=I|m=I} - \mu_{s=B|m=I}$, variability due to random effects, especially word effects, remains. There is little power in this test and it's doubtful that this is a reasonable model of supposed native speaker discrimination. The variance of the cross-entropy measure under the linear mixed model is,

$$Var(H(s, w|m = I)) = \Delta^2_{\mu'_{s|m=I}} + \sigma^2_{\mu'_{w(s)|m=I}} + \sigma^2_{\epsilon},$$

where[2] $\Delta^2_{\mu'_{s|m=I}}$ captures fixed effects variability and $\sigma^2_{\mu'_{w(s)|m=I}}$ random effects variability. The fixed effect difference $\Delta_{\mu'_{s|m=I}}$ has to dominate random effects such as word variability, in order to provide a powerful decision function.

Inherited and borrowed word competing cross-entropies. We use multiple lexical models to characterize the problem of discriminating between inherited and other word sources. In particular we consider the simplified case of an inherited word model trained on inherited words and a borrowed word model trained on borrowed words. For previously unseen words, to classify a word, the cross-entropy is calculated with each model and the results compared.

For inherited words, inherited word model estimates should generally be lesser than borrowed word model estimates. Analogously for borrowed words, borrowed word model estimates should generally be lesser than inherited word model estimates. This is captured as the following decision function,

$$s = borrowed \leftarrow (H(s, w|m = I) - H(s, w|m = B)) > 0.$$

This decision function turns out to be much more effective, because the differencing of cross-entropies by word, removes the main random effect of words

[2] terms are taken to be independent given the model hierarchy.

from the predictive model. We see this in the following linear mixed model representation of the decision function:

$$H(w|m = I) - H(w|m = B) =$$
$$\mu_{m|m=I} - \mu_{m|m=B} + \mu_{m,s|m=I} - \mu_{m,s|m=B} + \mu_{m,w(s)|m=I} - \mu_{m,w(s)|m=B} + \delta.$$

Word source and word within source effects drop out, and all other effects, retain inherited versus borrowed differences. This decision function is potentially much more powerful. The variance of the decision function under the linear mixed model is,

$$Var(H(w|m = I) - H(w|m = B)) =$$
$$\Delta^2_{\mu_m} + \Delta^2_{\mu_{m,s}} + \sigma^2_{\mu_{m,w(s)|m=I}} + \sigma^2_{\mu_{m,w(s)|m=B}} + \sigma^2_{\delta}.$$

2.1.2 Direct classification - borrowing models

Our models to directly predict borrowings represent two corner-points in experiments in lexical borrowing detection: 1. Bag of sounds, a simple non-sequential model based on only the set of sounds, consider only the phonology of words, and 2. Neural network sequence models predicting borrowings directly without an intervening segmented IPA prediction nor word cross-entropy calculation. These borrowing models are still lexical language model based but with the output no longer "string prediction" (Bender and Koller, 2020), but rather direct borrowing prediction.

Bag of sounds. The methods section of Miller et al. (2020) provides a detail explanation of the bag of sounds classifier method:

> Since the word forms in our data are available as harmonized phonetic transcriptions, it is straightforward to represent each word form in a given language as a vector indicating the presence and absence of distinct sound segments. Since the order of these sound segments is not important, and neither is their frequency considered, this vector can be thought of as a simple bag of sounds, in which the sounds making up a given word form are represented as a set. The task of distinguishing borrowed from inherited words can then be pursued with the help of a support vector machine with a linear kernel (Hastie, Tibshirani, and Friedman, 2001; Cristianini and Shawe-Taylor, 2000). The support vector machine identifies the plane which optimally separates inherited from borrowed words based on the set of sound segments. The bag of sounds method does not consider the order or the frequency of elements in a given sound sequence, and we did not expect it to perform extraordinarily well in all languages in our sample. The advantage of the model is that it is simple and fast in application. It also provides a baseline for those cases where peculiar sounds provide enough information to identify a given borrowed word. (Miller et al., 2020)

Direct neural network. An advantage of neural network models is that they ought to determine an appropriate decision function directly without having to insert human knowledge of theoretical intervening variables. In this case, even though there are plausible arguments for cross-entropy measures representing human lexical processing in borrowing detection, this may not be a necessary step if all we want to do is predict borrowing.

The direct model also uses the transformer architecture §2.1.1 where the output of the transformer module is either averaged (GlobalAveragePooling) to a 1-D vector the same size as the attention vectors, or flattened (Flatten) to a 1-D vector the size of the number of sounds by the size of the attention vectors. Given the relatively small size of attention/hidden layer vectors (length = 32 as seen in Tab. 2.1) and a similar word size limit, flattening is quite practical here. This output is connected to a dense output layer which predicts borrowed versus inherited status directly, without intervening word predictions, cross-entropy calculations, or separate decision function.

2.1.3 Donor focused borrowing models

Detection of lexical borrowing includes not only whether words are borrowed, but also from which language donor are they borrowed. This additional capability to take into account language donor is relatively straightforward for both the competing cross-entropy and the direct borrowing models.

The methods section of Miller, Pariasca, and Beltran Castañon (2021) introduces multiple donor and donor focused borrowing models:

> We broadened the problem definition to include donor source of the borrowed words. Output includes indication of word donor instead of simply inherited versus borrowed, where $d = 0$ designates an inherited word, and $d \in [1, D]$ designate which is the borrowed word donor. A minimum donor wordlist size of 75 words was used to assure enough data to fit corresponding donor models; less than 75 word donor sources were combined into a remaining wordlist.
>
> For competing cross-entropy models this results in $D + 1$ individual cross-entropy models per language — one for inherited words and one for each word donor. In the decision procedure, all models compete for which has the cross-lowest entropy to select the word. For the direct model, only one model is created and it directly discriminates between donors. (Miller, Pariasca, and Beltran Castañon, 2021)

This task of selecting the lowest cross-entropy donor can be described by an equation, $d = arg_d min(CE(w|m = d))$, where d is the donor, w is the word being tested, and m is the model out of $D + 1$ competing models.

2.1.4 Data augmentation borrowing models

Given the paucity of data, i.e., ≈1,500 words per language with commonly four to seven IPA sound segments per word, we tried to improve neural network

model results by artificially augmenting training data by performing the following experiments:

1. Added donor wordlist without alteration, as though it were an additional source of known borrowed words, and
2. Machine translation of donor wordlist to the target language, with subsequent inclusion as though a list of known borrowed words.

The first simple experiment developed a prototype method to include additional donor training data. This opened up the way for the very complex, labor and compute intensive experiment requiring machine translation.

Added Spanish donor wordlist

Miller, Pariasca, and Beltran Castañon (2021) describe the experimental process:

> Lack of sufficient borrowed words for training is a detractor so we enhanced our data methods to permit an additional donor wordlist source. This was a limited experiment where we include a Spanish wordlist from IDS (Key and Comrie, 2015), transcribe it to IPA, and combine it with Spanish donor borrowed words for target languages where Spanish is the primary donor language. Latin American languages in WOLD that have Spanish as the primary language donor are: Imbabura Quechua, Mapudungun, Otomi, Q'eqchi', Wichí, Yaqui, and Zinacantán Tzotzil. We apply our enhanced data methods with both competing cross-entropy and direct models to these seven languages. (Miller, Pariasca, and Beltran Castañon, 2021)

We shall see (§2.2.7) that simply adding a Spanish wordlist impacted precision and recall results, but with little overall improvement in borrowing detection.

Translated Spanish to target language wordlist

To improve performance, we augment borrowed word training data by simulating what borrowed words might look like in the word table. The expectation is that if we can simulate borrowed words well enough for the target language, then this would improve training of the borrowed word model.

The configuration of the competing entropies approach used in this experiment is shown in Tab. 2.2. The number of parameters varies with the vocabulary size, between 25 and 60 IPA segments in this study, and was on the order of 10,000 parameters each over target languages.

To simulate borrowed words we performed word translation from a Spanish donor wordlist to separate target languages (Left side of Fig. 2.3) via neural translators trained on a Trax Transformer model. Trax is "an end-to-end library for deep learning ... actively used and maintained in the Google Brain team" (Team, 2021). We configured training parameters for a single layer Trax Transfomer, "an encoder-decoder that performs tokenized string-to-string transduction", as shown in Tab. 2.3. Even though limited to a single layer, the number

Table 2.2: Data augmentation competing entropies configuration.

Parameter	Value
dimension	32
learning rate	0.0055
dropout	0.1
learning schedule	Transformer
number of epochs	100/150
number of parameters	$\approx$ 10,000

of training parameters was on the order of 30,000 over the full encoder-decoder model. While tiny for a typical text translation model, this seems overwhelming for the application of translating $\approx 1,400$ Spanish words to the different target languages.

Table 2.3: Translation model configuration.

Parameter	Value
dimension	40
learning rate	0.001
dropout	0.3
optimizer	Adafactor
learning schedule	multifactor
- warmup steps	200
- decay steps	100
number of steps	2000
loss function	cross-entropy
number of parameters	$\approx$ 30,000

The work flow for training and testing each target language in this experiment is shown in Fig. 2.3 and described as follows:

1. Separate the target language word table into train and test data.
2. Further divide train and test data into inherited and borrowed words.
3. For the borrowed words, identify the corresponding loaned Spanish language donor words.
4. Construct parallel train and test translation datasets consisting of Spanish donor words and corresponding target words.
5. Train Spanish donor to target language translator, and validate on test data.
6. Translate full Spanish donor wordlist to target language simulated borrowed words.
7. Train competing entropies borrowed and inherited words models on inherited words train data, and on borrowed words train data augmented by simulated borrowed words.

8. Validate borrowed word detection on test data.

Maintaining clear separation of train and test data across translation and borrowing detection, we selected optimal translation models based on internal training measures (15% sample of train) of maximum accuracy or minimum cross-entropy, or termination at 2000 training steps.

We used a train (90%) versus test (10%) split for each language table, where the split is preserved over the competing entropies approach and the simulated borrowed word translation. Because this is only a one-fold split, results from this experiment are not directly comparable with 10-fold cross validation studies of (Miller et al., 2020; Miller, Pariasca, and Beltran Castañon, 2021). There was, however, little cost to replicate training and test on the same partitioned train and test dataset, so each trial was replicated 10 times and the mean test results reported.

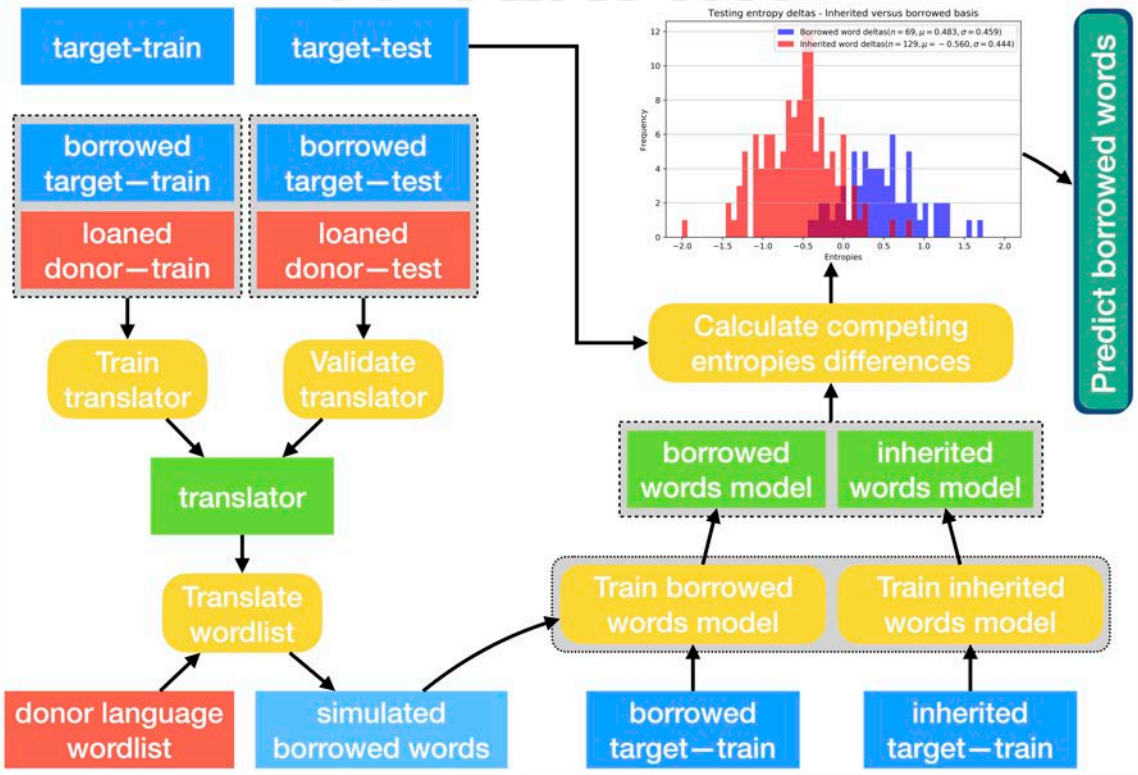


FIGURE 2.3: Competing entropies borrowing detection with data augmentation by simulated borrowed words.

2.1.5 Assessing detection performance

We have already mentioned precision and recall, and alluded to an overall measure of performance in detecting borrowings. Miller et al. (2020) defines these measures:

> We assess detection performance using *precision*, *recall*, and *harmonic mean* (*F1 score*), as well as *accuracy* measures based on frequency counts of borrowing detection by true borrowing status as defined

in Tab. 2.4. Following (Manning and Schütze, 2001), *precision* is the proportion of true positive borrowings out of all detected positives,

$$precision = tp/(tp + fp),$$

recall is the proportion of true positive borrowings out of all borrowings,

$$recall = tp/(tp + fn),$$

F1 score is the harmonic mean of precision and recall, and

$$F1 = (2 * precision * recall)/(precision + recall),$$

accuracy is the proportion of all detections that are correct,

$$accuracy = (tp + tn)/(tp + fp + fn + tn).$$

We consider F1, since it combines both precision and recall, as the primary measure. Accuracy does not specifically focus on borrowing detection and is of secondary importance. (Miller et al., 2020)

While *Borrowed* and *Inherited* are the category of interest and default category here, we could readily change this to focus on borrowings from a particular donor language, as we do in §3.1.2. In some experiments, we also measure execution time.

Table 2.4: Frequency counts of borrowing detection by true borrowing status.

Borrowing Detection	True borrowing status **Borrowed**	**Inherited**
Positive	tp=true positive	fp=false positive
Negative	fn=false negative	tn=true negative

From Miller et al. (2020).

2.2 Experiments and results

We use the World Loan Database (WOLD) (Tresoldi, Forkel, and Morozova, 2019) multilingual collection of wordlists as our primary data source for experiments in this chapter. The first several experiments have been previously documented in (Miller et al., 2020). We replicate these experiments and add experiments on inherited only lexical models, which we had explored before discovering the competing cross-entropies approach. We also report on competing cross-entropies and direct approaches based on our light-weight transformer lexical model previously documented in (Miller, Pariasca, and Beltran Castañon, 2021). Finally we report on our most complex approach where we augment training data based on translation of donor to target language wordlists.

2.2.1 Artificially seeded borrowings

The experiments and results section of Miller et al. (2020) reports the results of our artificially seeded borrowings experiments:

> To simulate a situation in which foreign words have recently entered a language without being modified by borrowed word nativization processes, we designed an experiment in which the wordlists in our base datasets were artificially mixed with words from another wordlist which was not part of the original WOLD collection. The idea to use "artificially seeded" borrowings instead of borrowings attested in actual language was originally proposed for evaluating methods for lateral gene transfer detection in biology (Dessimoz, Margadant, and Gonnet, 2008), and later tested on linguistic data in order to assess the power of phylogenetic methods for borrowing detection across multiple languages (List et al., 2014a). The advantage of this procedure is that it creates simulated data without requiring the efforts of detailed simulation experiments.
>
> Artificial borrowings were seeded into a wordlist in three steps. We first removed all borrowed words from the wordlist to guarantee that no recent borrowings from other languages could influence the results. We then added inherited words from the additional German list (see §1.2), which we created for testing purposes. Here, we tested three different proportions of borrowed words, 5%, 10%, and 20%, in order to allow to compare different degrees of contact. In a final step, we then split the resulting wordlist into a training and a test set (reserving 80% of the data for training and 20% for testing) and ran the three methods for monolingual borrowing detection, bag of sounds, Markov model, and neural network. (Miller et al., 2020)

Only the competing cross-entropies approach is included in (Miller et al., 2020). To this we add results for inherited only, Markov chain and neural network, lexical models. In the competing cross-entropies approach, the difference between borrowed and inherited cross-entropies is compared to zero to determine borrowed status. With the inherited only approach, the estimated cross-entropy is compared to a critical value to determine borrowed status. For this critical value, we use 20^{th} percentile of the training distribution of borrowed words which if the training and test distributions are similar, guarantees that about 80% of borrowings will be classified as such.

Experimental results are shown in Tab. 2.5 and Figs. 2.4, 2.5, and 2.6. Detail results by individual language and method are reported in appendix Tabs. A.1, A.2, and A.3. Bag of sounds and competing entropies models perform well at 20% and 10% borrowings achieving F1 scores or more than 0.90. At 5% borrowings, the competing entropies neural network method attains an F1 score of 0.89, the competing entropies Markov method degrades substantially with the poorest F1 score of 0.78, and bag of sounds and inherited only models all show high precision with F1 scores in the 80% range. High precision for the bag of sounds

TABLE 2.5: Borrowing detection results for artificially seeded borrowings.

Method	Rate	Precision	Recall	F1
Bag of Sounds	5	0.99	0.80	0.87
Inherited Markov	5	0.90	0.83	0.85
Competing Markovs	5	0.70	0.95	0.78
Inherited Neural	5	0.92	0.79	0.84
Competing Neurals	5	0.83	0.97	0.89
Bag of Sounds	10	0.99	0.87	0.92
Inherited Markov	10	0.91	0.82	0.86
Competing Markovs	10	0.87	0.97	0.91
Inherited Neural	10	0.93	0.79	0.84
Competing Neurals	10	0.91	0.98	0.94
Bag of Sounds	20	0.99	0.91	0.94
Inherited Markov	20	0.95	0.85	0.89
Competing Markovs	20	0.95	0.97	0.96
Inherited Neural	20	0.98	0.79	0.87
Competing Neurals	20	0.96	0.98	0.97

Results averaged over all languages for each method and borrowing rate. Updated from Miller et al. (2020).

model at 5% borrowings is likely due to the detection of sounds unique to the German artificial borrowings. The poorer competing entropies Markov models method at 5% borrowings is likely due to the poor estimation of the borrowed word model with so little data, given that the inherited only Markov model has not degraded.

Note that with commonly 20 to 50 distinct sound segments per language, and given the phonotactic restrictions on sound combinations, there are $\approx 1,750$ to $\approx 12,500$ non-zero conditional probability estimates for the Markov chain language models, depending on the number of distinct sounds and phonotactic restrictions. In contrast the bag of sounds model, where we don't see such model degradation, representing only the set of unique sounds per language, has ≈ 20 to ≈ 50 parameter estimates.

2.2.2 Borrowing detection on real language data (WOLD)

We experimented with more challenging real world data from WOLD wordlists, as reported by Miller et al. (2020):

> Our experiment on artificially seeded borrowings simulated an ideal situation of language contact in which new words were recently introduced into a given language without being adjusted (adapted) to the recipient language's target phonology. While this experiment provided high scores in our evaluation experiment, the experiment does not allow us to estimate how well the three borrowing detection methods will perform when being exposed to "real" data. For this

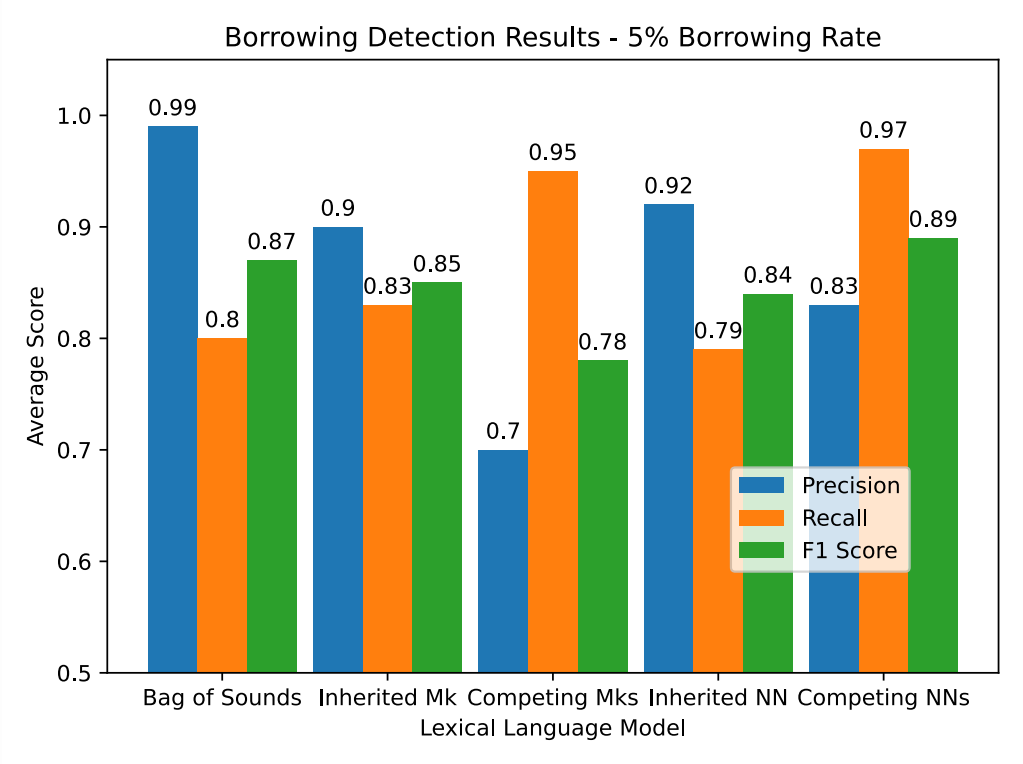


FIGURE 2.4: Borrowing detection results for 5% artificially seeded borrowings. From Miller et al. (2020).

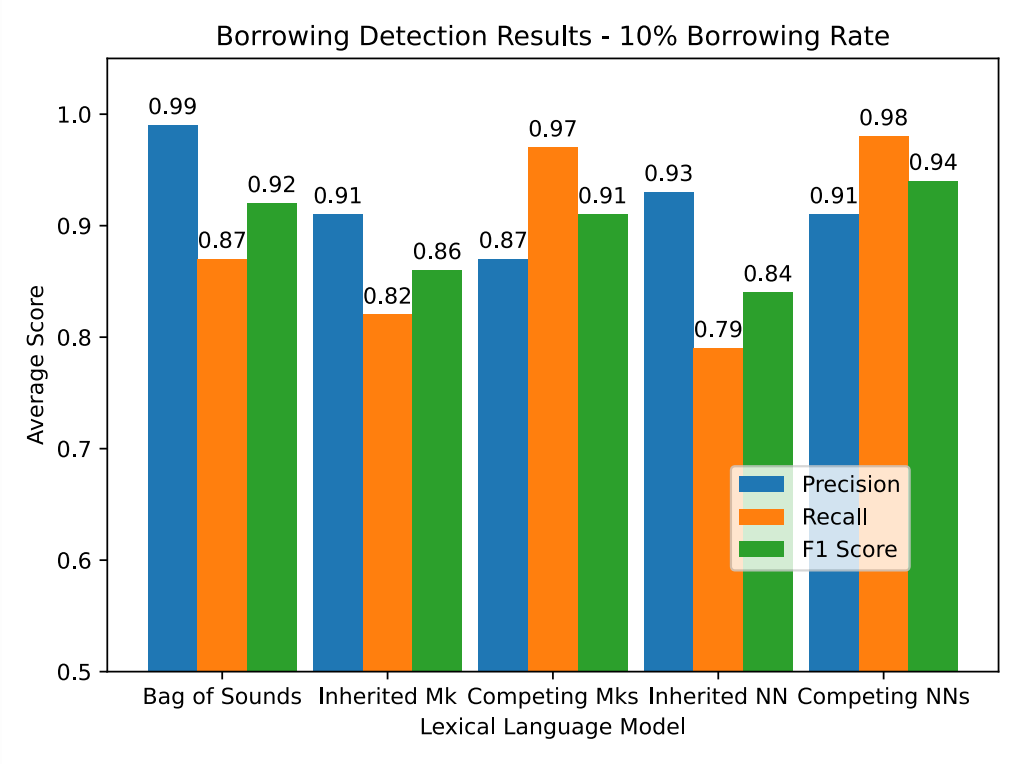


FIGURE 2.5: Borrowing detection results for 10% artificially seeded borrowings. From Miller et al. (2020).

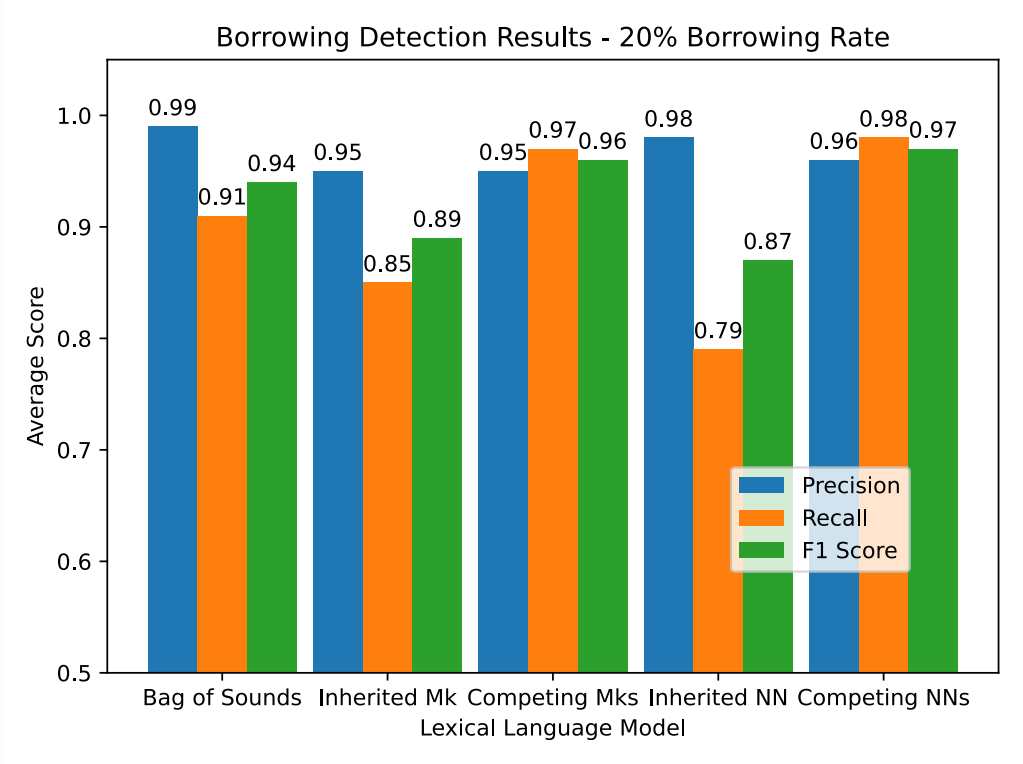


FIGURE 2.6: Borrowing detection results for 20% artificially seeded borrowings. From Miller et al. (2020).

> reason, we designed a second experiment on the WOLD data in their original form. Given that the wordlists are quite small, while specifically Markov Model and Neural Network language models tend to require larger amounts of data, we used cross validation techniques, in which the data are repeatedly partitioned into training and test data and evaluation results are measured for each trial and later summarized. We employed *ten-fold cross validation* for this experiment, where each word list was partitioned into 10 parts, and over 10 successive trials, one part was successively designated the test set while the remaining nine parts were designated the training set. This resulted in 10 separate estimates of borrowing detection performance, with each word appearing once in test sets and nine times in training sets. (Miller et al., 2020)

Tab. 2.6 shows the averages, standard deviations across languages, and pooled standard deviations across partitions within languages for cross-validation results (*precision*, *recall*, *F1 score*, *accuracy*) for each of our methods. As above, we also add the inherited only Markov chain and neural network lexical methods to that which was reported in (Miller et al., 2020). Fig 2.7 graphically summarizes the averaged results. Detail results by individual language and method are reported in appendix Tabs. A.4 and A.5, and Tabs. A.6 and A.7 for inherited only methods.

We observe from the table and figure:

TABLE 2.6: Borrowing detection results of the cross validation experiment.

Method	Statistic	Pecision	Recall	F1	Accuracy
Bag of Sounds	Mean	0.592	0.289	0.353	0.844
	Language SD	0.281	0.247	0.263	0.081
	Pooled SD	0.169	0.066	0.074	0.027
Inherited Markov	Mean	0.330	0.796	0.440	0.617
	Language SD	0.171	0.030	0.168	0.134
	Pooled SD	0.048	0.093	0.054	0.039
Competing Markovs	Mean	0.527	0.676	0.583	0.830
	Language SD	0.181	0.152	0.170	0.060
	Pooled SD	0.076	0.088	0.066	0.029
Inherited Neural	Mean	0.315	0.791	0.426	0.593
	Language SD	0.164	0.056	0.166	0.138
	Pooled SD	0.052	0.119	0.058	0.041
Competing Neurals	Mean	0.549	0.701	0.606	0.844
	Language SD	0.191	0.161	0.180	0.062
	Pooled SD	0.084	0.101	0.076	0.032

Mean and standard deviation over languages, and pooled standard deviation within language for each method over all languages. Updated from Miller et al. (2020).

- Performance is *poorer* versus that for the easier task of artificially seeded borrowings.
- Competing cross-entropies Markov chain and neural network methods substantially outperform inherited cross-entropy and the bag of sounds methods.
- Inherited cross-entropy methods both achieve average recall of ≈ 0.8 with their decision function's critical value based on the 20^{th} percentile of the training distribution of borrowed words, just as for the artificially seeded cases.
- Inherited cross-entropy methods outperform the phonology only Bag of Sounds method.
- Standard deviations across languages, ≈ 0.17 for cross-entropy methods and 0.26 for bag of sounds, show substantial variability in performance by language.

An examination of detail results for a few individual languages is reported by Miller et al. (2020):

> When examining the individual results achieved by each method for each individual language in our sample (Tab. A.4), one can find substantial variation in the results, ranging from results which one may consider as satisfying (such as the performance of the Neural Network on Zinacantán Tzotzil with an F1 score of 0.81) up to extremely

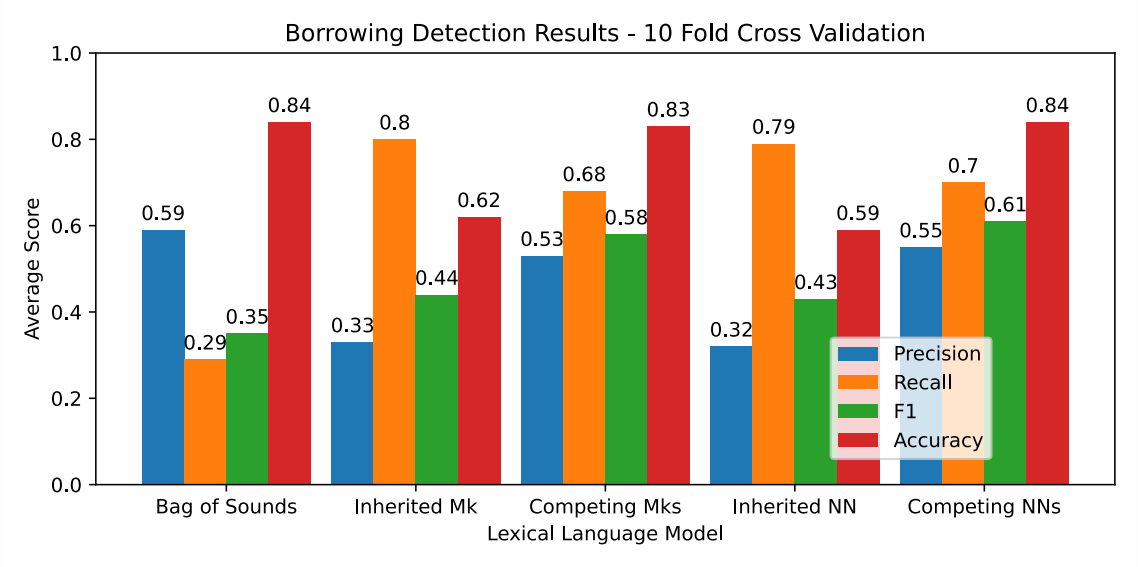


FIGURE 2.7: Results of the cross validation experiment.
Averaged for each method over all languages in our sample.
Adapted and updated from Miller et al. (2020).

> bad results (such as the performance of all methods on Mandarin Chinese, with F1 scores below 0.02). The reasons for the underwhelming results on Mandarin Chinese are twofold. On the one hand, the language rarely borrows words directly, but rather resorts to *loan translation*, by which new concepts are rendered with the help of the lexical material in the target language. As a result, Mandarin has the lowest amount of direct borrowings in our sample. On the other hand, Mandarin Chinese (as well as all Chinese dialects and many languages from Southeast Asia) has an extremely restricted syllable structure that makes it impossible to render most foreign words truthfully (Norman, 1988). As a result, words are usually directly adjusted to Chinese phonotactics when being borrowed and also written with existing Chinese characters, which again further masks their foreign origin (Sun, 2006). However, this very specific situation also makes it also difficult if not impossible for most Mandarin Chinese speakers to identify borrowings when considering phonotactic criteria alone. (Miller et al., 2020)

If we can figure out how to improve the results for poorer performing languages, maybe excluding the Chinese languages, we could have a more useful borrowing detection method.

2.2.3 Factors that influence borrowing detection

We then looked at factors by target languages that may have influenced borrowing detection, as reported by Miller et al. (2020) in their experiments and results section:

> Given that the performance of our supervised borrowing detection methods varied substantially, ranging from poor performance with F1 scores below 0.5, average performance with F1 scores between 0.5 and 0.8, and acceptable performance with F1 scores above 0.8, we performed analyses to assess to which degree certain factors might influence the borrowing detection methods.
>
> In concrete, we computed specific characteristics of each language variety in our sample and then checked to which degree these characteristics correlated with the test performance. As characteristics, we chose the proportion of borrowed words in a given language wordlist (since statistical and machine learning methods perform better with sufficient representation), and the proportions of unique sounds in borrowed words and in inherited words, as potential contributors to prediction performance. A higher proportion of borrowed words corresponded moderately to a lower proportion of unique sounds in inherited words, otherwise characteristics were independent. (Miller et al., 2020)

Characteristics by individual language that may impact borrowing are reported in Tab. A.8.

Statistical analyses (here correlational and matrix plots), were performed with JMP® Statistical Software (*JMP®, Version 17.0.0* 2022). A previous analysis for (Miller et al., 2020) used Minitab (Minitab, 2020). Correlation results, based on wordlists from the WOLD database, are reported in Tab. 2.7 with corresponding detailed plots in Figs. 2.8, 2.9, and 2.10.

Table 2.7: Correlations between phonological characteristics and performance of borrowing detection methods.

Proportion	Precision	Recall	F1
Bag of Sounds			
Borrowed words	**0.317**	**0.558**	**0.500**
Borrowed sounds	**0.370**	0.225	0.230
Inherited sounds	0.069	-0.059	-0.039
Markov Chain			
Borrowed words	**0.716**	**0.395**	**0.626**
Borrowed sounds	0.267	**0.343**	0.292
Inherited sounds	-0.283	-0.077	-0.228
Neural Network			
Borrowed words	**0.668**	**0.402**	**0.595**
Borrowed sounds	0.234	**0.359**	0.268
Inherited sounds	-0.163	-0.068	-0.132

Correlations with $|r| \geq 0.30$ are significant at $p < 0.05$.
Updated from Miller et al. (2020).

There is a moderate to strong positive correlation between the proportion of borrowed words and the borrowing detection performance for all methods. Opposed to this, there is little relationship between the proportion of unique inherited sounds and detection performance for any method. The proportion of unique borrowed sounds is more strongly related to precision for the bag of sounds method and to recall for competing cross-entropies methods.

The bag of sounds method has several languages where detection performance is zero or close to zero. Review of the borrowed sounds by recall and F1 score measures in Fig 2.8 shows most such languages have zero or close to zero borrowed sounds. This is consistent with the bag of sounds method which only considers the set of phonological symbols and so has little information to detect borrowed words without unique borrowed sounds. Both competing cross-entropy methods show non-linearities in their the borrowed words and borrowed sounds relationships with detection performance. For a very low proportion of borrowed words, detection performance is always lower than the expected linear relationship, and for an approximate zero proportion of unique borrowed sounds, evaluation scores may be almost any value in the range 0.0 to 1.0.

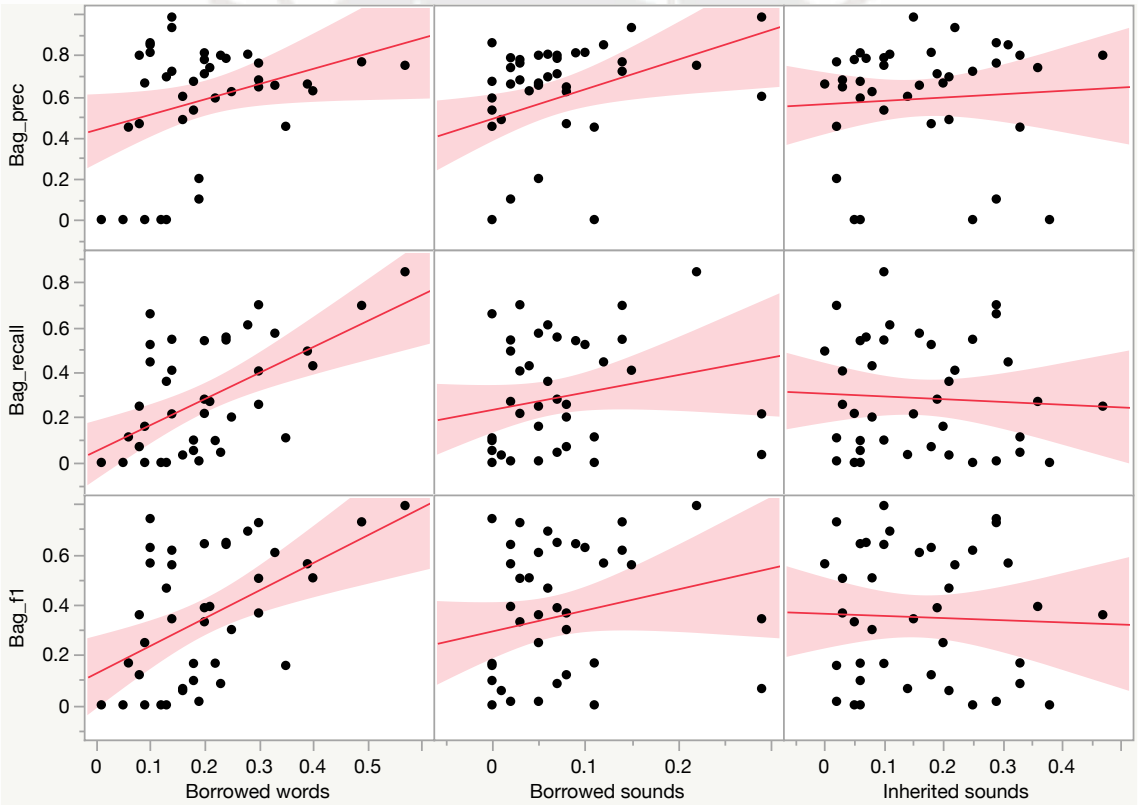


FIGURE 2.8: Determining characteristics that influence the performance of the bag of sounds.
Updated from Miller et al. (2020).

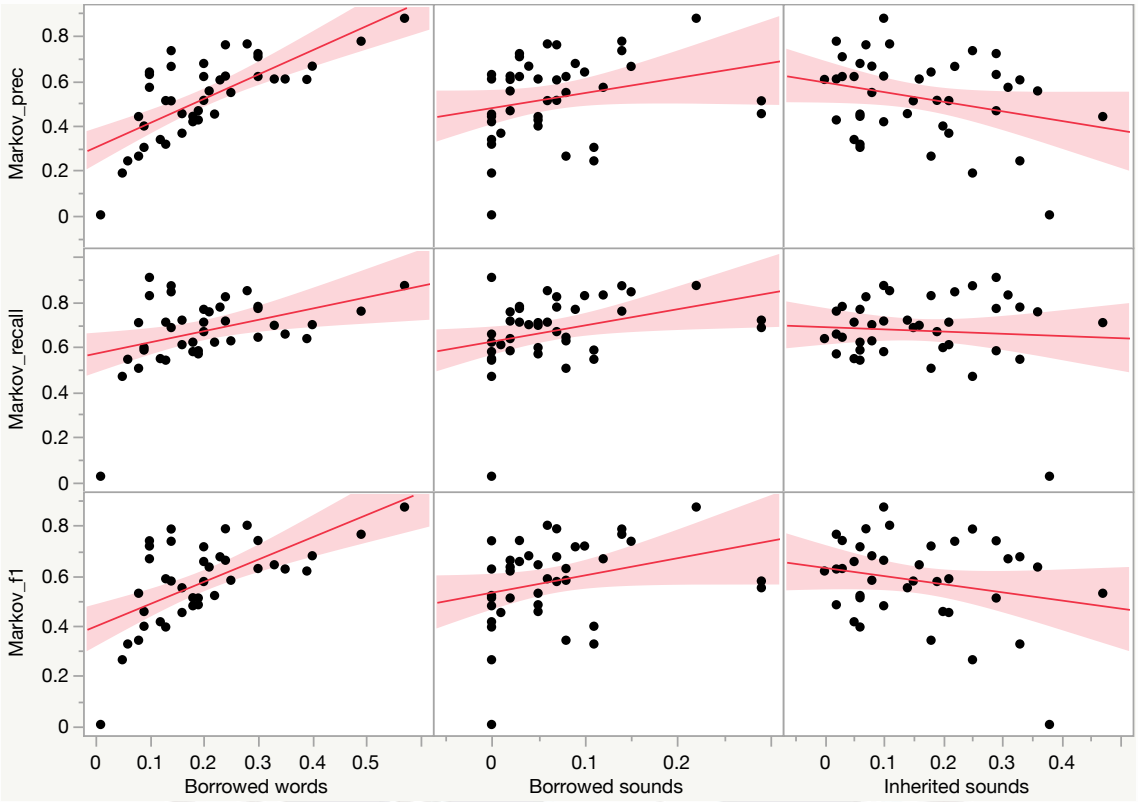


FIGURE 2.9: Determining characteristics that influence the performance of the Markov chain cross-entropies.
Updated from Miller et al. (2020).

2.2.4 Detecting borrowings when there is a dominant donor

Another factor which influences borrowing detection, is the intensity of the language contact situation, as quantified by the proportion of borrowed words from a single language donor. Quantity of lexical borrowing and primary language donor characteristics are reported by individual language in Tab. A.8. We look again at the relationship between proportion of borrowed words, a proxy for intensity of language contact, as reported by Miller et al. (2020):

> Testing our lexical language models on the WOLD data in their entirety could be considered as unfair to the methods, given that we know well that monolingual evidence for borrowing in phonotactics may get lost easily and that the WOLD database was never restricted to recent borrowings alone. Another problem of the data is that the distinction between inherited words on the one hand and borrowings on the other hand is as well a simplifying assumption, since we know that in intensive contact situations borrowings come from a specific donor language. As a result, it seems to be justified to test the three methods for monolingual borrowing detection with the help of more specific experiments in which the task consists in the detection of borrowings when there is a single or dominant language donor, as in intensive contact situations, versus the case when no language donor

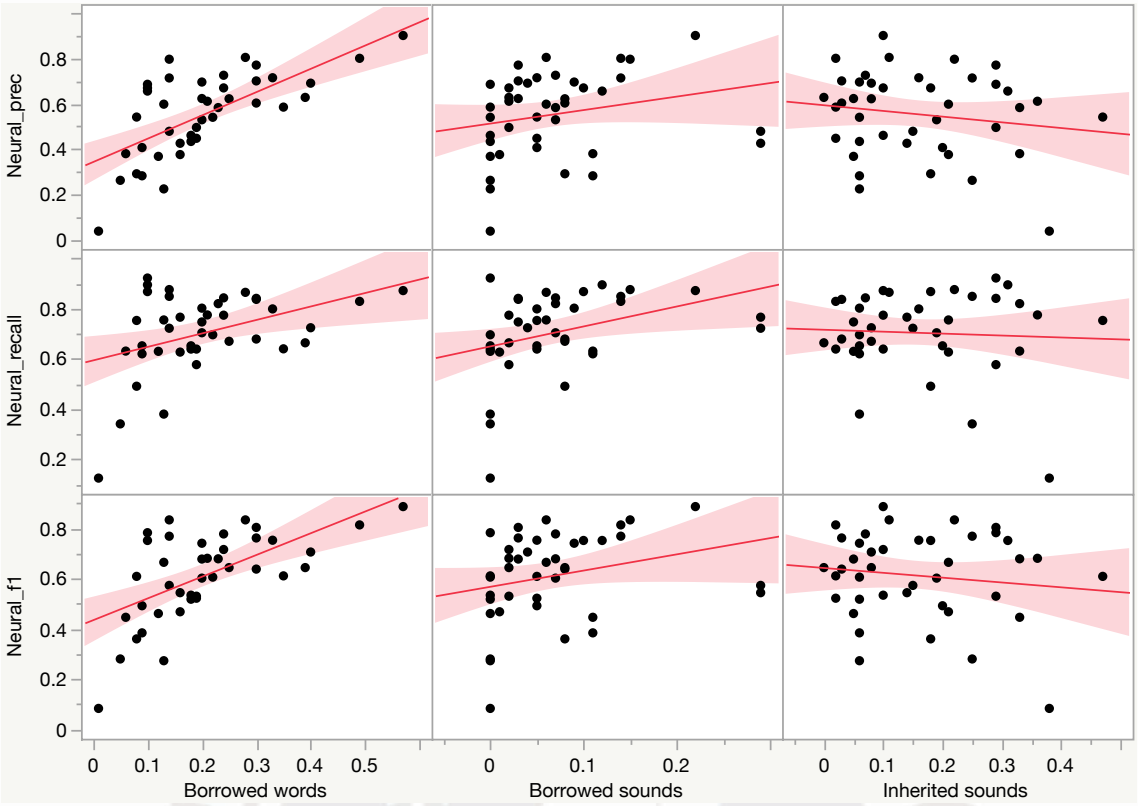


FIGURE 2.10: Determining characteristics that influence the performance of the neural network cross-entropies.
Updated from Miller et al. (2020).

dominates.

To test whether our methods show an improved performance when there is a dominant language donor as opposed to detecting borrowed words *per se*, we first created two subsets of the WOLD database, one containing languages with 300 and more borrowed words (17 language varieties), and one containing languages with 100 and more borrowed words (37 language varieties). We then searched for "dominant donor languages" in all wordlists in each sample, with dominant donor languages being defined as those donor languages (as identified in the WOLD database) that would account for two-thirds of all borrowings identified for a given language variety. For our sample of language varieties with 300 and more borrowings, this yielded a partition of the data into 8 language varieties for which a dominant donor could be identified and 9 for which none could be found. For the sample of language varieties with 100 and more borrowings, the partition yielded 20 language varieties with a dominant donor and 17 without. We were able to apply results of the 10-fold cross validation study for these two subsets of the data, which we had previously applied to all language varieties in the WOLD database. In order to test whether the observed differences between dominant donor and no

dominant donor categories were significantly different, we also performed randomization resampling tests of 5,000 iterations each, using Student's independent t statistic with unequal variances as our test statistic. We report p-values from the empirical distribution of t statistics calculated under the hypothesis of no difference due to dominant donor, i.e., dominant and no dominant categories are exchangeable. (Miller et al., 2020)

TABLE 2.8: Dominant donor and quantity of borrowed words; effect shown in 10-fold cross validation results.

Method	Dominant	Precision	p<	Recall	p<	F1	p<
$\geqq$ 300 **Borrowed words**							
Bag of Sounds	Yes (8)	0.716	NS	0.532	.03	0.586	.04
	No (9)	0.660		0.312		0.393	
Markov Chain	Yes	0.722	.003	0.771	.002	0.743	.003
	No	0.594		0.670		0.626	
Neural Network	Yes	0.742	.006	0.818	.001	0.775	.002
	No	0.622		0.707		0.659	
$\geqq$ 100 **Borrowed words**							
Bag of Sounds	Yes (20)	0.735	.002	0.413	.003	0.486	.004
	No (17)	0.504		0.192		0.253	
Markov Chain	Yes	0.612	.02	0.764	.001	0.671	.003
	No	0.512		0.639		0.561	
Neural Network	Yes	0.640	.02	0.800	.001	0.702	.003
	No	0.526		0.661		0.578	

Numbers in parentheses give the frequency of languages included in each group of $\geqq$ 300 or $\geqq$ 100 borrowed words and dominant language 'Yes' or 'No'. Updated from Miller et al. (2020).

As shown in Tab. 2.8, the performance of borrowing detection methods improves when borrowings come from a dominant language donor. In all comparisons, the dominant donor case has better detection performance than the non-dominant case. Performance also improves, as noted previously, with more borrowed words. While the bag of sounds method shows a strong increase in performance when most borrowings come from a single donor language for the $\geqq$ 100 partition, it is still not competitive with the cross-entropy based methods.

2.2.5 Why competing cross-entropies works

We previously provided a linear model, somewhat theoretical explanation, of why competing cross-entropies works well versus an inherited only cross-entropy method §2.1.1. Here we illustrate this with an anecdotal, graphical, and data-based explanation of why the competing cross-entropies methods work better from our explanation reported in (Miller et al., 2020):

> The Markov model and the neural network methods estimate word cross-entropy on a per sound basis given the inherited or borrowed words on which they are trained. Models trained on inherited words

should estimate lower cross-entropies for inherited words, and models trained on borrowed words should estimate lower cross-entropies for borrowed words. However, since words are borrowed over time and potentially also from various donor languages, using a single language model for borrowed words is not always optimal.

Our decision procedure for the Markov model and the neural network methods requires the comparison of competing cross-entropies for a given word, the cross-entropy of the lexical language model derived from inherited words and the cross-entropy of the lexical language model derived from borrowed words. If the difference between the cross-entropies is greater than zero, we designate the word as borrowed, and if it is smaller than or equal to zero, we designate the word as inherited.

In order to investigate the *discriminative force* of this procedure, it is useful to compare cross-entropy difference distributions of inherited and borrowed words for a given language variety. The distributions [of cross-entropy differences] for training and test data from the English wordlist in the WOLD database are shown in Figs. 2.11 and 2.12. While there is a certain overlap between cross-entropy difference distributions for inherited and borrowed words, the problem of discriminating between them based on cross-entropy differences seems tractable, and we can assume that improvements in cross-entropy estimation would have an immediate benefit on prediction. (Miller et al., 2020)

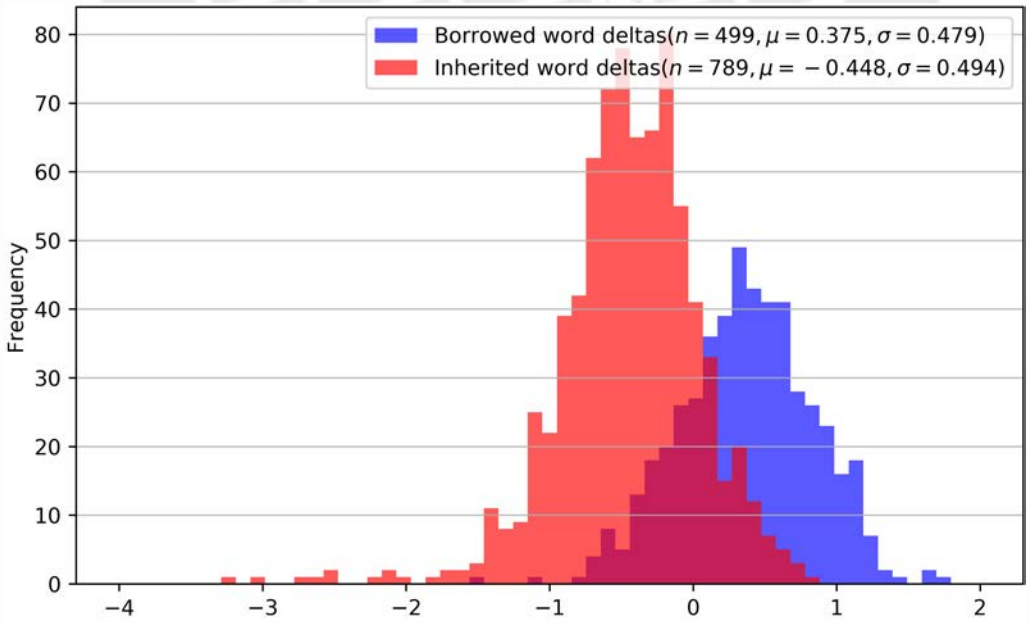


FIGURE 2.11: Distribution of training (85%) cross-entropy differences for English – Neural Network method.
From Miller et al. (2020).

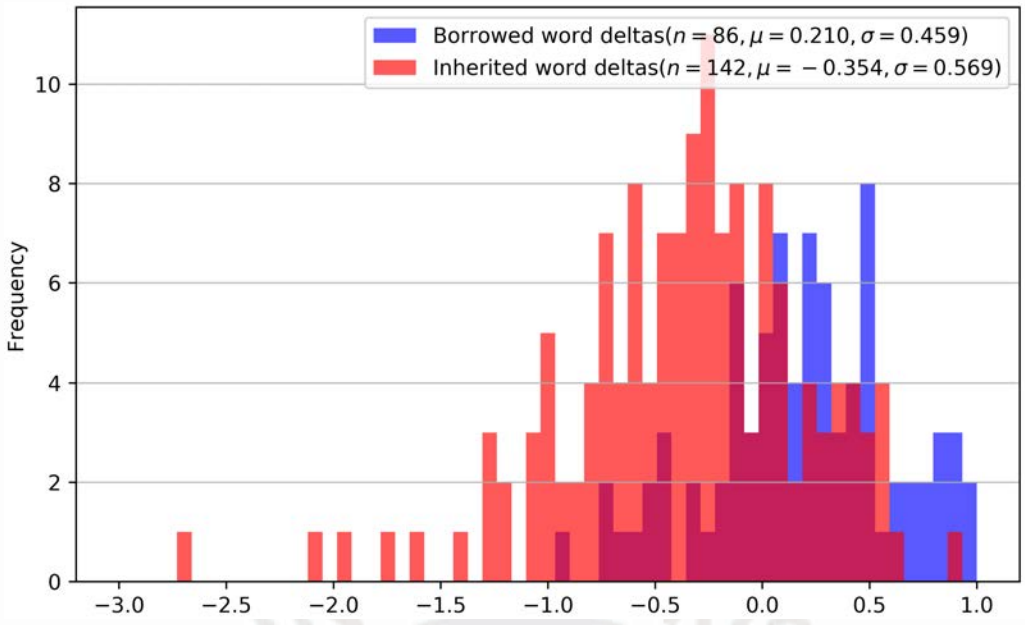


FIGURE 2.12: Distribution of testing (15%) cross-entropy differences for English – Neural Network method.
From Miller et al. (2020).

We noted in (Miller et al., 2020) that both the Markov chain and the neural network performed considerably well on Imbabura Quechua, a Quechua language spoken in Ecuador. With an F1 score above 0.8, it is not surprising to find a good separation between the cross-entropy difference distributions for inherited versus borrowed words, as shown in Figs. 2.13 and 2.14.

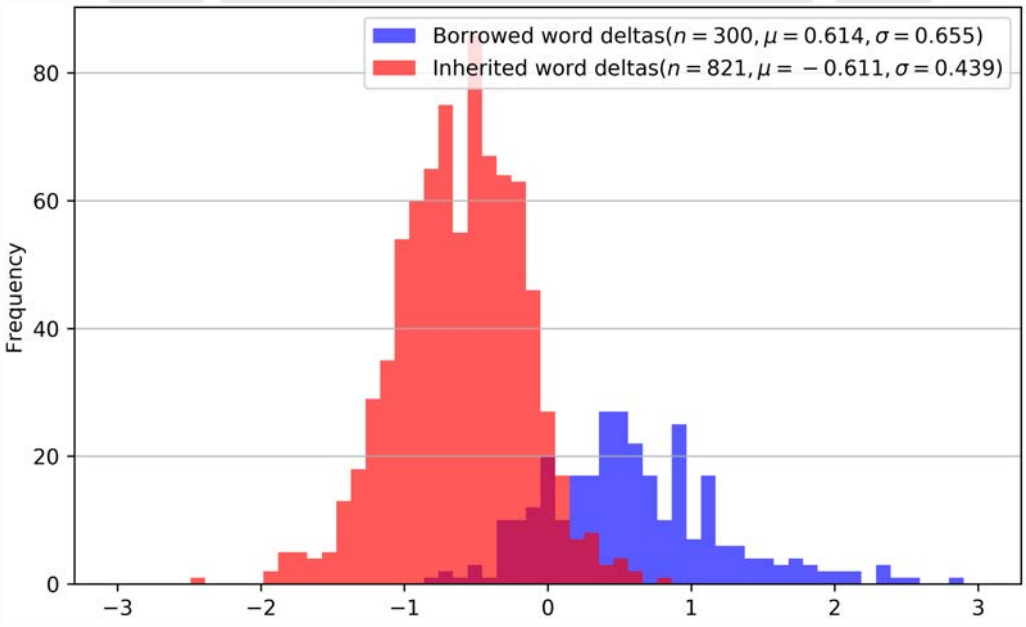


FIGURE 2.13: Distribution of training (85%) cross-entropy differences for Imbabura Quechua – Neural Network method.
From Miller et al. (2020).

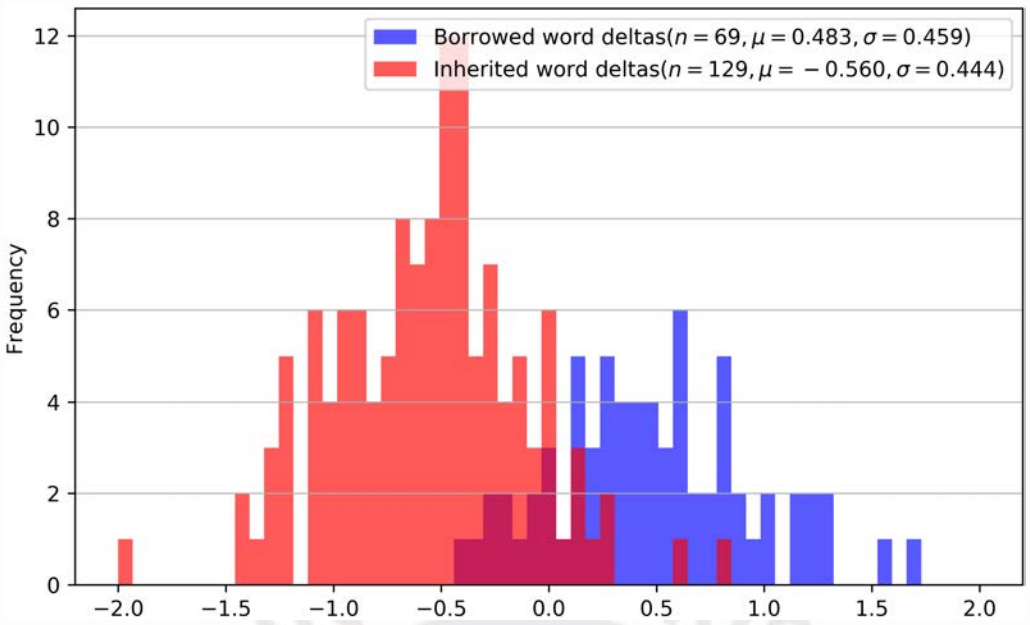


FIGURE 2.14: Distribution of testing (15%) cross-entropy differences for Imbabura Quechua – Neural Network method.
From Miller et al. (2020).

Likewise there are examples of poor performance, where we noted (Miller et al., 2020) that neither method performed very well on Oroqen, a Northern Tungusic language spoken in the Mongolian region of the People's Republic of China, with F1 scores below 0.36. As can be seen in Figs. 2.15 and 2.16 the cross-entropy difference distributions for inherited and borrowed words are poorly separated.

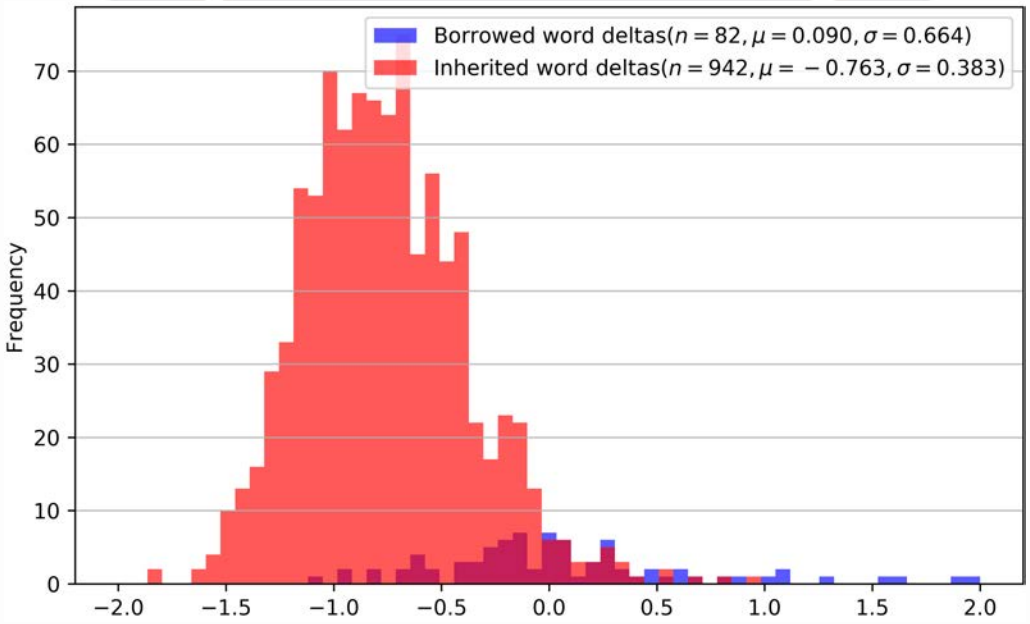


FIGURE 2.15: Distribution of training (85%) cross-entropy differences for Oroqen – Neural Network method.
From Miller et al. (2020).

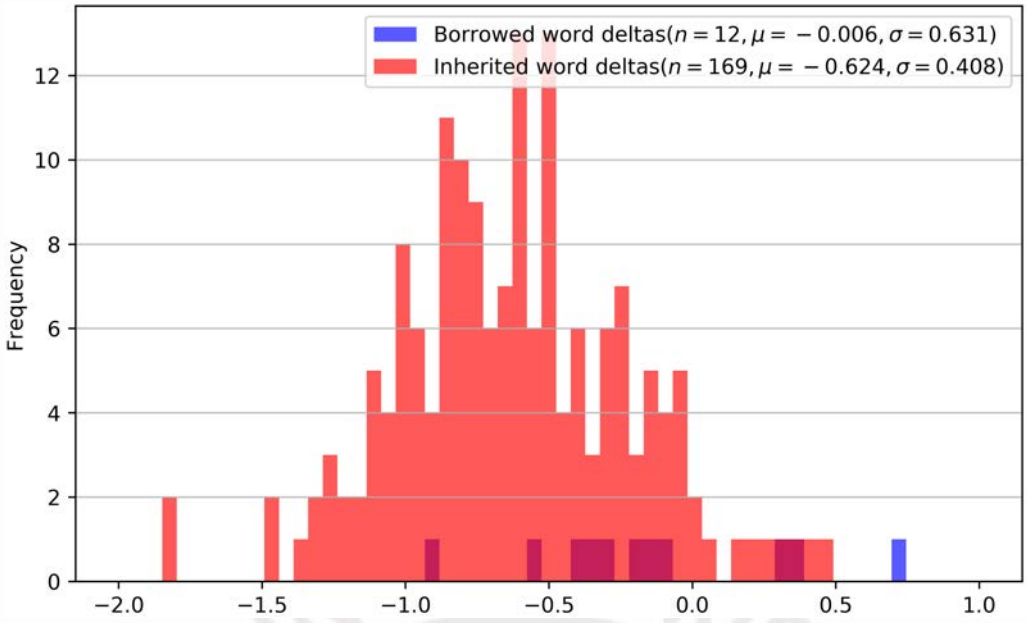


FIGURE 2.16: Distribution of testing (15%) cross-entropy differences for Oroqen – Neural Network method.
From Miller et al. (2020).

In the results section by Miller, Pariasca, and Beltran Castañon (2021), We elaborated on the usefulness of such graphical explanations:

> This strong relationship between the distribution of entropy differences and borrowing detection, indicates a tactic for improving monolingual lexical borrowing detection – increase the separation of difference distributions for inherited versus borrowed words. An examination of our sample cases reveals: 1. English and Imbabura Quechua, even though there were substantial borrowings, have reduced separation between inherited and borrowed word difference distributions for testing, resulting in reduced discriminative power, and 2. Oroqen, with few borrowings, has almost no separation between inherited and borrowed word distributions for testing, resulting in little discriminative power. Identification of problems permits trying to solve them, such as through improved training of neural networks, or by obtaining more borrowings, real or simulated, for training. (Miller, Pariasca, and Beltran Castañon, 2021)

The difference between train and test results is most likely due to insufficient training data given the large number of parameters being estimated by both Markov chain and recurrent neural network models. Random selection of train and test datasets makes non-representative sampling unlikely. Regulation in the neural network training and smoothing in the Markov chain training, was not sufficient by itself to eliminate test performance degradation.

Reasons for lack of discrimination between inherited and borrowed words for the Oroqen, English, Imbabura Quechua and similar cases even on the training data are: 1. in some cases the phonology and phonotactics between donor

and borrower are just not that different, so that similar words could easily be generated by either language, and 2. phonological and phonotactic adaption of different or strange sounding words effectively removes the distinct sound of borrowed words over time. Both reasons make lexical borrowing detection by language word models alone a difficult task.

2.2.6 Enhanced neural network experiments

While the neural competing cross-entropies model was little more than 2 percentage points better in F1 score than the Markov chain competing cross-entropies model, it offered an abundance of opportunities for experimentation and improvement. We performed the following experiments:

1. Develop a lightweight transformer substitute for the current recurrent neural network models,
2. Discriminate between language donor sources for borrowed words, and
3. Replace the competing entropies approach with a direct classification approach.

Tab. 2.9 shows results of these experiments, along with key parameter settings for the neural networks. Detail results by individual language for the competing cross-entropies transformer model are reported in appendix Tab. A.9.

The Baseline - recurrent group reports our previous result for the competing cross-entropies neural model and a recent replicate from our newer Pybor2 codebase. The Competing cross-entropies - transformer group reports results for our light-weight transformer with the competing cross-entropies approach. We enhanced the competing competing cross-entropies approach to optionally use *Donor* language source as borrowing category in place of the existing *Borrowed* category. Even though donor category is taken into account, only the overall precision, recall, and F1 scores are reported for borrowing detection. The Direct - transformer - flattened group reports on a direct classification approach, still using a lightweight transformer for processing word sound segments, and a final classifier layer which takes the flattened transformer output and classifies words as *Inherited* versus *Borrowed*. This direct method is also enhanced to optionally take into account *Donor* language category in place of the existing *Borrowed* category.

Results of these enhancements were previously reported in (Miller, Pariasca, and Beltran Castañon, 2021).

Cross-validation results are the same for original (Pybor1) and replicated (Pybor2) recurrent neural studies, while execution time is slightly reduced. This indicates the port from Pybor1 to Pybor2 remains faithful to the original.

Results for the light-weight Transformer with competing entropies approach whether classifying borrowings with just a *Borrowed* category or by *Donor* category, are on par with that for the baseline recurrent model, at least for the

TABLE 2.9: Competing cross-entropies and direct model experiments
10-fold cross-validation

Expmnt.	Epochs	Learning	Dropout			Prec.	Recall	F1	Time
		rate	embed	attn.	transf.				hr:min
Baseline - recurrent									
Pybor1	45	0.01				0.549	0.701	0.606	1:22
Pybor2	45	0.01				0.549	0.704	0.606	1:17
Competing cross-entropies - transformer									
Borrowed	50/80	.0075/.0035	0.4/0.2	0.2	0.1	0.556	0.709	0.613	0:57
Borrowed	50	0.0055	0.3	0.2	0.1	0.554	0.709	0.611	0:51
Borrowed	80	0.0055	0.3	0.2	0.1	0.556	0.712	0.614	1:22
Donor	50/80	.0075/.0035	0.4/0.2	0.2	0.1	0.527	0.724	0.599	0:59
Donor	50	0.0055	0.3	0.2	0.1	0.531	0.715	0.601	0:59
Donor	80	0.0055	0.3	0.2	0.1	0.535	0.739	0.610	1:16
Direct - transformer - flattened									
Borrowed	120	0.0025	0.3	0.3	0.3	0.506	0.653	0.556	1:32
Donor	120	0.0025	0.3	0.3	0.3	0.476	0.687	0.547	1:40

Updated from Miller, Pariasca, and Beltran Castañon (2021).

Borrowed case. There is no substantial improvement in adopting the transformer model, other than reduced execution times.

However, we do see a difference in precision versus recall for *Borrowed* versus *Donor* cases. The *Donor* case offers improved recall with reduced precision, and perhaps a slight reduction in F1 score too. The improvement in recall makes sense in that each major donor category has its corresponding language model trained just on words borrowed from that donor.

The expected positive effect of modeling borrowed words by donor, due to consistency of phonotactics, is likely offset by the reduction of number of words available for each entropy model. Paucity of donor data seems a stronger force than a more consistent language source in fitting the model.

The approach of competing cross-entropies is appealing in that cross-entropy seems a reasonable measure of how well we model the phonology and phonotactics of a language. In particular, by using separate language models for inherited versus borrowed words or donor words, the contrast in model predictions of cross-entropy for the same words, seems to capture discrepancies in phonology or phonotactics between models for the same words.

But maybe we are making the problem more complicated than it need be. While attractive, it is not essential that we model the human process in order to detect borrowings. We try with a direct neural classification model instead.

Results of our experiments with a light-weight transformer model to directly discriminate between inherited and borrowed words are reported under Direct - transformer - flattened. All direct trials used 120 training epochs and the same learning rate and dropout parameters. The direct transformer model scores $\approx$5

percentage points lower F1 score than the competing cross-entropies model and takes longer to execute.

The cross-validation results are perhaps not surprising since we have the collaboration of two models with the competing entropies approach instead of one model with the direct approach.

2.2.7 Additional wordlist for dominant donor language

There are many cases where a single or few donor languages account for the overwhelming majority of borrowings into the target language. We constructed the capability for our neural network approach to add a supplemental wordlist to the donor words for a target language. This capability permits us to run experiments with supplemental data such as the following:

1. Use a supplemental Spanish wordlist, represented as sound segments, as though borrowed words from target language, and,
2. Use a supplemental Spanish wordlist, with Spanish sound segments translated to target language, and use as though borrowed words from the target language.

Add supplemental wordlist to target language donor words

In the experiments section by Miller, Pariasca, and Beltran Castañon (2021), we describe this simple yet meaningful supplemental wordlist experiment:

> Several of the Latin American languages in WOLD [Imbabura Quechua, Mapudungun, Otomi, Q'qechi', Wichí, Yaqui, Zinacantán Tzotzil] have Spanish as the primary and only significant language donor. For each of these languages we added the Spanish wordlist in segmented IPA to the existing borrowed words of the training set, and then trained the models and evaluated test performance for these seven languages in a 10-fold cross-validation. This is a crude attempt to take advantage of data quantity with the hope that Spanish phonotactics would translate sufficiently into Spanish borrowed word phonotactics via our light-weight Transformer model. (Miller, Pariasca, and Beltran Castañon, 2021)

Tab. 2.10 shows results on cross-validations for competing cross-entropies and direct approaches using the supplementary Spanish wordlist. Also shown are the cross-validation results for competing cross-entropies and direct approaches using transformer models as well as our baseline recurrent model. F1 score performance is on par between supplementary and not supplementary Spanish wordlists, but there is substantial inversion of recall and precision results. Use of the Spanish supplementary wordlist reduces recall and increases precision. Indeed, with use of a Spanish wordlist, without going through some process of adaption, the *Donor* model should do very well identifying *unadapted* Spanish words, typical of recent borrowings, but miss out on words that have been

Table 2.10: Additional Spanish donor language table experiments - 10-fold cross-validation - over Latin American languages.

Expmnt.	Epochs	Learning rate	Dropout embed	Dropout attn.	Dropout transf.	Prec.	Recall	F1	Time min
With additional donor language table									
Competing	100/150	0.0055	0.1	0.1	0.1	0.825	0.761	0.788	37
Direct	120	0.0035	0.3	0.3	0.3	0.745	0.735	0.731	30
Without additional donor language table									
Competing - Borrowed - Recurrent						0.719	0.868	0.784	14
Competing - Borrowed - Transformer						0.715	0.862	0.777	9
Competing - Donor - Transformer						0.729	0.864	0.787	9
Direct - Borrowed - Transformer						0.664	0.806	0.721	14
Direct - Donor - Transformer						0.650	0.814	0.716	14

Languages: Imbabura Quechua, Mapudungun, Otomi, Q'qechi', Wichí, Yaqui, Zinacantán Tzotzil. Updated from Miller, Pariasca, and Beltran Castañon (2021).

adapted to the target language over substantial time. Thus recall is reduced, but detected borrowings are more likely from the donor language.

2.2.8 Translate donor wordlist to target language sound segments, and add to target donor words

The capability of adding a supplementary wordlist to the already existing borrowed words opens up the possibility of adding additional borrowed words from a borrowed language to that target language without having to restrict the source to wordlists. Here we explore the arduous, complicated, and time consuming experiment to translate Spanish wordlist from IDS into simulated target language borrowed words, and then add these simulated translated donor words as though borrowed words.

Results are reported for both the simulated borrowed word translation and for the competing entropies borrowing detection augmented with simulated word translations tasks.

Translate wordlist to target language wordlist sound segments

First translate the Spanish wordlist to target language sound segments. Test results by target language for each method of selecting the optimal model are shown in Tab. 2.11, with model selection based strictly on training data (§ 2.1.4). Cross-entropy and accuracy largely tracked together; the max accuracy criteria appears marginally better than min cross-entropy or 2000 step criteria.

Test results are disappointing en general in that even with the highest accuracy model, $\approx$ 80% for Imbabura Quechua, a five sound segment word would likely contain at least one error. [i.e., Assuming independence as a false but useful

simplification, $0.8^5 = 0.33$, is the probability of the word being completely accurately represented.] But maybe the simulated words don't have to be entirely accurate, just sufficiently similar to actual borrowed words.

TABLE 2.11: Translate Spanish wordlist to simulated borrowings - translation results.

Language	Opt-step	Test CE	Test Acc
Imbabura Quechua	min CE	0.97	0.79
	max Acc	0.91	0.84
	2000	1.03	0.81
Mapudungun	min CE	2.33	0.57
	max Acc	2.33	0.57
	2000	2.91	0.61
Otomi	min CE	2.05	0.60
	max Acc	2.35	0.67
	2000	2.61	0.66
Q'eqchi'	min CE	2.35	0.56
	max Acc	2.35	0.56
	2000	2.51	0.64
Wichí	min CE	1.30	0.71
	max Acc	1.48	0.73
	2000	1.74	0.77
Yaqui	min CE	1.09	0.71
	max Acc	1.57	0.71
	2000	1.34	0.73
Zinacantán Tzotzil	min CE	1.24	0.71
	max Acc	1.57	0.73
	2000	1.09	0.75

Word sound sequence translation accuracy for Quechua is better than for the remaining languages, but still inadequate to assure error free simulated borrowed words more than one-third of the time.

Add simulated words to target language donor words

Next add the translated simulated words to the target language donor words. Overall test results for borrowing detection using the competing cross-entropies approach are reported in Tab. 2.12. F1 scores for simulated borrowings are little better than baseline results, just a 1 to 2 percentage point improvement. It seems the translation results were not good enough to substantively improve borrowing detection.

However, there are substantive differences in the precision and recall. Just as for (Miller, Pariasca, and Beltran Castañon, 2021), the Spanish wordlist augmentation baseline shows lower recall and higher precision, likely indicating

adequate modeling of recent Spanish borrowings, but not so for better adapted older borrowings. Opposed to this are the results for simulated augmented borrowings where recall is several percentage points higher versus the unaugmented and Spanish augmented baselines. This suggests better modeling for Spanish borrowings in general, but not without errors, thus resulting in a loss of precision.

TABLE 2.12: Translation of Spanish donor table - overall borrowing detection.

Experiment	Precision	Recall	F1
One-fold CV Benchmarks			
No augmentation	0.793	0.773	0.781
Spanish augmented	0.829	0.748	0.784
Simulated Augmented Borrowings			
Max accuracy	0.778	0.827	0.801
Min cross-entropy	0.775	0.815	0.792
2000 training steps	0.755	0.829	0.787

Single-fold cross-validation benchmarks for comparison purposes and results for simulated borrowings for each criterion for selecting an optimum translation model.

Inclusion of simulated borrowings has a clear impact on the borrowed word language model which is now better adapted to recognize borrowings. This is made more obvious by contrasting the profiles of unaugmented and simulated augmented borrowings from Fig. 2.17. Even so, overall detection performance as shown by F1 score remains little changed.

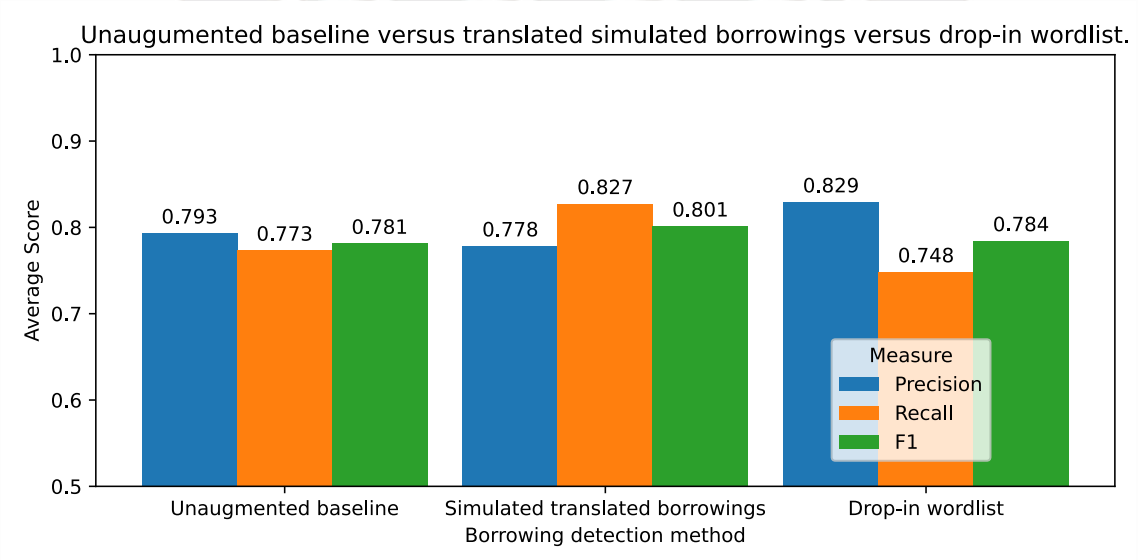


FIGURE 2.17: Borrowing detection for unaugmented versus simulated borrowings augmentation in training.

Test precision, recall, and F1 scores are reported in Tab. 2.13 for each target language for the unaugmented single-fold baseline, augmented translated simulated borrowed words, and for the augmented (drop-in) Spanish wordlist.

Average results for unaugmented single-fold baseline are similar to those for the 10-fold baseline previously reported. Comparisons of individual language results should be made with respect to their corresponding one-fold baseline. Neither augmented Spanish nor augmented translated borrowings consistently increase F1 score relative to their baseline. However, **in all cases**, recall improves for the augmented translated borrowings over both the unaugmented baseline and the simple drop-in of a Spanish wordlist. Similarly, in most cases precision decreases. The effect is consistent.

Table 2.13: Translation of Spanish donor table - borrowing detection by language.

Language	Unaug. baseline			Aug. Sp. translation			Aug. Sp. drop-in		
	Prec.	Recall	F1	Prec.	Recall	F1	Prec.	Recall	F1
I. Quechua	0.746	0.749	0.746	0.726	0.811	0.766	0.851	0.757	0.801
Mapudungun	0.786	0.767	0.775	0.827	0.826	0.826	0.829	0.689	0.752
Otomi	0.822	0.741	0.778	0.828	0.821	0.824	0.859	0.753	0.801
Q'eqchi'	0.679	0.665	0.671	0.626	0.687	0.654	0.721	0.565	0.632
Wichí	0.806	0.774	0.789	0.757	0.795	0.775	0.815	0.763	0.787
Yaqui	0.777	0.829	0.802	0.759	0.855	0.804	0.780	0.805	0.792
Z. Tzotzil	0.932	0.886	0.907	0.923	0.995	0.957	0.946	0.905	0.924
Average	0.793	0.773	0.781	0.778	0.827	0.801	0.829	0.748	0.784

Results from (Miller, Pariasca, and Beltran Castañon, 2021) for single-fold train and test split - unaugmented baseline, augmented with translated simulated borrowings from Spanish, augmented with unmodified Spanish wordlist.

Since we maintained the same train and test datasets over baseline and augmented training data studies, we can perform statistical tests of whether there is an improvement in detection versus the baseline. The paired t-test accommodates the case where sample groups are the same and an experimental factor varies (Dixon and Massey, Jr, 1983). [3] Differences were computed between unaugmented baseline and augmented translated borrowed words for F1 scores, precision and recall (Tab. 2.13). For each response the average and standard error of differences were calculated, which were then used to calculate paired t-statistics and p-values. There is a clear signal of increased recall, possible improved F1 score, with no significant change for precision versus the unaugmented baseline.

[3] See (Dror et al., 2018)'s "hitchhiker's guide" for a fuller discussion of statistical testing in NLP with mention of classification and cross-validation measures in particular. Our use of paired t-test and later use of mixed effects models addresses these issues.

TABLE 2.14: Borrowing detection statistical results - single fold - translated Spanish words versus unaugmented and versus drop-in Spanish words.

	Translation vs. unaug.			Translation vs. drop-in		
Statistic	Prec.	Recall	F1	Prec.	Recall	F1
Average Delta	-0.015	0.054	0.020	-0.051	0.079	0.017
Standard Error Delta	0.012	0.013	0.011	0.017	0.015	0.013
t-statistic	-1.209	4.283	1.734	-2.989	5.340	1.265
p-value	N.S.	< 0.001	< 0.1	< 0.05	< 0.001	N.S.

2.3 Discussion

Our most successful innovation in monolingual methods for detecting lexical borrowings has been the *competing cross-entropies approach*. It is enough better than the inherited words only approach, that we largely ignore the inherited words only approach in the following discussion. Yet we should recall that the original incentive for using cross-entropy word models is from the idea that a primary language speaker often just knows when a word sounds like it is part of the language. That is, our initial concept was that of an inherited word only approach inspired in an indigenous speakers knowledge of their language. The cognitive and linguistic explanation of lexical borrowing detection is seemingly more complex than our simplistic starting point.

2.3.1 Bag of sounds, Markov, and neural methods

Artificially seeded borrowings. As we saw previously and noted in Miller et al. (2020):

> In our artificially seeded borrowings experiment, we simulated very close, intensive, and recent language contact, where borrowed words were transferred without alteration. All methods performed well when the proportion of artificially borrowed words was high, and degraded differently when borrowings decreased. (Miller et al., 2020)

While bag of sounds outperformed cross-entropy methods on precision, the competing cross-entropies methods outperformed the bag of sounds method in recall. Inherited only models underperformed on F1 score even on this experiment.

Bag of sounds performs well on precision here because it decides whether the word is borrowed based just on the occurrence of sound segments seen only in borrowed words. So when it identifies a word as borrowed, it most likely is borrowed. It's recall performance suffers only when the phonotactics of words indicate borrowings without using specifically borrowed word only sound segments. Competing cross-entropies models perform well on recall, because the borrowed and inherited word models together take into account the phonology and phonotactics of borrowed and inherited words.

Borrowing detection in real world language data. In our real world borrowings experiment, we performed a 10-fold cross validation of lexical borrowing detection across all 41 WOLD wordlists. The competing cross-entropies methods performed much better than the bag of sounds method, with the recurrent neural network performed marginally better than the Markov chain.

This marginally better performance of the competing cross-entropies approach using neural network word models with precision $= 0.549$, recall $= 0.701$, and F1 score $= 0.606$ is not good enough to be useful, by itself, as a computer assisted tool in historical linguistic investigations. Somewhat arbitrarily, a goal precision, recall, and F1 score of 0.80 would seem the price of entry. So for the rest of this and subsequent chapters, we consider ways in which we can improve on these results.

Inherited only methods performed less well overall than their corresponding competing cross-entropies methods. However, they scored better on recall, a result of estimating from training data the inherited cross-entropy cutoff needed to achieve 80% recall on test data. But since the inherited and borrowed cross-entropy distributions overlap, there was a substantial reduction in precision and F1 score on test data using inherited only approach.

A key factor favoring the neural networks is that it includes conditional dependencies from all previous sound segments, without having to estimate extra parameters for this dependency. The marginal lack in performance of Markov chains may be due to estimation with limited conditional dependency (3-gram) and excess parameter estimates impacting reproducibility. The bag of sounds still maintains a better precision than competing entropies methods, but with very low recall. Because of its dependence on unlikely sound segments, it misses lots of even moderately adapted borrowings. When it does identify a borrowing, it's more likely to be correct, making the bag of sounds method more conservative than the other methods.

Inspection of individual language results revealed as we noted in Miller et al. (2020):

> When the overall proportion of borrowed words in wordlists is small, all models perform poorly. This is not surprising, since low borrowing proportions make it difficult to learn the phonotactics or phonology of borrowed words, if these can be identified at all. It is also not clear to which degree trained linguists would be able to identify borrowed words in the respective languages and even less so over entire wordlists instead of just recent borrowings, if they were given only monolingual information alone. (Miller et al., 2020)

Factors influencing borrowing detection. Because of the disappointing results with the real language data, we looked at major factors that might influence the performance of borrowing detection methods. Besides the obvious proportion of borrowings, we considered proportions of sounds occurring exclusively in

borrowed words and sounds occurring exclusively in inherited words (Miller et al., 2020).

We found "the effect of the proportion of borrowed words was remarkable, showing a strong linear increase in performance for all methods when the proportion of borrowed words was 5% and more" (Miller et al., 2020). The proportion of sounds occurring exclusively in borrowed words was positively correlated with performance of Bag of Words, and with neural network and Markov chain competing entropies methods, while the proportion of exclusively inherited sounds had little bearing on performance. This suggests that "modeling phonotactics with Markov Model and Neural Network methods also takes good advantage of the simple occurrence of borrowed sounds in words too."(Miller et al., 2020)

Detecting borrowings from a single dominant language. Following up to our finding of the importance of having a large proportion of borrowed words, we wanted to know if borrowing from a single dominant language donor resulted in better borrowing detection than just having lots of borrowings across various donors. The thinking here is that a singe dominant donor should result in more consistent and better trained language models for borrowed words.

We observed here and in Miller et al. (2020):

> Since we create lexical language models for borrowed and inherited words, it is straightforward to question why our basic approach would treat all borrowed words as if they come from a single donor language. While it may hold for specific contact situations that a given language is heavily influenced by one single, dominant donor language, it is also possible that borrowings form distinct layers in the lexicon of a given language, reflecting borrowings from different donor languages and different times. If the majority of the borrowings attested in a given language stem from a single donor, however, we would assume that our lexical language model approaches to monolingual borrowing detection would perform better, since the donor language which we access through the recipient language would provide a much more coherent and consistent picture than would a mix of words from different donor languages.
>
> We therefore systematically tested whether the performance of our methods would increase for those wordlists in our sample for which a dominant donor language could be identified. Our assumption, that the methods should show an increased performance for languages with a dominant donor language were largely confirmed, as reflected in substantially increased F1 scores of ≈ 0.75 for the Markov Model and the Neural Network methods in cases of high contact with more than 300 borrowings. While we still consider the overall performance of the monolingual borrowing detection disappointing, this experiment reflects the importance of having a consistent sample of the

> donor language when dealing with monolingual borrowing detection. (Miller et al., 2020)

With F1 scores at ≈ 0.75 for both Markov chain and neural network methods using the competing cross-entropies approach, we are within sight of our earlier informal goal of 0.80 to be useful in a computer assisted method of lexical borrowing detection. Of course, the context is pretty restrictive requiring more than 300 borrowed words and a dominant donor (i.e., $\geq 2/3$ of borrowed words from a single donor).

Comparing cross-entropy distributions. We previously discussed the usefulness of examining the distribution of cross-entropy differences in Miller et al. (2020).

> [Our graphical...] evaluation was intended to demonstrate how the Markov chain and neural network methods discriminate between inherited and borrowed words. We showed how plots of the distribution of cross-entropy differences between competing inherited and borrowed word models served to explain borrowing detection results. Comparing the distributions of cross-entropy differences, we saw that in cases where the proportion of borrowings was small, the discriminative force of the word cross-entropy differences dropped drastically for testing. Where the proportion of borrowings seemed adequate for training, we saw cases of a reduction in discriminative force for testing due to reduced separation of inherited and borrowed word cross-entropy difference distributions. (Miller et al., 2020)

So even though we understand the distribution of cross-entropy differences between inherited and borrowed language models, this does not solve our problem of too little borrowed word data for training. Even when training seems effective, there can be substantial variance in the test results, if training lacks sufficient data or appropriate tuning of architectural, learning, and regulation parameters.

2.3.2 Enhanced neural network experiments

Given results from bag of sounds, Markov chain, and neural network methods, especially our good results using the competing entropies approach with Markov and neural methods, we decided to focus on neural network methods. The competing entropies approach with neural network was only marginally better than even the Markov chain, but the neural network approach offers many more opportunities for experimentation and improvement such as in architecture, learning and regulation. So after replicating results from earlier our earlier study (Miller et al., 2020), we enhanced our baseline model and used that for further experimentation.

These results we previously discussed in Miller, Pariasca, and Beltran Castañon (2021):

> We developed a light-weight Transformer model (Bahdanau, Cho, and Bengio, 2015; Vaswani et al., 2017) and observed that it performed minimally better than par versus the recurrent model in borrowed word detection and was more responsive with reduced execution times. The light-weight Transformer offers a viable base for exploring different lexical borrowings detection approaches.
>
> Our meta-level experimental design, contrasts Transformer versus the recurrent model results, and forms a 2-factor design of 1. competing entropies versus direct approaches, and 2. inherited and borrowed versus inherited and word donor approaches.
>
> The light-weight Transformer with competing entropies approach performed five percentage points better than the light-weight Transformer with a direct approach. Competing entropies seems a useful approach to test for lexical borrowings, more so than fitting a larger but less meaningful [direct] neural model.
>
> With inherited versus donor models, we tested whether modeling donors separately would result in better prediction performance, on the basis that treating donors individually should give more coherent borrowed word samples. Resulting performance was just on par with the corresponding borrowed word approach; no benefit was conveyed by modeling donors separately. This suggests that any benefit due to modeling more coherent language subsets is offset by the reduced sample size for such subsets. This result seems more likely an artifact of insufficient data rather than a dismissal of the utility of modeling donors separately. (Miller, Pariasca, and Beltran Castañon, 2021)

2.3.3 Additional wordlist for dominant donor language

We previously commented: "Lack of sufficient data is a major detractor in obtaining good model fits and reproducibility on test cases; this is especially true for highly parameterized models where the parameter counts equal or exceed the data counts"(Miller, Pariasca, and Beltran Castañon, 2021). So "lack of sufficient data" becomes an improvement opportunity for us. We responded by creating the capability to augment borrowed wordlist training data with an *ad hoc* wordlist of presumedly borrowed words. Then we performed experiments on wordlists where Spanish is the dominant language.

In our first experiment we augmented borrowed wordlist training data with a Spanish language wordlist for target (recipient) language wordlists where Spanish was the dominant donor language – as we reported in Miller, Pariasca, and Beltran Castañon (2021).

> We added training data from the Spanish language wordlist as though they were borrowed words, fake borrowed words, for each of the seven WOLD language tables where Spanish is the primary donor language.

> Transformer models learned from actual and fake borrowed words produced F1 scores just on par with the competing entropies approach on borrowed words alone. Recall decreased and precision increased. This indicates that the Transformer model learned the Spanish wordlist so well that it no longer detected borrowed words that were better adapted to the recipient language. Similarly the better learning prevents language recipient inherited words from being confused with Spanish borrowed words. This suggests that with a table of fake borrowed words that conforms more faithfully to word adaption to each language, we could see a more sizable improvement in detection of borrowed words. We observed an improvement for direct detection too, but still poorer performance versus the competing entropies approach. (Miller, Pariasca, and Beltran Castañon, 2021)

This then sets us up for a followup experiment where we try to better construct *fake* borrowed words in the recipient (target) languages.

2.3.4 Translate donor wordlist to target language sound sequences and add to target donor words

After seeing the impact of using a Spanish wordlist as though borrowed words, we determined to simulate a wordlist of *translated fake* borrowed words for WOLD languages where Spanish was the dominant donor. We hoped this would show a more substantial improvement in detection when used to augment borrowings. We've termed this effort to create *fake* borrowed words as *translation*, but translation of sound segments from donor to target language adapted borrowings semblance.

Translation results for creating simulated borrowings for training the borrowed words language model were disappointing. It was likely that many simulated borrowings would not precisely simulate borrowed words in the target language based on the sound segment accuracy of $\leq 0.80\%$. Even so, we thought that it might be *close enough* that it could adequately represent what borrowed words would look like!

So we adopt simulated borrowed word translation in an effort to augment data for the borrowed word language model. But it's questionable that the augmented data is of adequate quality to meet that need. Just as our original borrowing detection approach suffered from lack of adequate data to train the model well, so it seems might our approach to simulate and augment training data.

It was not too surprising then when the results of data augmentation with simulated translated borrowed words showed little improvement in F1 scores over baseline datasets with unaugmented borrowed words. There was a substantial increase in recall on average and for each language. Use of augmented simulated borrowed words in training made the borrowed word language model fit actual borrowed words better, with lower cross-entropy, resulting in a higher

recall. This beneficial effect was partially offset by the model fitting some inherited words better as well, resulting in more inherited words also classified as borrowed, and so reducing precision in some cases.

The sample of seven South American indigenous languages individually and on average has a much better F1 score than the complete sample of WOLD languages. So there was also less opportunity to improve the overall result in any event. Maybe a language case with lower original F1 score, but sufficient borrowings to train a simulated borrowings models, would have fared better. Moreover, instead of being satisfied with just the words from WOLD tables, a more ambitious effort could sample more borrowed words for training a translation model. Alternatively, there may be better translation models or other methods for simulating borrowings that would improve recall without negatively impacting precision.

Scare resources learning. While training with simulated borrowings was effective in increasing borrowed word recall, it had little overall effect on overall borrowed word detection. We tried to squeeze more benefit out of the same data using similar methods, and saw little improvement.

2.4 Conclusions

Beginnings. We began with bag of sounds, Markov chain, and neural network supervised methods, for the detection of lexical borrowings in monolingual wordlists. Our rationale for this method selection and their use we explain in Miller et al. (2020):

> These methods are based on lexical language models and are intended to model specific aspects of phonology and phonotactics in the lexicon of spoken languages. Assuming that phonological and phonotactic properties of words in the lexicon of a spoken language can provide enough clues to identify borrowings by language-internal comparison of words alone, we designed workflows in which the lexical language models could be trained with monolingual wordlists, with borrowings are already annotated, and then used to detect borrowings when being confronted with so far unobserved words. (Miller et al., 2020)

Beyond the application of lexical language models to model sequences of sound segments (the phonology and phonotactics) of languages, our most important innovation was to develop and apply a *competing cross-entropies* approach. In this approach we trained models on training data for inherited and borrowed words separately. Then models competed over previously unobserved words to see which obtained the lower cross-entropy for each word, with the lower cross-entropy result serving to discriminate between inherited or borrowed words. Our competing cross-entropies approach was superior to the inherited word only approach. All lexical language model methods were superior to the weak

baseline bag of sounds method, which used only the *set* of sound segments to discriminate between inherited or borrowed words.

Overall, tests on real wordlists taken from the WOLD database revealed disappointing performance even on competing cross-entropies methods, Markov chain and neural network. Attempts in Miller et al. (2020) to identify the potential reasons for this inadequacy

> revealed two main factors that considerably influence how well the methods performed, namely (1) the amount of borrowings in a given language variety, and (2) the uniformity of the borrowings in a given language variety, as reflected in the presence of a dominant donor language. While the first factor reflects the importance of having enough training data when working in supervised learning frameworks, the second factor reflects more specific linguistic conditions of monolingual borrowing detection. Our methods identify borrowings primarily from phonological and phonotactic clues, and perform better in those cases where the words' properties are coherent and consistent. This is generally the case for inherited words, and also for words that were borrowed from the same donor language [within a given epoch]. (Miller et al., 2020)

The competing cross-entropies approach using Markov chain and neural network methods provided a valuable and promising baseline for the further exploration of monolingual approaches to lexical borrowing detection. From this point we chose to pursue more enhanced and state of the art neural network language models, as well as, in a subsequent chapter, look at methods that consider multilingual sources of information as well as mixed monolingual-multilingual solutions (§3).

Enhanced neural networks. Focusing on the Neural Network method as a useful technical direction to enhance our work, we constructed a responsive light-weight transformer model as a lexical language model for experimentation in the detection of lexical borrowings. With the transformer model we were able to continue experiments within the competing cross-entropies approach, where we considered multiple donor language models, for major donors, in place of a single borrowed word language model, combining all donors. This was to test our hypothesis resulting from earlier experiments, where we posited that borrowings from a single donor should be more uniform and so result in better language models with the end result of better borrowing detection. We also experimented with a direct classification model where the light-weight transformer still processes word sound segment sequences, but where output of the transformer feeds a logistic regression based classifier.

Detection performance of the competing cross-entropies light-weight transformer was on par with the recurrent neural network model. Direct detection of lexical borrowings using a light-weight transformer showed poorer performance than the competing cross-entropies approach. Competing cross-entropies seems to capture important evidence about lexical borrowings that our direct model

does not. Language donor based detection models performed on par with corresponding borrowed word models. Perhaps paucity of training data detracts more from the method than same donor language coherence contributes.

Training wordlist augmentation. In an attempt to solve the problem of sparse data, we developed the capability to add *ad hoc* wordlist data to augment borrowed word training data. In a preliminary experiment we added a dominant donor language (Spanish) wordlist to target language borrowed words for training, where Spanish was the dominant donor language, as though the added words were borrowings. This resulted in reduced recall and increased precision, but little impact on F1 score. We mext hypothesized that with better simulated (translated) borrowed words, rather than simple adoption of the donor wordlist, recall and precision would improve.

Training wordlist augmentation with translated word sound sequences In a much more ambitious experiment, we developed sound sequence translators from a dominant donor language (Spanish) to target languages, and then generated *fake* supplemental borrowed words lists for each target language. These fake borrowed wordlists were added to training data for borrowed word detection as though they were borrowed words. Translation quality for the wordlists was mediocre, because there was insufficient data to train translators with adequate accuracy. Even so, we tested the translated wordlists as augmented training data in borrowing detection. There was an increase in recall in all cases with a largely negative effect on precision and little overall impact on F1 score. Simulated borrowed words did improve recall, but translation quality seemed inadequate to make an overall improvement.

It's unclear whether augmentation of training data with simulated borrowed words via a translation model could produce substantially improved monolingual borrowing detection overall in spite of the improved recall effect. But there might be other options for developing augmented borrowed word training data. In trying to improve results from inventories of scant data, we should experiment with more dramatic changes in the methodologies we are using in order to increase our possibilities of success.

Subsequent opportunities. Multilingual or cross-linguistic methods alone or in combination with a monolingual based approach, such as inherited and borrowed word competing cross-entropies, offer other opportunities for improvement. Our monolingual methods add to the growing pool of automated approaches to lexical borrowing detection which could eventually be combined into an integrated workflow – a workflow in which evidence from monolingual and multilingual sources would form a unified picture of borrowing detection and more holistically, language contact.

Chapter 3

Multilingual borrowing detection

Lexical borrowing is a pervasive phenomenon and one of many results of language contact, often where one or a few languages dominate the rest. "Most computational approaches to borrowing detection treat all languages under study as equally important, even though dominant languages have a stronger impact on heritage [minority or ancestral] languages than vice versa" (Miller and List, 2023).

We explore and evaluate methods for lexical borrowing detection from wordlists in contact situations where dominant languages play an important role, for a sample of seven Latin American languages which have all borrowed extensively from Spanish (Miller and List, 2023). Methods include pairwise and cognate based classical sequence comparison methods with either normalized edit or sound class alignment distance measures, a cross-entropy comparison method based on language word models, and support vector machine and logistic regression machine learning meta-models combining the above comparison methods and distance measures.

In our initial multilingual borrowing detection foray (Miller and List, 2023), classical multilingual sequence comparison methods performed well with the support vector machine meta-model improving upon individual method performance. Error analysis showed, however, that absent donor words and divergent meanings of donor from recipient words accounted for the bulk of errors, and offered a significant improvement opportunity.

We explored possibilities of 1. augmenting donor words to address the problem of absent donor words, 2. relaxing the same concept requirement to address the problem of divergent donor and recipient concepts, and 3. adding a monolingual cross-entropy function to complement the multilingual methods. There is clear value to be gained by augmenting donor words. Relaxing the same concept requirement is not so clear cut and technically more complex, but it seems worth continued investigation. Adding the competing cross-entropies function to the existing meta-model combining methods and distance measures, results in a substantial improvement in borrowing detection over just multilingual methods.

Diversity wins.

Parts of this chapter were previously reported in (Miller and List, 2023) and will be cited as appropriate.

3.1 Materials and methods

3.1.1 Materials

With our focus on dominant donor languages in multilingual methods of borrowing detection, we constructed a database of seven Latin American languages, plus Spanish as the dominant donor language, with which to develop and prove out our methods. In the previous Monolingual chapter (§2.2.7), we selected these same Latin American languages from the WOLD database and added via and *ad hoc* procedure variants on a Spanish wordlist. Here we formalize this selection of languages and the addition of the Spanish wordlist as its own database for this part of our borrowing investigation. This initial foray of developing our research specific database is particularly appropriate in that in subsequent work we plan to develop a much more expansive Pano-Tacanan Borrowing database of focused on Pano-Tacanan languages with both Spanish and Portuguese as dominant donors, to showcase methods developed here and in the previous monolingual chapter (§2.1).

We quote from our "Detecting lexical borrowings from dominant languages" (Miller and List, 2023) paper to provide the new database details:

> For this study, a new comparative wordlist was created by taking data for seven Latin American languages from WOLD (https://wold.clld.org, Tresoldi, Forkel, and Morozova 2019) and combining them with a wordlist of Spanish derived from the Intercontinental Dictionary Series (https://ids.clld.org, Key and Comrie 2015). Phonetic transcriptions for the Latin American languages were added to WOLD (Miller et al., 2020). Latin American Spanish phonetic transcriptions were added for this study, and these could later be expanded by adding more transcriptions from historical varieties of Spanish.
>
> The resulting dataset conforms to the standards suggested by the Cross-Linguistic Data Formats initiative (CLDF, https://cldf.clld.org, Forkel et al. 2018). The data curation follows the Lexibank workflow (List and Forkel, 2022) and checks that data conform to certain standards, with languages being linked to Glottolog (https://glottolog.org, Hammarström, Forkel, and Haspelmath 2021, Version 4.7), concepts being linked to Concepticon (https://concepticon.clld.org, List et al. 2022a, Version 3.0), and transcriptions following the B(road)IPA conventions of the Cross-Linguistic Transcription Systems reference catalog (https://clts.clld.org, List et al. 2021, Version 2.2, see Anderson et al. 2019).
>
> Details of the resulting database are shown in the map of language locations along with percentages for borrowings from Spanish in Fig. 3.1 and in Tab. 3.1. Q'eqchi' and Zincantán Tzotzil are both Mayan languages, but appear substantially varied in the database. (Miller and List, 2023)

FIGURE 3.1: Map of languages with Spanish borrowing class.
Adapted from Miller and List (2023).

TABLE 3.1: Database details by language for seven Latin American languages plus Latin American Spanish.

Language	Concepts	Lexemes	Segments	Vocab.
Imb. Quechua	1,155	1,156	7,177	33
Mapudungun	1,040	1,242	7,356	33
Otomi	1,252	2,241	11,730	57
Q'eqchi'	1,211	1,773	10,367	49
Wichí	1,128	1,219	8,233	44
Yaqui	1,242	1,433	9,297	28
Zin. Tzotzil	955	1,266	7,129	41
Spanish	1,308	1,770	11,261	30
Aggregate	1,308	12,100	72,550	112

Adapted from Miller and List (2023)
Language details added.

3.1.2 Methods

Methods for multilingual borrowing detection makeup most of this subsection, but because of semantic restrictions on matching words from multilingual wordlists and our dominant donor language focus, adjustments are made to previous sampling and evaluation methods.

Dominant donor

In this part of our investigation, we have focused on detecting borrowings from dominant donors (previously defined as 67% of all borrowings in §2.2.4 for use with WOLD), but here simply defined as Spanish for the seven Latin American languages in our study. Given the dominant donor language we automatically infer the direction of borrowing as from the dominant donor language to the target language where the borrowing is detected.

The borrowing detection problem becomes – word not borrowed from dominant donor versus word borrowed from dominant donor. This has implications for borrowing detection methods as well as for the evaluation of these methods:

1. In training and test data, the detection distinction becomes – word not borrowed from dominant donor versus word borrowed from dominant donor. The criterion variable has changed from the previous *inherited* versus *borrowed* distinction.

2. In evaluation of borrowing detection on test data, performance, precision, recall, and F1 scores are calculated based on category distinctions of *not borrowed from dominant donor* versus *borrowed from dominant donor*. There is no change to the evaluation methods, but the criterion variable has changed.
3. A new classification error becomes – incorrectly inferring the direction of borrowing, when the borrowing is really to the dominant language.

The limitation to a single dominant donor and the simple inference rule that the borrowing comes from the dominant donor is an interim step in a more completely defined methodology of borrowing detection including direction of borrowing.

Same concept restriction

The sequence of sound segments (spoken word) used by a language to express a concept are largely independent of the concept meaning. As a result, in a multilingual context where some sound segments are similar across language, we expect to find many similar sounding words that have no semantic relationship whatsoever, and so not borrowings. To inoculate against this problem, the Comparative Method (Campbell, 2013) asks that semantics be taken into account when testing for cognates or borrowings. So for multilingual approaches, wordlists are organized by concepts, and comparisons between words can be restricted to just words expressing the same concept.

We further explore the impact of this same concept restriction and possible solutions in §3.2.5.

Methods for borrowing detection

The basis for multilingual borrowing detection methods described here was introduced in §1.3.2. Our "Detecting lexical borrowings from dominant languages" paper (Miller and List, 2023) describes methods for multilingual borrowing detection as follows:

> We develop three different methods for the detection of borrowings *from* a dominant language *to* non-dominant languages in multilingual wordlists. Following historical linguistics comparative method practice (Campbell, 2013), only word forms corresponding to the same concept are considered as candidates for borrowing.
>
> The first method, called *Closest Match* borrowing detection in the following, iterates over all word pairs that express the same concept in the dominant language and the heritage languages and then computes phonetic distances. Word pairs whose phonetic distance is below a certain threshold are judged to be borrowings from the dominant language. We test two phonetic distances, the normalized edit distance (NED) – the classical edit distance (Levenshtein, 1965) between two words, divided by the length of the longer word – and the SCA distance (List, 2012).

> The second method, called *Cognate-Based* borrowing detection in the following, follows the approach by (Hantgan, Babiker, and List, 2022): it first computes cognates using a cluster-based approach for automated cognate detection in which words expressing the same concept whose average phonetic distance is below a certain threshold are assigned to the same cognate set (List, Greenhill, and Gray, 2017), and then identifies all words assigned to cognate sets involving the dominant language as borrowings. We tested again normalized edit and SCA distances.
>
> The third method, called *Classifier* borrowing detection in the following, iterates over all word pairs with the same concept, but stores phonetic distance scores for various distance measures as vectors, which can then be used to train a classifier, firstly, a linear Support Vector Machine (SVM) (Cristianini and Shawe-Taylor, 2000), in a supervised setting. (Miller and List, 2023)

Our dominant donor focus has some implications for closest match and cognate based sequence methods. With closest match, even with multiple language targets for borrowing detection it is only necessary to compute pairwise distances between the dominant donor words and target language words, since distances between different target words are not useful in this context. With cognate-based, all words for a given concept are already considered and aligned when found to match. But we discard matches which do not include the dominant donor language and retain only matches with the dominant donor.

The sound class based alignment method (SCA) (List, 2012), used in both the closest match and cognate-based methods as well as indirectly through functions provided to the classifier, provides many options for configuring the alignment and measurement process. While we have experimented with several different options, for the purposes of this experiment, we accepted the default SCA configuration with the exception of tests at each of *global*, *overlap*, and *local* alignment modes. Alignment mode indicates how much of the sound sequences being aligned need be included in the final alignment and measurement, where *global* means everything, *local* means just the sequence in common without any overlaps, and *overlap* means everything less an overlap from one side.

We report only on the combination of normalized edit distance (NED) and the global SCA distance, since local and overlap alignment modes were never better than the global SCA distance. Both closest match and cognate-based methods require a fixed threshold which we estimate from the training data. So all three methods are considered as supervised (Miller and List, 2023).

Our use of the *classifier* method is a significant innovation. The *classifier* qualifies as a meta-method as it subsumes and combines finer grained closest match and cognate-based methods. Instead of having to chose between normalized edit (NED) or SCA distance from the closest match method, we can choose both and the classifier will optimize the combination to best predict borrowing detection. Once the classifier architecture was it place, it was straight forward to extend beyond the initial linear support vector machine (SVM) to a *radial basis function*

SVM, linear SVM with *weight balanced* sampling of borrowing categories, and a *Logistic regression*.

Besides incorporating effects for selected closest match functions, where the default is both the normalized edit distance (NED) and sound-class based alignment (SCA) distance, and cognate-based function, where none is the default, various other factors may be added to the classifier. By default, an indicator function is added for each target language to handle the possibility that different languages have a different propensity to borrow from the dominant donor language.

The classifier method could support more complex factors such as the pairing of target language indicator function with individual distance functions for the case where target languages vary in the weight that should be assigned to distance measures. We have not pursued such models at this time and think to handle this case with a more powerful neural network based classifier with a fully connected hidden layer between inputs and the final classification layer.

The cognate-based function as a data source for the classifier, provides indicator functions of which target languages borrowed from the dominant donor language. It would be practical to use the optimal threshold from the cognate-based (CB) method and provide a single indicator function of whether the CB classifies the word as borrowed from a dominant donor or not. However, since the classifier optimizes the the combination of inputs, our CB function provides four separate indicator variables based on a range of thresholds that includes the optimal CB method value.

The previous *Monolingual borrowing detection* chapter introduced the approach of *competing cross-entropies* using Markov chain and neural network language model methods (see §2.1.1). The Markov chain method has been re-implemented here as a stand-alone method called *Least Cross-entropy* (LCE) in the following, essentially replicating the previous competing cross-entropies method. Importantly, the least cross-entropy method makes the cross-entropy calculations available as an invocable separate function. With this, the *classifier* meta-method can incorporate both inherited and borrowed (or donor) cross-entropies from least cross-entropy along with already available phonetic distance measures from closest match or cognate-based methods. While the advantage offered with the classifier for combining distance measures was important, the incorporation of cross-entropy measures is a *significant upgrade to that innovation*.

Sampling and analysis

The restriction to comparison of words for the same concept effects how train and test data are split in order to evaluate reproducibility of a method. Our previous "Detecting lexical borrowings from dominant languages" paper (Miller and List, 2023) describes these effects for sampling and subsequent analysis and evaluation:

> Train-test splits are made based on concepts rather than individual word entries. This permits matching of words for the same concept

> in all methods without loss of candidate words. Treating train-test split as a nuisance variable takes into account differences between partitions across methods thus controlling for effects of sampling by concepts with differing borrowing behavior or statistical dependencies between test partitions due to sampling without replacement. See (Dror et al., 2018) for mention of the dependency problem with cross-validation. Our use of a fixed partition across treatments and analysis of variance controlling for partition as a nuisance or 'blocking' variable accounts for this dependency, and takes advantage of any systematic effects in borrowing behavior by partition.
>
> For the analysis of the cross-validation data, we use a randomized blocks design where experiment is the treatment or factor, and test partition is the randomized block or nuisance variable. A standard analysis of variance partitions treatment effects, nuisance variable, and error, and permits a more powerful test of treatment differences without the nuisance variable variance. We follow up statistically significant findings for treatment (experiment) with comparisons of experiments versus the overall average using a joint (family) error rate (Nelson, Wludyka, and Copeland, 2005). (Miller and List, 2023)

Implementation

Implementation tools used here are described in Miller and List (2023):

> Our methods are implemented in Python, specifically making use of the CLDFBench package (https://pypi.org/project/cldfbench/, Forkel and List 2020, Version 1.13.0) to provide command line access to all methods described here. For the computation of alignments and edit distances, LingPy (https://pypi.org/project/lingpy, List and Forkel 2021, Version 2.6.9) is used. SVM [also Logistic regression] and evaluation are realized with the help of Scikit-Learn (https://pypi.org/project/scikit-learn/, Pedregosa et al. 2011, Version 1.2.1). (Miller and List, 2023)

Furthermore, access to repositories of code and data for this study are also provided in Miller and List (2023):

> The data and code needed to replicate the results reported here, along with detailed information on installing and using the software is curated on GitHub (https://github.com/lexibank/sabor, Version 1.0) and has been archived with Zenodo (https://doi.org/10.5281/zenodo.7591335). (Miller and List, 2023)

3.2 Results

3.2.1 Detecting lexical borrowings in multilingual wordlists

We previously reported on closest match and cognate-based, and classifier experiments in Miller and List (2023). We gave a brief description of experimental methods and cross-validation:

> We tested our three methods with two distance measures in five experiments (normalized edit and SCA distances individually in both Closest Match and Cognate-Based methods, and combined in the Classifier-Based method) using a 10-fold cross-validation on our data and reporting precision, recall, F1 scores, accuracy, and execution times (mm:ss). F1 score is the primary result measure; accuracy and execution time are informational.
>
> The 10-fold cross-validation uses the same 10 fixed train and non-overlapping test splits for all experiments. With few parameter estimates (1 threshold each for Closest Match and Cognate-Based, 2 distance and 7 target language coefficients for Classifier), a separate *train* split into *fit/val* is not necessary. (Miller and List, 2023)

TABLE 3.2: Ten-fold cross-validation for three methods with NED (normalized edit) and SCA (Sound-Class based phonetic alignment) distance measures.

Method	Prec.	Recall	F1	Acc.	mm:ss
Closest Match					
NED	0.832	0.703	0.761	0.938	00:15
SCA	0.869	0.720	0.787	0.945	00:29
Cognate-Based					
NED	0.853	0.705	0.771	0.941	01:48
SCA	0.862	0.719	0.783	0.944	04:49
Classifier SVM (linear)					
NED, SCA	**0.931**	0.713	**0.806**	0.952	00:37

Bolded estimates are superior to and underlined estimates inferior to the the overall average using analysis of means (Nelson, Wludyka, and Copeland, 2005) with joint error rate $\alpha = 0.05$. Adapted and updated from Miller and List (2023).

We also reported out results from the experiments in Miller and List (2023):

> All methods perform well with less than 5 points separating the highest from the lowest F1 scores. Tab. 3.2 shows the results of the ten-fold cross validation of our three methods in five experiments. An analysis of variance,[1] with experiment as the effects variable and train-test split as the nuisance variable, shows highly significant effects for precision ($F_{4,36} = 25.74, p < 0.0001$) and F1 score

[1] Statistical analyses with JMP (*JMP®, Version 17.0.0* 2022).

($F_{4,36} = 14.3, p < 0.0001$).

Closest Match with normalized edit distance performs poorly, while Classifier-Based with combined normalized edit and SCA distances performs well. Classifier-Based performs better than the average of all experiments in F1 score, and substantially better in precision versus other experiments; the method is conservative, with a low number of false positives. Performance on remaining experiments is indistinguishable from the overall average of all experiments combined. The Cognate-Based method is compute intensive performing multiple alignment over all languages. Accuracy is well above the majority decision accuracy of 84.8% (100% − 15.2% borrowing) in all experiments. (Miller and List, 2023)

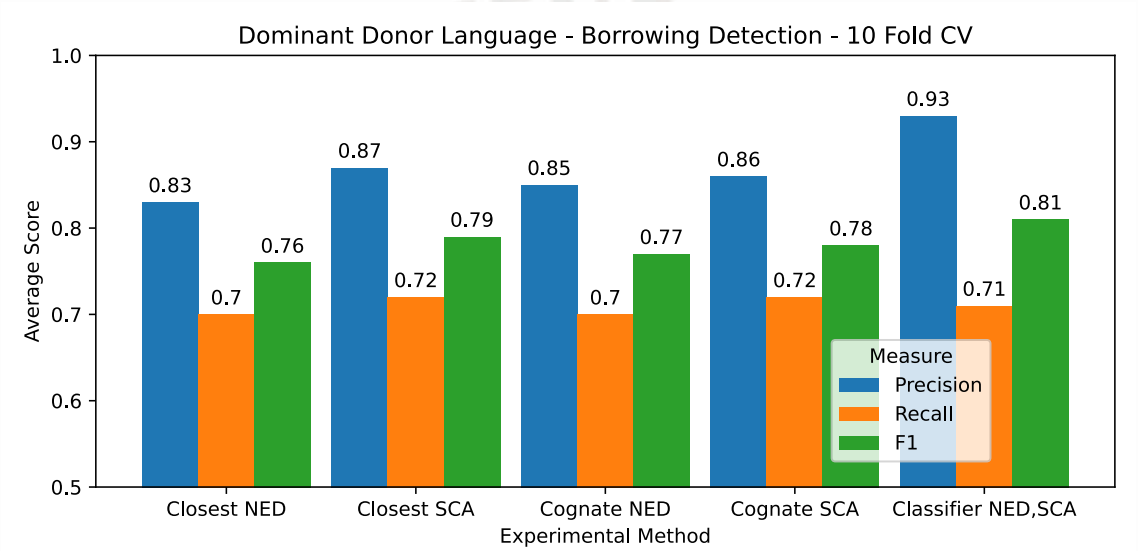


FIGURE 3.2: Results of the cross validation experiment.
Averaged for each method over all languages in our sample.
Adapted and updated from Miller and List (2023).

In Fig. 3.2 the differences in precision and F1 score stand out a bit more especially the high precision for the Classifier. The stand-along Closest SCA method looks pretty competitive, but not good enough to separate it from the Cognate-Based methods (NED and SCA) all performing near the average of all methods.

While not directly comparable, since the monolingual methods do not use a donor focused implementation or evaluation, the reported results for borrowing detection for the same set of seven languages of ≈ 0.80 (see §2.2.7, and Tab. 2.10) is little different from that seen here.

We also performed several *ad hoc* experiments in response to reviewer comments. These were summarized in (Miller and List, 2023):

> A search for classifier improvements prompted several *ad hoc* experiments (see Tab. 3.3). We observe: (1) A radial basis function (rbf) SVM

> classifier performs no better than our linear SVM. We suspect the estimated target language parameters do not generalize well to held-out data. (2) A logistic regression classifier performs on par with our linear SVM. (3) A weight balanced SVM classifier trades an increase in recall for a larger drop in precision. We also test whether using separate trials for each target language in Closest Match, would perform as well as all languages together. A combined trial performs better; a single threshold estimate appears to generalize better to held-out data than using individual language estimates. (Miller and List, 2023)

Results are similar for the classifier method with linear SVM and both NED and SCA functions. The combined analysis for all seven languages together appears slightly better than performing analyses separately.

TABLE 3.3: Ten-fold cross-validation for several *ad hoc* experiments with NED (normalized edit) and SCA (Sound-Class based phonetic alignment) distance measures.

Experiment	Prec.	Recall	F1	Acc.
Classifier variations - NED, SCA				
SVM (rbf)	0.945	0.694	0.799	0.951
Logistic regression	0.914	0.728	0.809	0.952
SVM (balanced)	0.613	0.826	0.704	0.902
Means over languages analyzed separately				
Closest match - SCA	0.860	0.707	0.770	0.941
Classifier (linear SVM) - NED, SCA	0.936	0.697	0.793	0.949
Previous analyses altogether				
Closest match - SCA	0.869	0.720	0.787	0.945
Classifier (linear SVM) - NED, SCA	0.931	0.713	0.806	0.952

Classifier experiments: SVM with radial basis function, Logistic regression, linear SVM with balanced class weights. Analyses for method by each language separately. Adapted and updated from Miller and List (2023).

We were also curious as to whether there was variation in borrowing detection by individual target language. Since highly parameterized language models are not used in multilingual methods, data paucity versus the number of estimated parameters should not come into play. Here are the results by target language for Classifier borrowing detection as we reported in Miller and List (2023):

> Tab. 3.4 shows the results of the Classifier method for the seven target languages in our sample with training and evaluation over the entire dataset. There is some variation in performance by language, in particular, with recall in $[0.615, 0.778]$. We detect a linear relation between the performance and the amount of borrowings from the dominant language in the target languages. (Precision: $r = -0.39$, NS; Recall: $r = 0.88, p < 0.01$; F1 score: $r = 0.85, p < .01$; 1-sided Pearson correlation tests with $df = 5$). Recall and F1 scores improve as borrowing increases. This is a curious finding given that there are only nine

parameters estimated over the entire dataset (one for each target language, and one each for NED and SCA functions). This correlation could be an artifact of higher borrowing resulting in better estimation of a target language coefficient, or more interestingly, a cultural process where more dominant-donor borrowing corresponds to reduced phonetic adaption into the target language. (Miller and List, 2023)

Table 3.4: By language results for the Classifier borrowing detection methods on the seven target languages in our sample.

Language	Prec.	Recall	F1	Acc.	Borr.
Imb. Quechua	0.921	0.773	0.841	0.924	26%
Mapudungun	0.944	0.716	0.814	0.950	15%
Otomi	0.932	0.692	0.794	0.968	9%
Q'eqchi'	0.934	0.615	0.742	0.961	9%
Wichí	0.952	0.658	0.778	0.953	12%
Yaqui	0.938	0.778	0.851	0.941	22%
Zin. Tzotzil	0.932	0.661	0.773	0.949	13%
Average	0.934	0.714	0.810	0.952	15%

Last column shows the proportion of Spanish borrowings.
Adapted from Miller and List (2023).

3.2.2 Error analysis

While pleased with the F1 score average of 0.81, we wanted see how we could improve borrowing detection even more, with the hope to achieve F1 scores exceeding 0.90. Such would make our methods truly useful to historical linguists and perhaps apt for inclusion in comparative method and phylogenetic modeling automated workflows.

So we performed a detail analysis of errors from the classifier on the combined database, and reported the results in Miller and List (2023):

> To get a better understanding about the different types of errors that our best performing experimental combination commits, we conducted a detailed error analysis from the Classifier-Based borrowing detection results. A spreadsheet snippet (Fig. 3.3), serves as a reference for several error types.
>
> For undetected borrowings (false negatives), we identified four error types: (1) cases where the borrowed form was not present in the donor wordlist, e.g., Mapudungun peso "coin" is borrowed from Spanish *peso* "peso", but our Spanish wordlist only has *moneda*, (2) cases where the form was present in the donor wordlist, but with a different concept, e.g., Wichi anio "age" is borrowed from Spanish *año* "year", while the Spanish word for "age" is *edad*, (3) cases of large phonetic distance between donor and recipient forms, e.g., Wichi *alulis* "adobe", which is somewhat distant from Spanish *adobe*, and

ID	DOCULECT	TOKENS	DONOR LANGUAGE	DONOR VALUE	DET STATUS
▼ ABSTAIN FROM FOOD					
	Spanish	a ʝ u n a ɾ			
6227	Qeqchi	a j uː n i n k + ɾ i ʃ	Spanish	ayunar	fn
▼ ADOBE					
	Spanish	a ð o β e			
8988	ImbaburaQuechua	a d u b i	Spanish	adobe	fn
10182	Wichi	a l u l i s	Spanish	adobe	fn
▼ AGE					
	Spanish	e ð a ð			
10531	Wichi	a n i o	Spanish	año	fn
▼ ANIMAL					
	Spanish	a n i m a l			
3351	ZinacantanTzotzil	tʃʰ a n u l i l			fp

FIGURE 3.3: Example collection of detection errors.
From Miller and List (2023).

> (4) cases of unrecognized partial borrowing, e.g., Qeqchi aiunink-riʃ "abstain from food", which is partially borrowed from Spanish *ajunar* "fast, abstain from food".
>
> For falsely detected borrowings (false positives), we identified three error types: (1) cases where the form was not borrowed from the dominant language but *vice versa*, e.g., Spanish *poroto* "bean" was borrowed *from* Quechua purutu, (2) cases of chance similarities between word forms, e.g., Spanish *animal* "animal" and Zinacantan Tzotzil tʃanulil, and (3) cases so improbably similar that we suspect errors in the original annotation, e.g., Spanish *pelota* "ball" and Wichi *pelutaj*. (Miller and List, 2023)

For 115 concepts with errors from 490 sampled concepts, there were 139 undetected (false negative) and 26 falsely detected (false positive) lexical borrowings. Note that only 37.5% of all concepts were sampled, 490 out of 1,308, and that only 23.5% of sampled concepts had errors. Note also that several concepts had multiple target language errors – so 165 total errors over 115 concepts with errors. Tab. 3.5 reports both the details and this summary.

The great majority of errors were in recall, and many of these borrowings were from lexemes **not** within the same concept (75) or **not** on the dominant donor wordlist (28). Note also that the not within same concept category, did not discern whether the form was available in another concept, so an undetermined amount of not same concept errors (semantic shift) might also be not in donor wordlist errors. There were also many recall errors (31) due to the classifier not matching on the donor lexeme (large phonetic distance). For the few falsely detected borrowings, most were due to the chance similarity of forms (10), or likely dataset error (9). Importantly, only 7 errors overall were because the *borrowing direction was not* from the dominant language.

TABLE 3.5: Summary of undetected (false negative) and falsely detected (false positive) borrowings over 115 concepts with detection errors from 490 sampled concepts.

Undetected Borrowings		
Error Type	Count	Pct
borrowed form not in donor list	28	17
different concept than recipient form	75	45
large phonetic distance	31	19
partial borrowing as only reason	5	3
Subtotal	139	84
Falsely Detected as Borrowings		
Error Type	Count	Pct
direction not from dominant donor	7	4
chance similarity of form	10	6
likely dataset error	9	5
Subtotal	26	16
Total	165	100

Adapted from Miller and List (2023).

3.2.3 Incorporating competing cross-entropies

A way to address the problem (see Tab. 3.5 of error diagnostics) of not matching on the donor lexeme, might be to add another method of matching to the classifier meta-method.

The least cross-entropy (LCE) method was introduced in §3.1.2 as a port of our previous Markov chain method of the competing cross-entropies for use in this multilingual wordlist and dominant language donor context. Here we prove out LCE both as a stand alone method and as a function used by the classifier meta-method. The already well performing classifier method with normalized edit (NED) and sound-class based phonetic alignment (SCA) distance functions now incorporates the least cross-entropy function (LCE) as well.

Results for the new experiments with LCE as well as a few previous experiments (for easy reference) are reported in Tab. 3.6. An analysis of variance, with experiment/method as the effects variable and train-test split as the nuisance variable, shows highly significant effects for precision ($F_{4,36} = 32.49, p < 0.0001$), recall ($F_{4,36} = 41.90, p < 0.0001$), and F1 score ($F_{4,36} = 30.13, p < 0.0001$). Comparisons between the overall average and individual experiments/methods, are shown for each measure at a joint error rate of 5%. Of particular note is the superior performance of the classifier with NED, SCA, and LCE functions on recall and F1, and for the classifier with just NED and SCA on precision.

Detail results by individual language and method are reported in appendix Table B.1.

TABLE 3.6: Ten-fold cross-validation for experiments with Least Cross-Entropy (LCE) and other methods.

Method	Prec.	Recall	F1	Acc.	mm:ss
Closest Match previous analyses					
- NED	0.832	0.703	0.761	0.938	00:15
- SCA	0.869	0.720	0.787	0.945	00:29
Least Cross Entropy					
- LCE	0.795	0.739	0.765	0.936	02:46
Classifier linear SVM					
- NED, SCA (previous)	**0.931**	0.713	0.806	0.952	00:37
- NED, SCA, LCE	0.871	**0.827**	**0.848**	0.958	04:34

There were $\approx 25,000$ non-zero count parameters overall LCE language models, i.e., $\approx 1,750$ per language model. **Bolded** estimates are superior to and underlined estimates inferior to the the overall average using analysis of means (Nelson, Wludyka, and Copeland, 2005) with joint error rate $\alpha = 0.05$. Format adopted from Miller and List (2023).

Fig. 3.4 presents the tabular results grouped by performance measure – this better shows the contrast between experimental methods on each measure. Note: just as for Tab. 3.6, that previous results without LCE are shown for contrast.

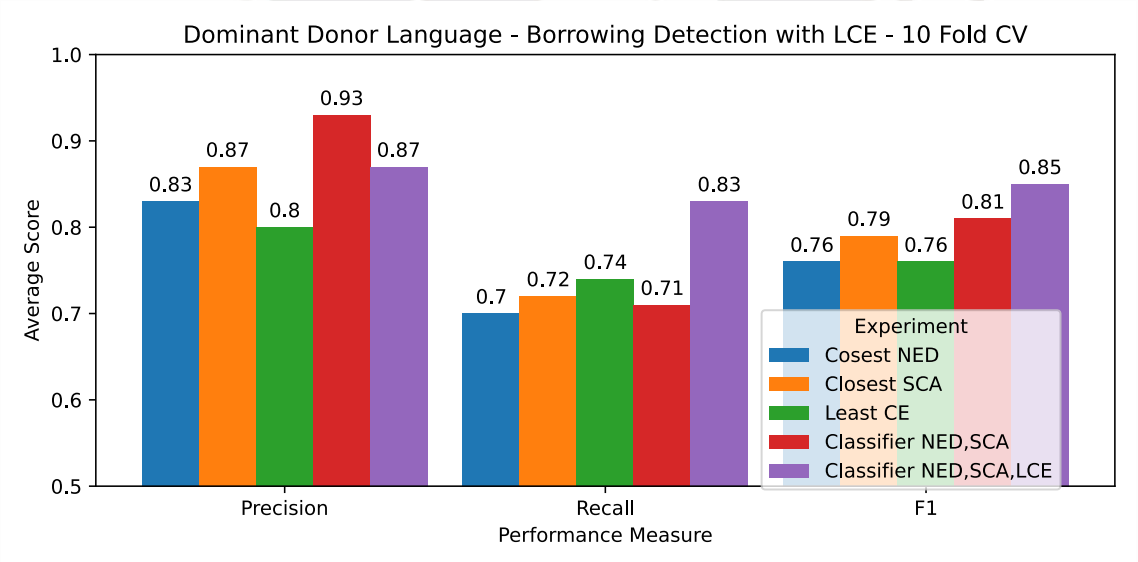


FIGURE 3.4: Results of the cross validation experiment.
Averaged for each method over all languages in our sample.
Format adopted from Miller and List (2023).

There is an impressive increase in recall and recall and F1 score for the classifier meta-method with NED, SCA, and LCE functions versus the classifier with just NED and SCA functions. While precision dropped from the previous high level, it is still well above our acceptable limit of 0.8. Stand-alone least cross-entropy (LCE) merely performs on par with the closest match NED. Again we find that the combination of diverse useful methods complement each other to give overall better borrowing detection performance than any stand alone method.

The execution time for performing the cross-validation with incorporated least cross-entropy method now measures in minutes instead of just seconds. This is due to the fitting of the 3rd order Markov chain language models for dominant donor and not dominant donor borrowed word models for each target language.

Risk of overfitting with least cross-entropy. With traditional multilingual sequence models with few parameters, there was little difference between borrowing detection on training and test datasets. With LCE either stand-alone or incorporated into the classifier, there is risk of substantial overfitting, and so it becomes essential to report results from test datasets in order to adequately represent expected borrowing detection performance. Results for borrowing detection methods are reported for the initial fold of our 10-fold cross-validation train and test datasets for comparable cases with and without LCE in Tab. 3.7. Note the large difference in number of estimated parameters and the substantial variance between train and test F1 scores for methods with LCE.

TABLE 3.7: Difference between train and test F1 scores depending on use of Lowest Cross-Entropy method. Trials made on fold 00 from a 10-fold cross-validation train and test split.

Method	Parameters	Train F1	Test F1
Closest match - SCA	1	0.790	0.788
Least cross-entropy	$\approx 25,000$	0.939	0.776
Classifier - linear SVM - NED, SCA	9	0.811	0.809
Classifier - linear SVM - NED, SCA, LCE	$\approx 25,000$	0.959	0.834

Smoothing parameter for Markov chain. The previous Markov chain language model for monolingual borrowing detection (§2.1.1), used Kneser-Ney smoothing with a smoothing value of 0.3. This was determined in an *ad hoc* manner based on early screening trials before the more rigorous comparison of Markov chain, neural network, and bag of words methods. For this reimplementation of Markov chain borrowing detection as the least cross-entropy method and function for incorporation into the classifier meta-method, we conducted experiments on just the training partitions of the database to determine an optimal setting for the smoothing parameter. Kneser-Ney smoothing of 0.9 is optimal for this application based on the results with the highest F1 score in Tab. 3.8. This results in improved recall and approximately the same precision versus the 0.3 smoothing used previously.

3.2.4 Augment donor wordlist coverage

A way to address the problem (see Tab. 3.5 of error diagnostics) of not having the borrowed lexeme in the donor wordlist, would be to augment the donor wordlist so that it is more likely to include borrowed forms.

Given the set of seven target languages plus Spanish in the SABor database with borrowing already annotated, it would be a straight forward task to augment the

TABLE 3.8: Borrowing detection and Kneser-Ney smoothing parameter for Least Cross-Entropy method

Method	precision	recall	F1 score
0.1	0.781	0.667	0.718
0.3	0.793	0.695	0.740
0.5	0.802	0.692	0.742
0.7	**0.821**	0.691	0.748
0.9	0.790	**0.730**	**0.758**

Spanish wordlist with all the missing lexemes from the Spanish donor wordlist. However, this would not serve to develop and prove out methods for borrowing detection to be used by historical linguists when wordlists have not yet been annotated for borrowing. In the *real life* of historical linguists, there are no pre-existing *oracles* to say what donor language words/lexemes have been borrowed.

We investigate how difficult a problem it would be to improve borrowing detection from a dominant donor, by comparing a few pre-existing Spanish wordlists with a list of borrowed words in the SABor database. The critical measure in this case is the coverage of known borrowed words by the candidate Spanish wordlists, where better coverage should translate into better borrowing detection. **Beware.** We do not consider whether the donor word is for the same concept as the target. So coverage results will provide an optimistic view of what coverage is possible.

Materials and Methods for Donor Wordlist Coverage Experiment

Materials. Specific datasets used in this experiment are:

1. Currently defined SABor database of seven indigenous Latin American languages and their dominant language donor (Spanish). Key fields from language forms include: ID, Language_ID, Parameter_ID (concept), Form (orthographic form), Segments (sounds), Borrowed (True/False), Borrowed_Score (0, .25, .5, .75, 1.0), Donor_Language ('Spanish'), and Donor_Value (orthographic form).
2. IDS database including the Spanish wordlist used as the base for the Spanish wordlist included in SABor (Key and Comrie, 2015).
3. Concepticon database of concepts including language forms in several major languages (List et al., 2022a).
4. Big dictionary of the 8,600 most frequently occurring written words from a large sample of Spanish text (Neri, 2018).

Methods. The workflow process to describe wordlist coverage follows:

1. Get the wordlist of all borrowings from Spanish into the seven target languages for the SABor database defined by this study. This list consists of all Donor_Value where the Donor_Language is Spanish.
2. Get the Spanish source wordlists of Forms (orthographic forms) for these cases: (a) Extract Spanish wordlist from the IDS database - both Forms (orthographic forms) and Parameter_IDs (concepts). (b) Extract Spanish wordlist from Concepticon - both forms and concepts. Concepticon is likely to provide good coverage for wordlists such as that derived from IDS or WOLD. (c) Extract Spanish wordlist from a big freely available dictionary of forms - just forms as the dictionary is not indexed by concept.
3. Match words between the wordlist of all borrowings from Spanish and each case of Spanish source wordlists, as well as the union of IDS and Concepticon Spanish source, and the union of all Spanish source wordlists cases. Express the matches as a percent coverage of the wordlist of all borrowings from Spanish.
4. After stemming all wordlists using the NLTK Spanish stemmer, report the match between wordlist of all borrowings from Spanish and the Spanish source wordlists. Since Spanish borrowings and Spanish source wordlists may use different word form conventions reflected in Spanish word inflections, stemming should remove such differences. Likewise, multilingual matching methods, would assign more value for matching on word stems than suffixes, and so stemming should be more representative of potential word matching and borrowing detection potential.

Results

Coverage of borrowed forms from Spanish in the seven target languages of the SABor database is shown in Tab. 3.9 for original orthographic forms and for stemmed orthographic forms. A graph shows the distinction between wordlist sources and effect of stemming even more clearly in Fig. 3.5. The IDS Spanish wordlist source is the basis for the Spanish wordlist source used in SABor and so percentage coverage here should be similar to coverage that available to borrowing detection methods reported previously in this chapter. As mentioned above, no consideration is given here for whether the form is used with the same concept shared between target and donor language, so coverage here is more a best case scenario.

In all cases stemmed forms show a several point advantage in coverage versus original forms. This should be reflected in detection methods that more effectively ignore less important differences between source and target words, such as terminations for a right-hand side inflected language such as Spanish. Looking at the IDS Spanish source, the difference between original and stemmed coverage is like the difference between closest NED (normalized edit distance) versus classifier NED, SCA performance. The Concepticon Spanish source also adds several points advantage over the baseline IDS source. However, the Big (8,600) Spanish source performs poorly. Combining IDS with Concepticon source adds

TABLE 3.9: Donor wordlist coverage of orthographic forms for Spanish borrowings (1,480) from target languages.

Wordlist source	Size	Match	No match	% coverage
Original orthographic forms				
IDS Spanish	1,683	1,112	368	75
Concepticon Spanish	2,620	1,244	236	84
Big (8,600) Spanish	8,645	990	490	67
IDS, Concepticon	2,832	1,253	227	85
Altogether	10,108	1,313	167	89
Stemmed orthographic forms				
IDS Spanish	1,571	1,193	287	81
Concepticon Spanish	2,335	1,317	163	89
Big (8,600) Spanish	4,114	1,094	386	74
IDS, Concepticon	2,529	1,327	153	90
Altogether	5,319	1,372	108	93

little to Concepticon source alone, but the union of all three sources produces a several point advantage over the the Concepticon source alone or IDS and Concepticon. Even so, still ignoring the issue of same concept requirement, the maximum coverage attained is 93%.

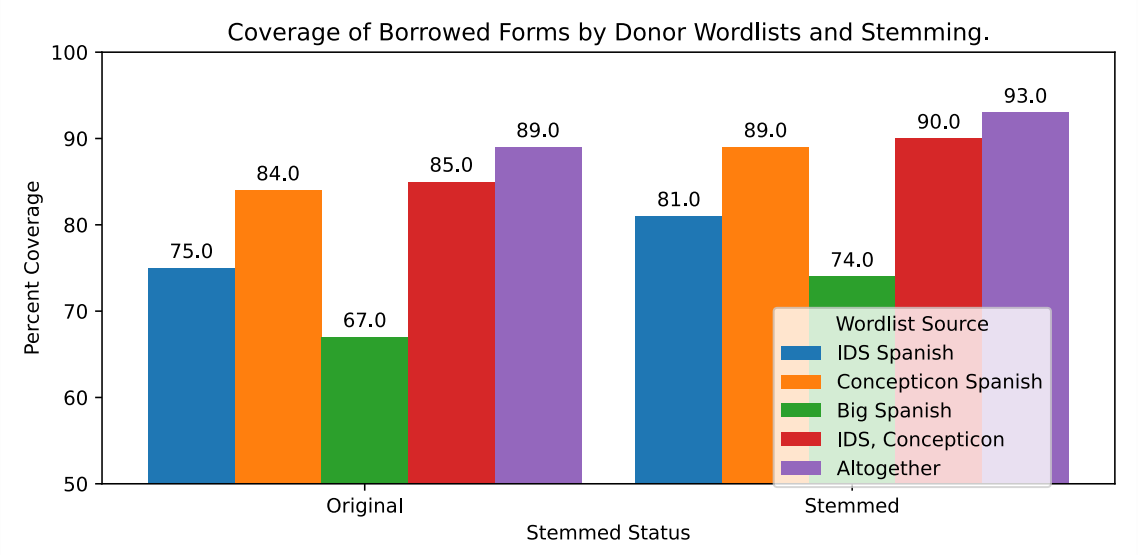


FIGURE 3.5: Results of augment donor wordlist coverage experiment.

3.2.5 Relax same concept restriction

A way to address the problem (see Tab. 3.5 of error diagnostics) of the borrowed (recipient) lexeme not coming from the same concept as the donor lexeme, would be to relax the requirement that borrowed and donated lexeme must be from the same concept.

We investigate a simplified version of this problem using an experimental version of the closest match method with the sound-class based phonetic alignment (SCA) measure. This experimental version supports the following variants of a concept restriction between borrowed and donor lexeme: 1. same concept restriction (current standard), 2. major concept restriction, where major concept is available from Concepticon, 3. no concept restriction at all, 4. same concept restriction as priority with fallback to major concept restriction. The critical measure of success in this experiment is borrowing detection performance as measured by F1 score. Corresponding recall and precision are also reported as well as execution time.

Materials and Methods

In addition to the currently defined SABor database of seven indigenous Latin American languages and their dominant language donor (Spanish), the "Rzymski-2020-1624" concept list from the Concepticon database (List et al., 2022a) is used to implement a more abstract (higher level) *major* concept restriction. This extra concept list provides a mapping between standard Concepticon concepts already available in SABor and their major concept categories.

The existing closest match method (subcommand *closest*) is used to define an experimental *closest_exp* subcommand which implements the concept restriction variants enumerated above. Besides implementing the concept restrictions, the report out for *closest_exp* calculates borrowing detection performance at multiple threshold values in order to determine the optimum for each concept restriction variant. So optimum threshold is another response variable from this experiment.

Results

Since only one parameter is being estimated from this experiment for each concept restriction variant, there should be little problem of reproducibility of results and so all data is used in training with results reported from training. Experiments were run for each concept restriction variant, with systematic sampling from thresholds between 0.05 and 0.5 with step-size of 0.05. Borrowing detection results for the optimum threshold are reported for each method in Tab. 3.10 and shown graphically in Fig. 3.6.

Table 3.10: Borrowing detection by concept restriction – threshold selected to optimize F1 score.

Concept Restriction	Prec.	Recall	F1	Acc.	Threshold	Time
Target (default)	0.872	0.721	0.789	0.945	0.40	0:10
Central	0.848	0.676	0.752	0.936	0.3	0:25
No restriction	0.358	0.582	0.443	0.791	0.15	35:00
Target-Central	0.866	0.741	0.799	0.947	0.35	0:25

No concept restriction results in poor borrowing detection, poor accuracy, and a prolonged execution time. When a lexeme on the target wordlist can match with any lexeme on a corresponding donor wordlist, without concept restriction, there will often be closer matches found than for the same concept. The execution time is increased by a significant fraction of the donor wordlist size, since without a concept restriction, each donor lexeme has to be checked for each target word.

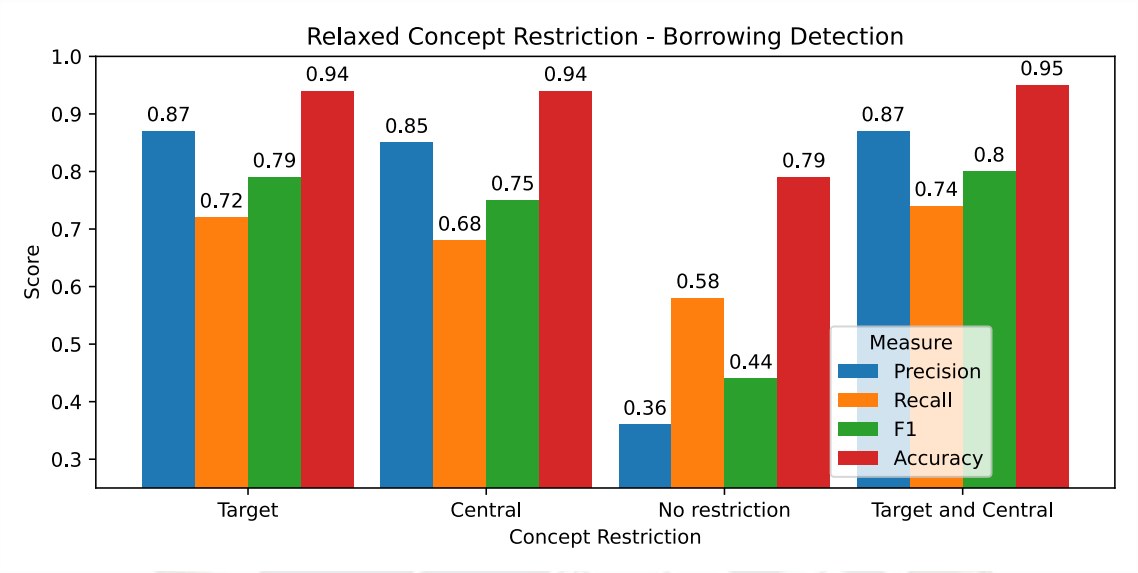


FIGURE 3.6: Results of relaxed concept requirement experiment.

Even limited to just the central concept, there will be some lexemes that are a closer match than lexemes for the same concept. Interestingly, the recall for the central concept restriction is less than with the same concept restriction. The restriction to same concept doesn't seem so bad after all. And the combination of restriction to target concept with fallback to central concept when no match is found, might offer a path for improvement.

3.3 Report out to historical linguist

Something applied this way comes...

In all the talk of wordlists, borrowing detection methods and results, experiments and cross-validations, we had almost forgotten to show what a useful report out of borrowing detection looks like.

Each of the multilingual methods, e.g., classifier, produces a brief report of train and test datasets, as well as creates a tab delimited rectangular file organized by concept and language of input forms and predictions of borrowing, including the donor language, and reference to the likely donor form. A subsequent evaluation step, for when borrowings have already been annotated, as is the case here with SABor, reports detection error status and the donor language and form. Here only show the donated form since Spanish is the dominant donor

and hence the only donor language to report. Detection error status is coded as: fn=false negative (undetected borrowing), tn=true negative, fp=false positive (incorrectly predicted borrowing), and tp=true positive (correctly predicted borrowing).

Various example snippets from the evaluation file are given for the best case `classifier NED, SCA, LCE` prediction and followup evaluation in the following.

ID	LANGUAGE	FAMILY	CONCEPT	FORM	IPA	PR_LANG	PR_ID	DET	DONATED
▼ BED									
718	Spanish	Indo-European	BED	cama	kama				
2386	Yaqui	Uto-Aztecan	BED	kaamam	kaːmam			fn	cama
3773	ZinacantanTzotzil	Mayan	BED	tem	tem			tn	
3774	ZinacantanTzotzil	Mayan	BED	vayebal	vajebal			tn	
5171	Qeqchi	Mayan	BED	ch'aat	tʃ'aːtʰ			tn	
5172	Qeqchi	Mayan	BED	warib'	warib			tn	
7063	Otomi	Otomanguean	BED	nt'ots'i	nt̰'ots'ĭ			tn	
7064	Otomi	Otomanguean	BED	xifi	xifi			tn	
8970	ImbaburaQuechua	Quechuan	BED	kama	kama	Spanish	718	tp	cama
10165	Wichi	Matacoan	BED	tomow'et	tomowʔet			tn	
11434	Mapudungun	Araucanian	BED	kawitu	kawitu	Spanish	718	fp	kawitu
▼ BOOK									
1588	Spanish	Indo-European	BOOK	libro	liβro				
3074	Yaqui	Uto-Aztecan	BOOK	liprom	liprom	Spanish	1588	tp	libro
4378	ZinacantanTzotzil	Mayan	BOOK	vun	vun			tn	
4379	ZinacantanTzotzil	Mayan	BOOK	livro	livro	Spanish	1588	tp	libro
6082	Qeqchi	Mayan	BOOK	hu	hu			tn	
6083	Qeqchi	Mayan	BOOK	liib'r	liːbr	Spanish	1588	tp	libro
8235	Otomi	Otomanguean	BOOK	hę̱'mi	hæʔmi			tn	
9542	ImbaburaQuechua	Quechuan	BOOK	kamu	kamu	Spanish	1588	fp	
10761	Wichi	Matacoan	BOOK	liwulu	liwulu	Spanish	1588	tp	libro
12028	Mapudungun	Araucanian	BOOK	lifru	lifɻu	Spanish	1588	tp	libro

FIGURE 3.7: Snippet of borrowing report output.
Spreadsheet edited to remove unnecessary columns and provide highlights. Green columns are predicted donor language and form ID in the report. Blue columns are output from the evaluation only when borrowings are already annotated.

In the example of Fig. 3.7, the report is organized by concept with language forms and borrowing predictions reported for each concept. Most language forms are not borrowed and so are not predicted to have a language donor or corresponding form ID. These show a "tn" detection status.

The concepts of *BED* and *BOOK* were selected to show off other detection statuses as well. Imbabura Quechua borrowed the word "cama" from Spanish. This is correctly predicted with language and ID (Spanish, 718). The ID points to the row of the form for the same concept, as seen here. Mapudungun borrowed the word "kawitu" from Quechua. But it is incorrectly predicted as borrowed from Spanish "cama", so this has a "fp" detection status. The word "libro" was borrowed by most of the indigenous languages in the SABor database, and correctly

predicted so. However "kamu" was incorrectly predicted as a borrowing from Spanish, when indeed it seems an innovation of the Imbabura Quechua.

If we were working as historical linguists and using the classifier to show likely words borrowed from a dominant language, then we would of course not have detection and donated word outcomes available as with the evaluation report. Instead we would work with the prediction report from the classifier. To focus on just the borrowing predictions in this case we could filter on the predicted donor language, PR_LANG in this example.

Since we are working with an evaluation report, we can choose to ignore the less interesting detection status of "tn", words not borrowed and correctly predicted as not borrowed. To do this we filter the evaluation spreadsheet on detection status (DET) other than "tn". We present several example concepts with explanations as appropriate.

ID	LANGUAGE	FAMILY	CONCEPT	FORM	IPA	PR_LANG	PR_ID	DET	DONATED
▼ COOKHOUSE									
699	Spanish	Indo-European	COOKHOUSE	cocina	kosina				
2370	Yaqui	Uto-Aztecan	COOKHOUSE	kosina	kosina	Spanish	699	tp	cocina
3762	ZinacantanTzotzil	Mayan	COOKHOUSE	kusina	kusina	Spanish	699	tp	cocina
7046	Otomi	Otomanguean	COOKHOUSE	nthokụkomida	nθokukomida			fn	comida
ID	LANGUAGE	FAMILY	CONCEPT	FORM	IPA	PR_LANG	PR_ID	DET	DONATED
▼ CUSTOM									
1644	Spanish	Indo-European	CUSTOM	costumbre	kostumbre				
1645	Spanish	Indo-European	CUSTOM	hábito	aβito				
3108	Yaqui	Uto-Aztecan	CUSTOM	kojtumrem	kojtumrem			fn	costumbre
9570	ImbaburaQuechua	Quechuan	CUSTOM	koshtumbri	koʃtumbri	Spanish	1644	tp	costumbre
ID	LANGUAGE	FAMILY	CONCEPT	FORM	IPA	PR_LANG	PR_ID	DET	DONATED
▼ DRUM									
1591	Spanish	Indo-European	DRUM	bombo	bombo				
1592	Spanish	Indo-European	DRUM	tambor	tambor				
4381	ZinacantanTzotzil	Mayan	DRUM	tampol	tampol			fn	tambor
6085	Qeqchi	Mayan	DRUM	tamb'or	tambor	Spanish	1591	tp	tambor
6086	Qeqchi	Mayan	DRUM	moor	moːr	Spanish	1591	fp	
8240	Otomi	Otomanguean	DRUM	tambo	tambo	Spanish	1591	tp	tambor
9544	ImbaburaQuechua	Quechuan	DRUM	tambur	tambur	Spanish	1591	tp	tambor

FIGURE 3.8: Snippets of borrowing report output for concepts *cookhouse*, *custom*, and *drum*. Snippets show expected detection behavior of the classifier.

The snippets of Fig. 3.8 show the expected behavior of the classifier with NED, SCA, LCE functions. For the concept *COOKHOUSE* Yaqui and Zinacantan Tzotzil languages borrow "cocina" from Spanish, and the borrowing is correctly predicted with language and ID (Spanish, 699). However, Otomi borrows "comida" which is not present in the Spanish wordlist for this concept. Even if it were, it's not clear that the classifier would have identified "comida" from the composite Otomi word. For the concept *CUSTOM*, the classifier recognizes the borrowing of "costumbre" into Imbabura Quechua, but not into Yaqui. For the concept *DRUM*, the classifier recognizes the borrowing of "tambor" into Qeqchi,

Otomi, and Imbabura Quechua, but not into Zinacantan Tzotzil. Also, "moor" seems problematic, it is not identified as donated in the wordlist, but the classifier detects it as a borrowing.

ID	LANGUAGE	FAMILY	CONCEPT	FORM	IPA	PR_LANG	PR_ID	DET	DONATED
▼ COUNT									
1235	Spanish	Indo-European	COUNT	contar	kontar				
4129	ZinacantanTzotzil	Mayan	COUNT	pas_kwenta	pʰas+kʷenta	Spanish	1235	fp	
10512	Wichi	Matacoan	COUNT	kunta	kunta	Spanish	1235	tp	contá
ID	LANGUAGE	FAMILY	CONCEPT	FORM	IPA	PR_LANG	PR_ID	DET	DONATED
▼ CHEAP									
1105	Spanish	Indo-European	CHEAP	barato	barato				
9206	ImbaburaQuechua	Quechuan	CHEAP	baratu	baratu	Spanish	1105	tp	barato
11684	Mapudungun	Araucanian	CHEAP	pichi_falin	pitʃi+falin	Spanish	1105	tp	valer
11685	Mapudungun	Araucanian	CHEAP	faratu	faɻatu	Spanish	1105	tp	barato
ID	LANGUAGE	FAMILY	CONCEPT	FORM	IPA	PR_LANG	PR_ID	DET	DONATED
▼ DOORPOST									
704	Spanish	Indo-European	DOORPOST	jamba_de_puerta	xamba+ðe+pwɛrta				
705	Spanish	Indo-European	DOORPOST	larguero	laryɛro				
2375	Yaqui	Uto-Aztecan	DOORPOST	marko	marko	Spanish	704	tp	marco
5155	Qeqchi	Mayan	DOORPOST	champa	tʃampa			fn	jamba
7052	Otomi	Otomanguean	DOORPOST	poste	poste	Spanish	704	tp	poste

FIGURE 3.9: Snippets of borrowing report output for concepts *count*, *cheap*, and *doorpost*. Snippets show more problematic detection behavior of the classifier.

The snippets of Fig. 3.9 show a somewhat problematic behavior of the classifier with NED, SCA, LCE functions, which we hope to resolve in subsequent work. For the concept *COUNT* Wichi borrows "contar" from the Spanish and this is correctly predicted. However, Zinacantan Totzil borrows "cuenta" from the Spanish, and this is not recognized, both because "cuenta" is not included as a word for *COUNT* and also because because it is a partial borrowing. For the concept *CHEAP*, Imbabura Quechua and Mapudungun borrow "barato" from Spanish and this is correctly predicted. Interestingly, "pichi falin" is also correctly recognized as a borrowing of from Spanish "valer" although misidentified as from "barato". This seems the effect of the LCE function which can identify borrowings based on lexical cross-entropies, such as for "pichi falin" which obeys Spanish phonology and phonotactics. For the concept *DOORPOST*, "macro" and "poste" are both correctly detected as borrowings from Spanish, even though they don't correspond at all to the Spanish "jamba de puerta". Again this is an effect of the LCE function based on lexical cross-entropies. However, the borrowing of "jamba" is not recognized likely because it is just a partial borrowing.

These examples show how the report out of the classifier or similar command of the SABor database could be very useful in computer assisted detection of borrowings, or in the case of already annotated borrowings, for the evaluation of borrowing detection. Moreover, while the examples shown are anecdotal, they suggest further work that could be done to improve our borrowing detection methods. A more complete error assessment as performed previously would

give a better insight into the gains and losses made by including the LCE function in the classifier.

3.4 Discussion

How well did we do automatically detecting borrowings from dominant languages based on wordlist data?

3.4.1 Initial effort

In our multilingual only approach, with implementation and evaluation focused on a dominant donor language, we devised three general methods to detect borrowed words from dominant languages, closest match and cognate-based multilingual sequence comparisons, and a classifier meta-model. The classifier meta-model showed the best performance, with F1 score of 0.81, and high precision of 0.93. This method in its current state could already prove very useful in computer-assisted workflows, at least for the case of high contact language events with a dominant language donor.

Borrowing detection results are similar to those obtained using the neural network competing cross-entropies method from our monolingual borrowing detection chapter §2. However, the methods are not directly comparable, since the monolingual method did not use a dominant donor focused implementation or evaluation. Nevertheless this suggests that each approach and method is useful in its own right. It might be useful to offer a donor focused implementation for monolingual methods as well. Or general borrowing detection in multilingual methods in addition our donor focused efforts.

Since in our *ad hoc* experiments we saw little difference between our default linear SVM classifier versus a radial basis function SVM classifier or logistic regression classifier, we could continue with anyone of these classifier methods. The logistic regression classifier might be preferable in a neural network implementation, opening up the possibility to add a fully connected hidden layer between inputs and logistic regression. Such configuration might learn an improved predictor for dominant donor borrowing. Since *balanced* training data, between borrowed or not borrowed from dominant donor categories, gave poor results on test data, we continue with unweighted data in general.

We observed a positive correlation between dominant donor borrowing percentage and borrowing detection performance over languages. With data hungry methods such as Markov chains and neural networks used in monolingual borrowing detection, an obvious explanation for improved detection would be the better fit of models resulting from increased data. But here, only the threshold parameter is being estimated for closest match and cognate based methods, and just nine factor coefficients for the classifier meta-model. These multilingual methods aren't data hungry. There is an abundance of training data versus the few parameters being estimated.

So a more likely explanation, or at least plausible explanation, is that there is some other variable related to the percentage of borrowing that better explains why the borrowing detection improves with the percentage of borrowing. A few plausible and untested hypotheses are: 1. With higher percentage of borrowing, there is lesser adaption of borrowed words to the target language. 2. Higher percentage of borrowing corresponds to borrowing over a shorter time period, and so borrowings are more consistently and similarly adapted. These possibilities might even apply to the monolingual case, where it was so much easier to attribute borrowing percentage and borrowing detection to a machine learning with insufficient data problem.

3.4.2 Error analysis

Our investigation of detection errors for the multilingual dominant donor approach showed several opportunities for improvement. In particular, there seems substantial potential for improvements that account for borrowing accompanied by *semantic shift*, i.e., where the donor and target language concepts are not the same. Designing such methods is not trivial, since an unconstrained comparison of word forms independent of meaning also dramatically increases the number of falsely detected borrowings (see §3.2.5). Also important were the lack of borrowed word forms (lexemes) in the dominant donor wordlist, and the failure of even the classifier to detect the similarity of some donor forms that were present in the same concept as the target language form.

3.4.3 Improvements based on error analysis

Based on the error analysis, we explored these opportunities:

1. Fit language models to wordlists and add word cross-entropy to the classifier. This adds a complementary borrowing detection that is independent of concept and individual donor form.
2. Augment donor wordlist for increased coverage of possible forms.
3. Relax "same" to "similar" concept restriction for matching target and donor language forms.

Least cross-entropy function. After adding least cross-entropy function results, i.e., competing cross-entropies based on Markov chain language models, to the classifier inputs, dominant donor borrowing detection improved dramatically (4 percentage points). This shows that monolingual functions are not only equally powerful, but also complementary with multilingual functions! While there was substantial overfitting using least cross-entropy (LCE), due to the extreme increase in estimated parameters, borrowing detection performance on test (held-out) data is clearly superior when LCE is used in the classifier metamethod along with normalized edit distance (NED) and sound-class based phonetic alignment (SCA) distance.

The least cross-entropy method, only considers the language models and calculated cross-entropies, but is not matching on specific forms or concepts. So it serves as a valuable complement to sequence matching methods in detecting borrowing, as seen by the success of the classifier-NED,SCA,LCE.

Augment donor wordlist. Exploration of Spanish donor wordlist coverage of borrowings in the seven indigenous languages, showed what would be possible in borrowing detection, were there no issues of semantic shift nor not matching related forms (due to phonetic distance or incapable methods). Without such issues, 93% coverage with all available forms in our collection of databases should map to 93% recall with F1 score moving up or down depending on the precision attained. For similar wordlists to those used here, Concepticon source by itself might suffice to provide 89% coverage and so potentially 89% recall too. The point is, even for wordlists similar to those used here (IDS, WOLD, or less inclusive), it would take a large dictionary of forms to reach the 90% range, and this by ignoring the issues of semantic shift (not matching on concept) and not matching related forms.

This finding is based on the classifier meta-model using normalized edit distance (NED) and sound class-based phonetic alignment (SCA) distance functions. Inclusion of the least cross-entropy (LCE) method, may change that calculus as LCE does not depend on matching words across languages, but rather measuring fit of words to a language model. With LCE in the classifier, words may be correctly classified as borrowed even when there is no similar form in the donor wordlist. This is a big advantage and reduces the need to shoot for higher and higher donor vocabulary coverage.

Relax same concept restriction for match. Relaxing the same concept restriction between target borrowed word and donor word, showed using no restriction as harmful as it results in poor precision (excessive false positives) and mediocre recall. Optimal performance resulted from using the same concept restriction with fall-back to a central concept restriction if no borrowing was detected for the same concept. However there was only a single percentage point improvement, so relaxing the same concept restriction was at most marginally effective. The fall-back to a central concept restriction provides an increase in recall with a slight decrease in precision.

There is an interesting interplay here between the tightness of the concept restriction (semantics), recall, and phonetic distance, based on our experiment using closest match SCA. Loosening the restriction up to the central concept, opens up the possibly to correctly detect semantically related borrowings, but it also opens up more matching possibilities for non-borrowings. A more subtle way of relaxing the same concept restriction, such as consider donor concepts that are semantically similar to the target concept, by some measure of semantic similarity, versus the phonetic distance between forms as measured by closest match. A classifier meta-model could integrate both phonetic distance and semantic similarity measures to optimize borrowing detection.

3.4.4 Report out to historical linguist

This was the most *applied* section of this thesis in that we showed what the results of borrowing detection look like and how they could be used. There is abundant information in the prediction and evaluation reports to *connect the dots* between target and donor language lexical forms. Even when the corresponding donor lexical form is not given, whether because it is not on the donor wordlist at all or because the meaning varies from the given target concept, i.e., semantic shift, borrowings may be detected based on the incorporation of the least cross-entropy (LCE) function into the classifier. This is the big advantage of adding the LCE function.

Some problems remain: 1. Donor wordlist coverage is not complete, as noted previously. 2. Lexical borrowings with semantic shift are not recognized, as noted previously. 3. Lexical borrowings detected because of the LCE function may reference a dissimilar donor word for the same concept.

3.5 Conclusions

Classical sequence comparison measures, such as normalized edit distance (NED), or sound-class based phonetic alignment (SCA), incorporated into multilingual methods, such as *closest match* with pairwise alignment, or *cognate-based* with multiple alignment, of target and donor language sequences, offers the advantages that (1) comparisons are largely independent from one anther given a borrower concept, and (2) only a single fixed threshold is necessary to make cognate or borrowing decisions. However, they suffer the disadvantages that (1) a restrictive borrower concept needs be specified to carry out a small number matching attempts, and (2) donor language sequences need to include all possible donor words that might correspond to borrowings in order to be matched.

In contrast, sequence cross-entropy measures, such as in Markov chain or neural network language (word) models, in inherited only or competing cross-entropies methods offer the advantages that (1) comparisons are independent of borrower concept, and (2) the monolingual only language sample need only sufficiently represent the phonology and phonotactics of the borrowed and non-borrowed words of the target language. However, they suffer the disadvantages that (1) a large and phonetically representative sample of words is necessary to estimate the borrowed and non-borrowed word language models, and (2) with so many estimated parameters, borrowing detection performance needs to be demonstrated in separate test (hold-out) datasets for each language.

Our most important **discovery** in this chapter, is that the classifier meta-method, combining classical sequence comparison measures with sequence cross-entropy measures via corresponding closest match and least cross-entropy functions, produces superior borrowing detection versus without such a combination of complementary functions.

Dominant donor focused borrowing detection, an *innovation*, finesses the problem of identifying the direction of borrowing by simply attributing direction

of borrowing as from dominant donor to target language. While sometimes in error, this was a small count out of all tallied errors.

Our error analysis exposed several problems of errors resulting from the use of closest match, cognate-based, and classifier methods before adding the least cross-entropy method. We further explored some of the problems revealed by the error analysis, but more work needs to be done. In the following, we further clarify our conclusions and contributions, and discuss possible research directions to further improve borrowing detection.

Chapter 4

Conclusions and path forward

Our objective was to develop methods for automatic or semi-automatic detection of lexical borrowings from other lexical origins and apply these methods to wordlists organized by language, or by language and concept, where detection of lexical borrowings includes not only the borrowing decision itself, but also the donor language and even the likely donor form.

Detection of lexical borrowings is an essential part of and crucial for the successful application of the comparative method in historical linguistics (Campbell, 2013), where the comparative method seeks to reconstruct ancestral language and describe language relationships and events. Detection of borrowings is also crucial for phylogenetic reconstruction which seeks to identify probable language phylogenies by which a family of languages evolved to their current state (Gray, Greenhill, and Atkinson, 2013).

Common approaches to lexical borrowing detection from wordlists are:

- Monolingual borrowing detection looks for similarities and differences between words from the same wordlist based on the phonology and phonotactics of words, and designates words that are dissimilar from inherited words as borrowings. This was covered in *monolingual borrowing detection*, chapter §2.
- Multilingual borrowing detection looks for similarities between words across languages and designates similar words which cross language families as borrowed words. We dealt with the special case of dominant language donor as a means to focus the search for borrowings and designate the direction of borrowing. This was covered in *multilingual borrowing detection*, chapter §3.
- When incorrectly annotated words are used in language relationship models, e.g., using borrowed words in phylogenetic models, such words may appear very discrepant versus the model, and so are likely incorrectly annotated. We did not cover this approach in our investigation.
- Hybrid approaches correspond to how historical linguists actually perform their work – employing a toolkit of methods across monolingual, multilingual, and discrepancy based approaches. While we did not cover an overall holistic approach to borrowing detection in our investigation, we

did indeed demonstrate the integration of monolingual with multilingual methods in subsections §3.1.2 and §3.2.3 of *multilingual borrowing detection*, chapter §3.

4.1 Monolingual borrowing detection

The primary data source for our research in monolingual borrowing detection was the World Online Loan Database (WOLD) (Tresoldi, Forkel, and Morozova, 2019), consisting of 41 language wordlists from all over the world – annotated with word form, language, concept, borrowing information for each lexical entry, including segmented broad IPA – all packaged as a Cross-Linguistic Data Format (CLDF) database. For our purposes, all of this detail information was reduced to multiple training and test tests per language of word_id, segmented_IPA, and borrowed_status for training and testing of monolingual methods. Percentage of borrowed words ranged from 0.7%, for Mandarin Chinese, to 56.3%, for Selice Romani, over 1,308 standardized concepts, with the number of lexical entries varying between $\approx$ 950, for CeqWong, and $\approx 2,560$, for Otomi.

The investigation included a rudimentary baseline support vector machine (SVM) classifier with features composed from the set of phonemes in a word, the bag of sounds model. More fruitful models and the subject of our research were Markov chain and neural network based language word models used to estimate word cross-entropies.

The inherited word cross-entropy approach, fit language models, both Markov chain and neural network, based just on inherited words from training data, and then compared estimated cross-entropies for words from test data versus a specified critical value to determine borrowing. This was not a very effective approach with an average F1 score $\approx$ 0.44 overall languages for Markov chain and $\approx$ 0.43 for neural network models.

However, we realized we could fit both inherited and borrowed language word models, and then compare the cross-entropies calculated by each model where the least cross-entropy result determined whether the word is inherited or borrowed. This was much more effective with an average F1 score $\approx$ 0.58 overall languages for Markov chain and $\approx$ 0.61 for neural networks.

This **important and effective yet simple innovation** named *competing cross-entropies* works because by using paired model estimates of cross-entropy for each word, the variability across words is controlled for – just as in a paired or blocked trials experiment. Comparisons are made for each word between borrowed or inherited models, largely eliminating word variability as a source of error. The comparison is more powerful. However, there still are unaccounted for sources of error, so the approach does not perform well enough to suffice by itself for borrowing detection.

We performed several experiments with this approach, and learned that performance improves when there are more borrowings and when there are single donors accounting for more than 67% of the borrowings. We subsequently named such super-majority donors as *dominant donors*. This improved result seems due to: 1) having more borrowing data from which to estimate the borrowing model, 2) having borrowings from largely a single donor, so that the borrowing model is estimated for essentially a single language, and maybe 3) language contact was over a shorter time, resulting in more consistent borrowing and adaption of borrowed words.

The original recurrent neural network model, used an embeddings layer and recurrent sequence layer per target language for borrowed or inherited word subgroups. The *transformer* model is closer to state of the art, and replaces the recurrent sequence layer with a multi-head attention layer, feed forward layer, and add and normalization steps. We expected this change to result in reduced execution time and improved detection performance, but there was only a modest reduction in execution time and an insignificant improvement in F1 scores. However, the introduction of the transformer model did set us up better for further experimentation. So while not so successful, it was an **enabling step** in or research.

With the transformer, we tried out **innovative** enhancements of: 1) distinguishing borrowings by individual donors, and 2) direct classification with the model sending flattened transformer output to a fully connected logistic layer to make detection decisions. The competing cross-entropies model with inherited versus individual donors gave slightly reduced F1 scores versus the simple inherited versus borrowed model. Perhaps, the reduced number of training examples for individual donors offset the expected gain from more consistent individual donor alternatives. This approach was innovative, but ineffective in improving borrowing detection. In the direct model case, F1 score performance was reduce by 5 percentage points versus the competing cross-entropies models. Either the classification layer needs to be improved beyond the fully connected logistic function, or individual cross-entropies capture vital information that is not captured by the classification layer.

Experiments to artificially augment data to supplement borrowings, were **innovative**, but ineffective in improving borrowing detection. The simple case of treating a Spanish wordlist as though it were partially adapted borrowings from Spanish, made negligible difference in F1 score, but did substantially impact the precision and recall. Precision was improved and recall reduced with borrowed word training data augmented by a Spanish wordlist.

The much more complex attempt of translating a Spanish wordlist to borrowed words, based on a transducer trained on borrowed words, resulted in a negligible one percentage point improvement, for a very complex and lengthy process of training translators, translating wordlists, and trying out detection with augmented wordlists. In this case, recall was improved while precision suffered.

While there are still many possibilities for further research in monolingual borrowing detection, looking to multilingual and cross-linguistic approaches seemed

to offer richer possibilities. This was the conclusion of our papers on borrowing detection using monolingual lexical models (Miller et al., 2020; Miller, Pariasca, and Beltran Castañon, 2021) and reaffirmed here.

4.2 Multilingual borrowing detection

For the primary data source of our research in multilingual borrowing detection, we selected the seven Latin-American indigenous languages from the World Online Loan Database (WOLD) (Tresoldi, Forkel, and Morozova, 2019), added in a Spanish dominant donor wordlist, and saved this in a cross-linguistic data format (CLDF) database for development of multilingual wordlist methods of borrowing detection.

We limited our study to the special case of multiple recipient languages with a dominant donor, Spanish in this case. Detected borrowings are assumed to come from the dominant donor, and we detected and measured detection performance based on a "borrowed or not-borrowed from the dominant donor" distinction. Both this dominant donor policy and the definition of detection relative to the dominant donor were **innovative**, and subsequently effective.

The closest match method performed pairwise alignments and matched between the dominant language and each recipient language – for lexical entries ordered by concept. We tried both normalized edit distance (NED), and sound class based phonetic alignment (SCA) distance functions. The cognate-based method used multiple alignments built from pairwise alignments, either NED or SCA, to identify likely cognate sets. Cognate sets were qualified as borrowings from a dominant donor if they included the dominant donor.

Finally the classifier based method used a classifier as a meta-model to combine NED and SCA functions, and target language indicators as features. Use of a meta-model of multiple distance measures was **innovative** and successful. The cross-validation approach of using the same fixed partitions of train and test over all methods, allowed us to use a more powerful analysis of variance (ANOVA) factoring out nuisance effects of the partition and giving more statistical power to the method comparisons. The statistical methods are standard, and the application was **innovative** and effective.

Error analysis revealed several significant opportunities for improvement of our methods: 1) increase coverage of donor wordlist, 2) relax same concept restriction, and 3) add alternative distance measures. While performing an error analysis is hardly innovative, its application here was very informative and influential on subsequent effort.

Preliminary investigation of increased wordlist coverage, showed that recall percentage could be increased by 10 percentage points by adding the curated Concepticon (List et al., 2022a) Spanish wordlist to the already present IDS (Key and Comrie, 2015) Spanish wordlist. Up to 90% coverage could be possible using using this combined wordlist based on results using *stemmed* forms, which

should be similar to matching on segmented broad IPA sound sequences where right-hand side results are trimmed or discounted.

Preliminary investigation of relaxation of same concept restriction, revealed a more complex and difficult picture, where simply loosening up the same concept restriction, results in longer execution time and poorer borrowing detection as false positives for similar sounding words from dissimilar concepts overwhelms the benefit of a few more true positive matches. This needs to be done carefully. These results, confirm our prior expectations in a resounding way through an **innovative** and effective use of Concepticon functions coupled with modifications to the closest match method.

Finally, we added a least cross-entropy (LCE) distance measure based on Markov chain language word models. Using the same train, test partitions as before, we tested LCE both stand-alone and as a function added to the classifier. The classifier with NED, SCA, and LCE functions performs substantially better than other methods. This was both nicely **innovative** and highly successful.

An important general learning, going beyond our work in multilingual borrowing detection, has been that "combining a variety of complementary functions in a meta-model classifier can result in substantially improved results versus any single function."

4.3 Path forward

Borrowing detection applied to Pano-Tacanan languages. We are in the process of curating a borrowing database of Pano-Tacanan and neighboring languages from the region of Peru and neighboring countries. Wordlists come originally from the Intercontinental Dictionary Series (IDS) (Key and Comrie, 2015) which are similar in size to WOLD (Tresoldi, Forkel, and Morozova, 2019) wordlists, and share largely the same concept space. Curation includes: 1. Annotate forms as segmented Broad IPA, 2. Include Spanish and Portuguese language wordlists from IDS as likely donor languages, as well as Imbabura Quechua from WOLD, 3. Annotate borrowings including donor language, form and meaning, with emphasis on Spanish, Portuguese, and Quechua as important donors, 4. Compile the database into a shareable cross linguistic data format (CLDF) (Forkel and List, 2020).

Our multilingual methods used only dominant donor based borrowing detection, whereas a general borrowing solution should support general and donor specific donor detection - both in the detection process and in the evaluation of borrowing detection. We expect to generalize the *classifier* from multilingual borrowing detection to function in general borrowing detection as well as donor dominant borrowing detection, and provide evaluations for both cases.

Borrowing detection based on the generalized *classifier* would be applied to the Pano-Tacanan borrowing database in order to illustrate this approach on our Peruvian and neighboring countries' languages. The resulting database with software will be put into the public domain for all to use.

Make borrowing detection tools usable by historical linguists. This thesis has focused on development and evaluation of monolingual and multilingual methods of lexical borrowing detection. However, the utility of this work is in the application of these methods to real world wordlists such as from IDS (Key and Comrie, 2015) and the planned Pano-Tacanan borrowing database. To make our work usable by historical linguists, we need to demonstrate borrowing analyses on real world data, using an application interface understandable by historical linguists without requiring specialists in computational historical linguistics. While our applications already provide for borrowing analyses on accompanying databases, this needs to be made even more accessible, and the output of such analyses focused more on historical linguistic use.

Augment donor wordlist. Using Concepticon (List et al., 2022a) and IDS (Key and Comrie, 2015) wordlists combined could achieve 90% coverage for many common wordlists and vocabulary. To get beyond 90% coverage or beyond the concept coverage of Concepticon, would require a substantially greater effort if the experience with *Big (8,600) Spanish* wordlist (Neri, 2018) applies more generally. In that case it would take a large dictionary of tens of thousands of words with corresponding concept and segmented sounds to go the extra kilometer to 95% or greater coverage. This may be doable, given a focus on dominant donor languages, most of which are majority languages with sufficient linguistic resources, but it would still be a substantial effort.

Relax concept restriction. The Central concept restriction does not offer the fine granularity necessary to optimize the interplay between semantic and phonetic distances of words. Instead a more general vector based model of meaning might be more workable, where vector cosine similarity could suffice as a semantic similarity measure, and semantic similarity might combine in a classifier with various monolingual and multilingual phonetic distance measures to predict borrowing. Instead of a curated conceptual structure, there is a machine learned vector representation of tens of thousands of words, combined with a broad IPA sound segment encoding, itself created from an orthographic to sound segment mapping of those tens of thousands of words as text. This could provide the basis to solve both the augmented donor wordlist and relaxed concept restriction problems.

Dominant donor and general borrowing detection. Monolingual borrowing evaluated borrowing in general with a brief foray into borrowing by donor, while multilingual borrowing evaluated dominant donor borrowing without considering borrowing in general. It would be useful to evaluate both borrowing by language donor, whether dominant or not, and overall, in monolingual and multilingual contexts.

What might that look like? This would also support the case of multiple donor languages. Dominant donor would no longer be a special case, but just one donor among others. Our Pano-Tacanan borrowing database (above path) anticipates at least a partial solution to this problem.

Enhance classifier for multilingual wordlists. The logistic regression (LR) *classifier* performs on par with the support vector machine (SVM) *classifier*. However, if the LR *classifier* were implemented as a neural network, this would open up the possibility to stack a fully-connected hidden layer between the input layer of *closest match*, *least cross-entropy*, target language and other inputs, and a logistic classification layer. This extra hidden layer, for being fully connected, would add substantially more parameters to the *classifier*, but this might also be what is needed to improve borrowing detection.

Enhance classifier for monolingual wordlists. The direct classifiers used with monolingual wordlists performed five percentage points lower than corresponding competing cross-entropies methods. While we attributed this in part to the specialness of competing cross-entropies, that is not necessarily the case. With much recent effort and articles written on neural network based *classifiers*, it's still possible that a direct classifier could be made to outperform competing cross-entropy methods.

Enhance language word models to use a common embedding layer. A common embedding layer for broad IPA segments might permit better learning of embeddings through sharing across languages and borrowing categories. Language specific or category specific, e.g., inherited, borrowed, language donor, differences could enter with the hidden layers. The model would become more complex, but by leveraging embeddings over all languages and categories, the overall neural borrowing detection model could gain greater power to discriminate between borrowing categories.

Drop the wordlist requirement. If donor wordlists grow substantially, and concept restrictions are relaxed, and dominant donor becomes just another donor instance in borrowing detection, then the next step could be to leave behind the requirement of curated wordlists altogether. Image borrowing detection that is more than 90% effective (F1 score > 0.90) and applicable to general text. Still not ready to replace an experienced historical linguist, but now an important tool for work in historical linguistics.

Not the end state. Next would be to go beyond lexical borrowing.

Appendix A

Monolingual detail results

Here are detail results by language for the major experiments from monolingual borrowing detection §2.2.

Detail results for artificially seeded borrowings from §2.2.1 are given in Tables A.1, A.2, and A.3 for 20%, 10%, and 5% borrowings correspondingly.

Detail results for real world borrowing from the world online loan database (WOLD) from §2.2.2 are given in Tabs. A.4, A.5, A.6, and A.7 for competing entropies and inherited only language model approaches with means and standard deviations reported separately.

Followup investigations on detection performance are documented in §2.2.3, factors influencing borrowing detection, and §2.2.4, detecting borrowings with dominant donor language. Detail information used in these investigations including borrowing percentages and phonological characteristics by language are reported together in Tab. A.8.

The neural network transformer language model was adopted to replace the original recurrent language model in §2.2.6. Detail by language results, both mean and standard deviations, are reported for the competing cross-entropies approach using the transformer model in Tab. A.9.

TABLE A.1: Artificially seeded 20% borrowing - 10-fold cross-validation.

	Neural net				Markov chain				Bag of sounds			
Language	**Prec.**	**Recall**	**F1**	**Acc.**	**Prec.**	**Recall**	**F1**	**Acc.**	**Prec.**	**Recall**	**F1**	**Acc.**
Archi	0.98	0.98	0.98	0.99	0.92	0.98	0.95	0.98	0.91	0.91	0.91	0.96
Bezhta	1.00	1.00	1.00	1.00	0.93	0.96	0.95	0.98	1.00	0.96	0.98	0.99
Ceq Wong	0.96	0.93	0.95	0.98	0.84	0.97	0.90	0.95	1.00	0.79	0.88	0.96
Dutch	0.87	0.86	0.87	0.94	0.77	0.89	0.83	0.92	1.00	0.69	0.82	0.93
English	0.85	1.00	0.92	0.96	0.86	0.92	0.89	0.95	1.00	0.84	0.91	0.97
Gawwada	0.95	0.98	0.97	0.98	0.93	0.98	0.96	0.98	0.95	0.90	0.92	0.97
Gurindji	0.98	0.98	0.98	0.99	0.98	1.00	0.99	1.00	1.00	1.00	1.00	1.00
Hausa	1.00	0.99	0.99	1.00	1.00	0.98	0.99	1.00	0.98	0.87	0.92	0.97
Hawaiian	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	0.99	1.00
Hup	1.00	1.00	1.00	1.00	1.00	0.98	0.99	1.00	1.00	0.89	0.94	0.98
Imbabura Quechua	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.92	0.96	0.98
Indonesian	0.99	1.00	0.99	1.00	1.00	1.00	1.00	1.00	0.98	0.97	0.98	0.99
Iraqw	0.93	0.98	0.96	0.98	0.92	0.92	0.92	0.97	1.00	0.87	0.93	0.97
Japanese	1.00	1.00	1.00	1.00	0.99	1.00	0.99	1.00	0.99	0.95	0.97	0.99
Kali'na	1.00	1.00	1.00	1.00	0.98	0.98	0.98	0.99	1.00	0.98	0.99	1.00
Kanuri	1.00	1.00	1.00	1.00	1.00	0.98	0.99	1.00	1.00	0.91	0.95	0.98
Ket	0.96	0.98	0.97	0.99	0.86	0.98	0.92	0.97	1.00	0.90	0.95	0.98
Kildin Saami	0.99	1.00	0.99	1.00	0.93	0.96	0.95	0.98	0.95	0.84	0.89	0.96
Lower Sorbian	0.98	0.97	0.98	0.99	0.95	0.99	0.97	0.99	0.97	0.92	0.95	0.98
Malagasy	1.00	1.00	1.00	1.00	0.99	1.00	0.99	1.00	1.00	0.92	0.96	0.99
Manange	0.93	1.00	0.96	0.98	0.98	0.96	0.97	0.99	1.00	0.93	0.96	0.98
Mandarin Chinese	0.99	1.00	0.99	1.00	0.99	1.00	1.00	1.00	0.99	0.89	0.94	0.97
Mapudungun	0.97	1.00	0.98	0.99	0.98	1.00	0.99	1.00	1.00	0.98	0.99	1.00
Old High German	0.91	0.91	0.91	0.97	0.85	0.92	0.89	0.95	0.90	0.86	0.88	0.95
Oroqen	0.95	0.98	0.97	0.99	0.90	0.93	0.92	0.96	0.98	0.68	0.80	0.93
Otomi	0.99	0.99	0.99	1.00	0.98	1.00	0.99	1.00	1.00	0.97	0.98	0.99
Q'eqchi'	0.99	0.95	0.97	0.99	0.98	0.98	0.98	0.99	1.00	0.93	0.96	0.99
Romanian	0.96	0.99	0.97	0.99	0.92	0.97	0.95	0.97	1.00	0.91	0.95	0.98
Sakha	0.96	1.00	0.98	0.99	0.93	0.98	0.96	0.98	0.98	0.90	0.94	0.98
Saramaccan	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.97	0.88	0.92	0.97
Selice Romani	0.90	1.00	0.95	0.98	0.94	0.97	0.96	0.98	1.00	0.94	0.97	0.99
Seychelles Creole	0.99	0.99	0.99	1.00	0.96	1.00	0.98	0.99	1.00	0.92	0.96	0.99
Swahili	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.94	0.96	0.99
Takia	0.96	0.96	0.96	0.98	0.97	1.00	0.98	0.99	1.00	0.96	0.98	0.99
Tarifiyt Berber	0.97	0.91	0.94	0.98	0.97	0.95	0.96	0.99	0.98	0.98	0.98	0.99
Thai	0.98	0.98	0.98	0.99	0.91	0.95	0.93	0.97	0.97	0.90	0.93	0.97
Vietnamese	1.00	0.98	0.99	1.00	0.93	0.91	0.92	0.98	1.00	0.98	0.99	1.00
White Hmong	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	0.99	1.00
Wichí	0.98	1.00	0.99	1.00	0.98	1.00	0.99	1.00	1.00	0.91	0.95	0.98
Yaqui	1.00	0.97	0.98	0.99	0.93	1.00	0.97	0.98	1.00	0.96	0.98	0.99
Zinacantán Tzotzil	0.97	0.99	0.98	0.99	0.98	1.00	0.99	1.00	1.00	0.96	0.98	0.99
Mean	0.97	0.98	0.98	0.99	0.95	0.98	0.96	0.98	0.99	0.91	0.95	0.98

TABLE A.2: Artificially seeded 10% borrowing - 10-fold cross-validation.

	Neural net				Markov chain				Bag of sounds			
Language	**Prec.**	**Recall**	**F1**	**Acc.**	**Prec.**	**Recall**	**F1**	**Acc.**	**Prec.**	**Recall**	**F1**	**Acc.**
Archi	0.91	1.00	0.95	0.99	0.70	0.95	0.81	0.96	1.00	0.91	0.95	0.99
Bezhta	0.83	1.00	0.91	0.98	0.72	1.00	0.84	0.97	1.00	0.95	0.98	1.00
Ceq Wong	0.68	0.94	0.79	0.94	0.69	0.95	0.80	0.94	1.00	0.73	0.85	0.97
Dutch	0.58	0.71	0.64	0.94	0.58	0.81	0.68	0.93	0.95	0.64	0.77	0.96
English	0.95	0.95	0.95	0.99	0.54	1.00	0.70	0.91	1.00	0.61	0.76	0.97
Gawwada	0.81	0.95	0.88	0.97	0.93	1.00	0.97	0.99	1.00	0.84	0.91	0.98
Gurindji	1.00	1.00	1.00	1.00	1.00	0.94	0.97	0.99	1.00	0.91	0.95	0.99
Hausa	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.81	0.90	0.98
Hawaiian	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Hup	1.00	1.00	1.00	1.00	0.90	1.00	0.95	0.99	1.00	0.86	0.93	0.99
Imbabura Quechua	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.85	0.92	0.98
Indonesian	0.97	0.97	0.97	0.99	1.00	0.97	0.98	1.00	1.00	0.76	0.87	0.97
Iraqw	0.82	1.00	0.90	0.98	0.86	0.96	0.91	0.98	1.00	0.92	0.96	0.99
Japanese	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	0.91	0.94	0.99
Kali'na	1.00	1.00	1.00	1.00	1.00	0.93	0.97	0.99	1.00	0.93	0.96	0.99
Kanuri	0.97	1.00	0.98	1.00	1.00	1.00	1.00	1.00	1.00	0.92	0.96	0.99
Ket	0.87	0.96	0.91	0.98	0.70	0.86	0.78	0.96	1.00	0.79	0.88	0.98
Kildin Saami	0.88	1.00	0.94	0.99	0.82	1.00	0.90	0.98	1.00	0.69	0.81	0.96
Lower Sorbian	0.96	0.96	0.96	0.99	0.92	0.97	0.95	0.99	1.00	0.88	0.94	0.99
Malagasy	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.87	0.93	0.98
Manange	1.00	1.00	1.00	1.00	0.96	0.96	0.96	0.99	1.00	0.93	0.96	0.99
Mandarin Chinese	0.94	1.00	0.97	1.00	0.90	1.00	0.95	0.99	1.00	0.80	0.89	0.98
Mapudungun	0.95	1.00	0.98	1.00	0.96	1.00	0.98	1.00	1.00	0.96	0.98	1.00
Old High German	0.91	0.97	0.94	0.98	0.81	0.94	0.87	0.97	1.00	0.73	0.84	0.98
Oroqen	0.81	1.00	0.90	0.98	0.67	1.00	0.81	0.94	1.00	0.85	0.92	0.98
Otomi	0.98	1.00	0.99	1.00	1.00	1.00	1.00	1.00	0.96	0.89	0.92	0.98
Q'eqchi'	0.87	1.00	0.93	0.99	0.94	0.94	0.94	0.99	1.00	0.83	0.91	0.98
Romanian	0.84	0.97	0.90	0.98	0.84	0.91	0.88	0.97	1.00	0.80	0.89	0.98
Sakha	0.91	0.97	0.94	0.98	0.84	0.96	0.90	0.98	1.00	0.84	0.91	0.98
Saramaccan	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.87	0.93	0.99
Selice Romani	0.81	1.00	0.90	0.98	0.92	0.86	0.89	0.98	1.00	0.83	0.91	0.98
Seychelles Creole	0.98	0.98	0.98	1.00	0.91	0.98	0.94	0.99	1.00	0.91	0.96	0.99
Swahili	0.96	0.96	0.96	0.99	1.00	1.00	1.00	1.00	0.96	0.93	0.94	0.99
Takia	0.97	0.97	0.97	0.99	0.92	1.00	0.96	0.99	1.00	1.00	1.00	1.00
Tarifiyt Berber	0.88	0.96	0.92	0.98	0.84	1.00	0.91	0.98	0.91	0.83	0.87	0.97
Thai	0.94	0.91	0.93	0.99	0.76	0.89	0.82	0.96	1.00	0.90	0.95	0.99
Vietnamese	0.97	1.00	0.98	1.00	0.91	0.97	0.94	0.99	1.00	0.85	0.92	0.99
White Hmong	1.00	1.00	1.00	1.00	1.00	0.96	0.98	1.00	1.00	0.95	0.98	1.00
Wichí	1.00	0.97	0.98	1.00	0.97	1.00	0.98	1.00	1.00	0.93	0.96	0.99
Yaqui	1.00	1.00	1.00	1.00	0.92	0.96	0.94	0.99	1.00	0.88	0.94	0.99
Zinacantán Tzotzil	0.97	1.00	0.99	1.00	0.96	0.96	0.96	0.99	1.00	0.96	0.98	1.00
Mean	0.92	0.98	0.95	0.99	0.89	0.97	0.92	0.98	0.99	0.86	0.92	0.98

TABLE A.3: Artificially seeded 5% borrowing - 10-fold cross-validation.

	Neural net				Markov chain				Bag of sounds			
Language	**Prec.**	**Recall**	**F1**	**Acc.**	**Prec.**	**Recall**	**F1**	**Acc.**	**Prec.**	**Recall**	**F1**	**Acc.**
Archi	0.50	1.00	0.67	0.97	0.27	1.00	0.43	0.88	1.00	0.80	0.89	0.99
Bezhta	0.71	1.00	0.83	0.98	0.60	0.92	0.73	0.96	1.00	0.83	0.91	0.99
Ceq Wong	0.40	1.00	0.57	0.96	0.31	1.00	0.47	0.87	1.00	0.67	0.80	0.98
Dutch	0.56	0.83	0.67	0.96	0.52	0.86	0.65	0.95	0.86	0.50	0.63	0.97
English	0.67	1.00	0.80	0.98	0.40	1.00	0.57	0.94	1.00	0.67	0.80	0.99
Gawwada	0.94	0.89	0.91	0.99	0.82	1.00	0.90	0.99	1.00	0.50	0.67	0.97
Gurindji	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Hausa	0.86	1.00	0.92	0.99	0.81	0.94	0.87	0.98	1.00	0.64	0.78	0.99
Hawaiian	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Hup	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.75	0.86	0.99
Imbabura Quechua	1.00	0.92	0.96	0.99	0.86	1.00	0.92	0.99	1.00	1.00	1.00	1.00
Indonesian	1.00	1.00	1.00	1.00	1.00	0.95	0.97	1.00	1.00	0.86	0.92	0.99
Iraqw	0.65	0.85	0.73	0.97	0.67	1.00	0.80	0.98	0.93	0.88	0.90	0.99
Japanese	0.95	1.00	0.97	1.00	0.78	1.00	0.88	0.99	1.00	0.84	0.91	0.99
Kali'na	1.00	1.00	1.00	1.00	0.89	0.89	0.89	0.99	1.00	0.86	0.92	0.99
Kanuri	0.78	1.00	0.88	0.99	0.90	1.00	0.95	0.99	1.00	0.77	0.87	0.99
Ket	0.55	1.00	0.71	0.96	0.42	0.93	0.58	0.92	1.00	0.88	0.93	0.99
Kildin Saami	0.73	1.00	0.85	0.98	0.57	0.92	0.71	0.96	1.00	0.83	0.91	0.99
Lower Sorbian	0.78	0.91	0.84	0.97	0.71	1.00	0.83	0.99	1.00	0.75	0.86	0.99
Malagasy	0.95	1.00	0.97	1.00	1.00	1.00	1.00	1.00	1.00	0.93	0.96	1.00
Manange	0.75	1.00	0.86	0.99	0.69	1.00	0.81	0.98	1.00	0.92	0.96	1.00
Mandarin Chinese	0.90	1.00	0.95	1.00	0.89	1.00	0.94	0.99	1.00	0.77	0.87	0.99
Mapudungun	0.90	1.00	0.95	1.00	0.81	1.00	0.90	0.99	1.00	1.00	1.00	1.00
Old High German	0.62	0.83	0.71	0.97	0.58	0.92	0.71	0.96	1.00	0.73	0.84	0.99
Oroqen	0.67	1.00	0.80	0.97	0.61	0.93	0.74	0.96	0.71	0.56	0.63	0.97
Otomi	1.00	1.00	1.00	1.00	0.94	1.00	0.97	1.00	1.00	0.85	0.92	0.99
Q'eqchi'	0.88	1.00	0.94	0.99	0.82	0.93	0.87	0.99	1.00	0.71	0.83	0.99
Romanian	0.76	0.84	0.80	0.97	0.67	0.93	0.78	0.97	1.00	0.87	0.93	0.99
Sakha	0.69	1.00	0.81	0.98	0.61	0.93	0.74	0.96	1.00	0.73	0.85	0.98
Saramaccan	0.44	0.80	0.57	0.97	0.69	1.00	0.82	0.98	1.00	0.89	0.94	0.99
Selice Romani	0.50	0.88	0.64	0.95	0.50	1.00	0.67	0.94	1.00	0.67	0.80	0.98
Seychelles Creole	0.87	0.87	0.87	0.99	0.71	0.95	0.82	0.98	1.00	0.67	0.80	0.98
Swahili	0.78	1.00	0.88	0.99	0.95	1.00	0.97	1.00	1.00	0.85	0.92	0.99
Takia	0.85	1.00	0.92	0.99	0.52	1.00	0.69	0.95	1.00	0.65	0.79	0.97
Tarifiyt Berber	0.58	1.00	0.74	0.97	0.24	1.00	0.39	0.86	0.83	0.45	0.59	0.96
Thai	0.74	0.82	0.78	0.98	0.70	0.95	0.81	0.97	1.00	0.87	0.93	0.99
Vietnamese	0.93	1.00	0.97	1.00	0.61	1.00	0.76	0.96	1.00	0.71	0.83	0.98
White Hmong	1.00	1.00	1.00	1.00	0.61	1.00	0.76	0.97	1.00	0.83	0.91	0.99
Wichí	0.81	1.00	0.90	0.99	0.92	0.92	0.92	0.99	1.00	0.92	0.96	1.00
Yaqui	0.93	1.00	0.97	1.00	0.88	1.00	0.94	0.99	1.00	0.91	0.95	1.00
Zinacantán Tzotzil	1.00	1.00	1.00	1.00	0.88	1.00	0.93	0.99	1.00	0.75	0.86	0.99
Mean	0.80	0.96	0.86	0.99	0.72	0.97	0.81	0.97	0.98	0.79	0.87	0.99

TABLE A.4: Real world borrowing from the World Online Loan Database (WOLD) – 10-fold cross-validation - means.

	Neural net				Markov chain				Bag of sounds				Prop.
Language	**Prec.**	**Recall**	**F1**	**Acc**	**Prec.**	**Recall**	**F1**	**Acc**	**Prec.**	**Recall**	**F1**	**Acc**	**Inher.**
Archi	0.593	0.748	0.655	0.841	0.549	0.757	0.634	0.817	0.687	0.267	0.375	0.817	0.786
Bezhta	0.760	0.814	0.784	0.868	0.699	0.759	0.726	0.832	0.772	0.705	0.735	0.849	0.701
Ceq Wong	0.716	0.773	0.741	0.823	0.614	0.682	0.641	0.752	0.660	0.575	0.611	0.762	0.667
Dutch	0.470	0.585	0.512	0.800	0.442	0.566	0.493	0.784	0.200	0.007	0.013	0.814	0.814
English	0.615	0.660	0.634	0.709	0.609	0.635	0.620	0.701	0.667	0.517	0.578	0.712	0.614
Gawwada	0.367	0.628	0.456	0.864	0.352	0.572	0.427	0.867	0.618	0.167	0.252	0.915	0.909
Gurindji	0.244	0.468	0.318	0.753	0.330	0.578	0.416	0.798	0.000	0.000	0.000	0.875	0.875
Hausa	0.550	0.654	0.594	0.826	0.525	0.683	0.591	0.815	0.705	0.290	0.397	0.832	0.802
Hawaiian	0.438	0.773	0.557	0.805	0.457	0.720	0.556	0.817	0.600	0.034	0.063	0.845	0.839
Hup	0.658	0.883	0.750	0.940	0.562	0.829	0.658	0.919	0.921	0.448	0.590	0.940	0.899
Imbabura Quechua	0.784	0.845	0.811	0.892	0.771	0.846	0.803	0.886	0.808	0.604	0.687	0.849	0.720
Indonesian	0.611	0.676	0.640	0.771	0.602	0.635	0.617	0.765	0.641	0.269	0.378	0.734	0.698
Iraqw	0.578	0.742	0.638	0.895	0.501	0.727	0.582	0.871	0.689	0.375	0.470	0.895	0.870
Japanese	0.720	0.823	0.766	0.851	0.712	0.771	0.738	0.839	0.676	0.406	0.506	0.768	0.704
Kali'na	0.482	0.688	0.559	0.853	0.510	0.649	0.563	0.858	0.967	0.200	0.327	0.883	0.855
Kanuri	0.432	0.627	0.506	0.789	0.437	0.596	0.497	0.792	0.717	0.053	0.098	0.830	0.823
Ket	0.461	0.780	0.551	0.908	0.433	0.702	0.530	0.899	0.647	0.212	0.301	0.928	0.916
Kildin Saami	0.448	0.575	0.501	0.789	0.425	0.551	0.479	0.775	0.050	0.004	0.007	0.807	0.810
Lower Sorbian	0.612	0.738	0.665	0.857	0.590	0.692	0.633	0.845	0.692	0.191	0.293	0.824	0.803
Malagasy	0.380	0.627	0.470	0.823	0.375	0.580	0.453	0.826	0.000	0.000	0.000	0.875	0.875
Manange	0.320	0.613	0.411	0.889	0.262	0.579	0.345	0.866	0.400	0.069	0.114	0.939	0.935
Mandarin Chinese	0.019	0.092	0.031	0.950	0.004	0.050	0.007	0.816	0.000	0.000	0.000	0.993	0.993
Mapudungun	0.650	0.820	0.723	0.875	0.656	0.803	0.720	0.877	0.815	0.528	0.637	0.882	0.800
Old High German	0.200	0.335	0.241	0.887	0.190	0.433	0.261	0.866	0.000	0.000	0.000	0.947	0.947
Oroqen	0.305	0.596	0.388	0.864	0.266	0.507	0.341	0.854	0.600	0.088	0.150	0.928	0.922
Otomi	0.704	0.911	0.792	0.953	0.621	0.892	0.729	0.936	0.874	0.688	0.764	0.959	0.902
Q'eqchi'	0.658	0.845	0.735	0.937	0.609	0.822	0.693	0.927	0.807	0.513	0.624	0.937	0.895
Romanian	0.697	0.713	0.704	0.761	0.663	0.720	0.690	0.741	0.634	0.421	0.505	0.671	0.600
Sakha	0.612	0.687	0.644	0.812	0.555	0.640	0.592	0.783	0.684	0.217	0.323	0.777	0.751
Saramaccan	0.583	0.637	0.606	0.707	0.584	0.629	0.605	0.709	0.597	0.076	0.134	0.653	0.645
Selice Romani	0.902	0.877	0.889	0.876	0.887	0.877	0.882	0.865	0.746	0.834	0.787	0.742	0.427
Seychelles Creole	0.274	0.573	0.364	0.828	0.315	0.594	0.409	0.849	0.200	0.011	0.020	0.911	0.911
Swahili	0.687	0.775	0.725	0.860	0.636	0.692	0.661	0.828	0.795	0.540	0.637	0.854	0.758
Takia	0.607	0.795	0.684	0.836	0.593	0.752	0.660	0.822	0.800	0.042	0.078	0.777	0.768
Tarifiyt Berber	0.813	0.789	0.798	0.806	0.774	0.771	0.772	0.778	0.773	0.697	0.730	0.749	0.511
Thai	0.541	0.654	0.590	0.805	0.459	0.648	0.531	0.759	0.591	0.104	0.175	0.791	0.785
Vietnamese	0.449	0.641	0.524	0.789	0.421	0.595	0.489	0.777	0.617	0.113	0.186	0.824	0.817
White Hmong	0.373	0.589	0.451	0.784	0.383	0.656	0.482	0.782	0.325	0.024	0.045	0.846	0.845
Wichí	0.751	0.900	0.817	0.941	0.729	0.839	0.772	0.933	0.726	0.523	0.602	0.904	0.857
Yaqui	0.746	0.836	0.786	0.893	0.754	0.804	0.776	0.890	0.780	0.560	0.649	0.857	0.760
Zinacantán Tzotzil	0.748	0.905	0.815	0.942	0.696	0.852	0.762	0.924	0.931	0.412	0.565	0.912	0.857
Mean	0.550	0.700	0.606	0.845	0.526	0.675	0.581	0.830	0.595	0.287	0.351	0.844	0.797

TABLE A.5: Real world borrowing from the World Online Loan Database (WOLD) – 10-fold cross-validation - standard deviations.

	Neural net				Markov chain				Bag of sounds			
Language	**Prec.**	**Recall**	**F1**	**Acc.**	**Prec.**	**Recall**	**F1**	**Acc.**	**Prec.**	**Recall**	**F1**	**Acc.**
Archi	0.113	0.072	0.082	0.026	0.048	0.098	0.056	0.027	0.211	0.119	0.138	0.048
Bezhta	0.058	0.076	0.053	0.026	0.095	0.089	0.087	0.045	0.056	0.065	0.048	0.030
Ceq Wong	0.084	0.062	0.063	0.036	0.102	0.036	0.069	0.040	0.093	0.075	0.065	0.032
Dutch	0.115	0.115	0.101	0.024	0.096	0.099	0.090	0.045	0.422	0.015	0.029	0.031
English	0.055	0.071	0.047	0.051	0.075	0.073	0.067	0.048	0.063	0.078	0.057	0.037
Gawwada	0.097	0.133	0.091	0.040	0.085	0.160	0.096	0.029	0.407	0.121	0.170	0.020
Gurindji	0.084	0.161	0.106	0.041	0.083	0.115	0.091	0.046	0.000	0.000	0.000	0.031
Hausa	0.113	0.071	0.089	0.035	0.084	0.071	0.072	0.033	0.166	0.108	0.122	0.038
Hawaiian	0.075	0.065	0.070	0.026	0.068	0.079	0.059	0.032	0.516	0.042	0.074	0.021
Hup	0.109	0.087	0.091	0.028	0.152	0.133	0.117	0.025	0.114	0.148	0.131	0.019
Imbabura Quechua	0.085	0.050	0.061	0.028	0.066	0.063	0.041	0.022	0.084	0.109	0.086	0.034
Indonesian	0.063	0.054	0.047	0.040	0.056	0.078	0.061	0.027	0.074	0.029	0.039	0.034
Iraqw	0.146	0.091	0.099	0.027	0.115	0.128	0.087	0.022	0.185	0.146	0.134	0.036
Japanese	0.062	0.068	0.047	0.031	0.059	0.053	0.039	0.021	0.092	0.051	0.057	0.013
Kali'na	0.143	0.130	0.132	0.023	0.054	0.154	0.079	0.023	0.105	0.069	0.098	0.022
Kanuri	0.117	0.091	0.107	0.030	0.115	0.073	0.075	0.034	0.343	0.033	0.058	0.022
Ket	0.211	0.181	0.219	0.028	0.160	0.171	0.161	0.029	0.403	0.154	0.199	0.019
Kildin Saami	0.081	0.136	0.098	0.024	0.072	0.094	0.078	0.021	0.158	0.011	0.021	0.034
Lower Sorbian	0.073	0.073	0.056	0.012	0.082	0.112	0.081	0.032	0.177	0.079	0.101	0.021
Malagasy	0.079	0.114	0.083	0.041	0.058	0.100	0.062	0.027	0.000	0.000	0.000	0.025
Manange	0.086	0.168	0.093	0.024	0.107	0.218	0.120	0.039	0.516	0.103	0.165	0.020
Mandarin Chinese	0.040	0.217	0.066	0.018	0.012	0.158	0.022	0.028	0.000	0.000	0.000	0.007
Mapudungun	0.049	0.066	0.033	0.023	0.081	0.079	0.067	0.021	0.107	0.086	0.084	0.031
Old High German	0.099	0.097	0.102	0.027	0.111	0.245	0.146	0.027	0.000	0.000	0.000	0.025
Oroqen	0.094	0.226	0.121	0.018	0.065	0.167	0.078	0.040	0.516	0.094	0.155	0.024
Otomi	0.044	0.053	0.031	0.008	0.104	0.066	0.091	0.020	0.098	0.074	0.052	0.009
Q'eqchi'	0.122	0.063	0.087	0.019	0.097	0.101	0.075	0.024	0.100	0.091	0.086	0.014
Romanian	0.055	0.035	0.037	0.027	0.047	0.064	0.048	0.042	0.078	0.053	0.058	0.035
Sakha	0.057	0.080	0.043	0.033	0.080	0.085	0.071	0.031	0.149	0.054	0.059	0.024
Saramaccan	0.074	0.048	0.053	0.044	0.057	0.057	0.052	0.042	0.218	0.031	0.052	0.032
Selice Romani	0.026	0.033	0.021	0.019	0.035	0.023	0.023	0.027	0.032	0.043	0.025	0.025
Seychelles Creole	0.096	0.102	0.103	0.021	0.048	0.089	0.051	0.024	0.422	0.024	0.045	0.014
Swahili	0.077	0.056	0.049	0.019	0.067	0.043	0.036	0.026	0.061	0.098	0.065	0.022
Takia	0.106	0.065	0.081	0.030	0.087	0.057	0.064	0.034	0.422	0.035	0.064	0.039
Tarifiyt Berber	0.057	0.051	0.032	0.027	0.033	0.046	0.031	0.024	0.051	0.063	0.027	0.027
Thai	0.080	0.075	0.067	0.029	0.096	0.076	0.067	0.031	0.087	0.035	0.048	0.024
Vietnamese	0.080	0.095	0.072	0.039	0.079	0.114	0.077	0.029	0.104	0.043	0.062	0.032
White Hmong	0.064	0.140	0.068	0.033	0.049	0.079	0.052	0.027	0.472	0.034	0.063	0.028
Wichí	0.085	0.055	0.067	0.028	0.136	0.092	0.093	0.020	0.189	0.064	0.102	0.019
Yaqui	0.062	0.056	0.039	0.017	0.079	0.069	0.063	0.031	0.056	0.062	0.050	0.018
Zinacantán Tzotzil	0.055	0.074	0.029	0.015	0.061	0.092	0.049	0.017	0.093	0.099	0.096	0.015
Mean	0.085	0.092	0.074	0.028	0.079	0.098	0.072	0.030	0.181	0.064	0.073	0.026

TABLE A.6: Real world borrowing from the World Online Loan Database (WOLD) – 10-fold cross-validation - means [inherited only methods].

	Neural net - inherited				Markov chain - inherited				Prop.
Language	**Prec.**	**Recall**	**F1**	**Acc**	**Prec.**	**Recall**	**F1**	**Acc**	**Inher.**
Archi	0.349	0.788	0.483	0.643	0.384	0.807	0.516	0.683	0.786
Bezhta	0.432	0.767	0.549	0.632	0.469	0.810	0.593	0.671	0.701
Ceq Wong	0.389	0.792	0.519	0.522	0.424	0.792	0.549	0.572	0.667
Dutch	0.244	0.788	0.370	0.506	0.249	0.813	0.379	0.507	0.814
English	0.458	0.798	0.579	0.557	0.475	0.806	0.595	0.581	0.614
Gawwada	0.127	0.828	0.219	0.467	0.149	0.781	0.248	0.585	0.909
Gurindji	0.137	0.793	0.233	0.349	0.145	0.796	0.245	0.382	0.875
Hausa	0.292	0.810	0.428	0.572	0.314	0.789	0.449	0.618	0.802
Hawaiian	0.193	0.798	0.309	0.425	0.196	0.804	0.314	0.442	0.839
Hup	0.273	0.770	0.402	0.774	0.242	0.821	0.368	0.727	0.899
Imbabura Quechua	0.621	0.805	0.696	0.806	0.685	0.788	0.730	0.839	0.720
Indonesian	0.445	0.795	0.570	0.641	0.430	0.795	0.557	0.619	0.698
Iraqw	0.196	0.770	0.309	0.557	0.254	0.821	0.385	0.663	0.870
Japanese	0.463	0.806	0.586	0.665	0.456	0.793	0.577	0.658	0.704
Kali'na	0.172	0.798	0.280	0.420	0.201	0.798	0.318	0.505	0.855
Kanuri	0.255	0.805	0.387	0.551	0.265	0.802	0.396	0.574	0.823
Ket	0.159	0.804	0.261	0.629	0.175	0.793	0.284	0.672	0.916
Kildin Saami	0.233	0.814	0.361	0.452	0.220	0.800	0.344	0.425	0.810
Lower Sorbian	0.303	0.798	0.439	0.599	0.339	0.802	0.477	0.654	0.803
Malagasy	0.173	0.782	0.283	0.510	0.175	0.802	0.285	0.504	0.875
Manange	0.173	0.787	0.279	0.749	0.188	0.785	0.298	0.768	0.935
Mandarin Chinese	0.010	0.450	0.019	0.513	0.009	0.617	0.018	0.445	0.993
Mapudungun	0.351	0.798	0.484	0.665	0.399	0.810	0.533	0.720	0.800
Old High German	0.077	0.805	0.139	0.474	0.086	0.814	0.155	0.535	0.947
Oroqen	0.140	0.810	0.235	0.595	0.142	0.791	0.236	0.608	0.922
Otomi	0.488	0.798	0.599	0.896	0.497	0.798	0.607	0.900	0.902
Q'eqchi'	0.387	0.826	0.523	0.842	0.374	0.807	0.508	0.837	0.895
Romanian	0.501	0.802	0.616	0.600	0.531	0.796	0.637	0.640	0.600
Sakha	0.321	0.819	0.460	0.523	0.360	0.803	0.496	0.594	0.751
Saramaccan	0.390	0.802	0.522	0.483	0.397	0.803	0.529	0.498	0.645
Selice Romani	0.772	0.796	0.783	0.748	0.815	0.795	0.803	0.778	0.427
Seychelles Creole	0.137	0.804	0.232	0.534	0.144	0.803	0.243	0.565	0.911
Swahili	0.330	0.809	0.468	0.557	0.336	0.800	0.472	0.571	0.758
Takia	0.321	0.788	0.454	0.564	0.373	0.817	0.512	0.640	0.768
Tarifiyt Berber	0.632	0.787	0.700	0.673	0.652	0.792	0.714	0.692	0.511
Thai	0.280	0.782	0.411	0.518	0.288	0.805	0.422	0.527	0.785
Vietnamese	0.191	0.810	0.307	0.338	0.196	0.798	0.314	0.361	0.817
White Hmong	0.175	0.789	0.284	0.387	0.169	0.788	0.277	0.368	0.845
Wichí	0.393	0.805	0.520	0.793	0.444	0.807	0.568	0.831	0.857
Yaqui	0.561	0.805	0.656	0.801	0.546	0.796	0.645	0.791	0.760
Zinacantán Tzotzil	0.382	0.831	0.520	0.783	0.317	0.784	0.450	0.727	0.857
bf Mean	0.315	0.791	0.426	0.593	0.330	0.796	0.440	0.617	0.797

TABLE A.7: Real world borrowing from the World Online Loan Database (WOLD) – 10-fold cross-validation - standard deviations [inherited only methods].

	Neural net - inherited				Markov chain - inherited			
Language	**Prec.**	**Recall**	**F1**	**Acc.**	**Prec.**	**Recall**	**F1**	**Acc.**
Archi	0.043	0.086	0.053	0.025	0.061	0.085	0.056	0.023
Bezhta	0.071	0.089	0.066	0.029	0.037	0.096	0.050	0.038
Ceq Wong	0.047	0.139	0.063	0.030	0.075	0.110	0.077	0.061
Dutch	0.044	0.070	0.054	0.039	0.051	0.071	0.064	0.054
English	0.070	0.069	0.066	0.053	0.074	0.029	0.061	0.047
Gawwada	0.033	0.132	0.052	0.051	0.051	0.186	0.078	0.035
Gurindji	0.029	0.099	0.042	0.044	0.033	0.057	0.050	0.049
Hausa	0.051	0.050	0.059	0.053	0.036	0.062	0.043	0.028
Hawaiian	0.053	0.133	0.076	0.070	0.033	0.081	0.047	0.042
Hup	0.090	0.216	0.128	0.025	0.070	0.117	0.092	0.038
Imbabura Quechua	0.096	0.064	0.054	0.037	0.080	0.084	0.071	0.044
Indonesian	0.049	0.070	0.054	0.033	0.048	0.042	0.047	0.032
Iraqw	0.052	0.128	0.069	0.058	0.056	0.139	0.073	0.047
Japanese	0.027	0.070	0.022	0.023	0.067	0.056	0.063	0.041
Kali'na	0.046	0.129	0.066	0.050	0.063	0.068	0.081	0.063
Kanuri	0.041	0.062	0.051	0.026	0.048	0.068	0.058	0.037
Ket	0.053	0.136	0.075	0.045	0.052	0.109	0.073	0.043
Kildin Saami	0.035	0.056	0.044	0.044	0.021	0.075	0.028	0.021
Lower Sorbian	0.043	0.082	0.056	0.036	0.034	0.079	0.045	0.034
Malagasy	0.034	0.108	0.051	0.030	0.035	0.092	0.045	0.051
Manange	0.049	0.146	0.066	0.039	0.060	0.155	0.083	0.032
Mandarin Chinese	0.009	0.445	0.018	0.058	0.008	0.458	0.016	0.064
Mapudungun	0.056	0.062	0.055	0.033	0.047	0.068	0.047	0.021
Old High German	0.031	0.155	0.053	0.042	0.015	0.130	0.025	0.045
Oroqen	0.046	0.141	0.068	0.065	0.040	0.133	0.053	0.071
Otomi	0.070	0.100	0.048	0.019	0.078	0.087	0.063	0.014
Q'eqchi'	0.074	0.077	0.071	0.029	0.056	0.072	0.050	0.025
Romanian	0.020	0.027	0.019	0.024	0.049	0.054	0.051	0.036
Sakha	0.049	0.049	0.055	0.037	0.044	0.044	0.042	0.042
Saramaccan	0.072	0.047	0.072	0.061	0.057	0.065	0.054	0.047
Selice Romani	0.055	0.041	0.040	0.041	0.049	0.050	0.036	0.035
Seychelles Creole	0.031	0.048	0.045	0.028	0.033	0.160	0.052	0.039
Swahili	0.037	0.062	0.045	0.033	0.045	0.066	0.052	0.021
Takia	0.042	0.077	0.042	0.036	0.032	0.047	0.035	0.032
Tarifiyt Berber	0.050	0.055	0.043	0.032	0.044	0.053	0.037	0.028
Thai	0.028	0.078	0.035	0.043	0.040	0.062	0.046	0.040
Vietnamese	0.031	0.064	0.039	0.028	0.034	0.088	0.049	0.047
White Hmong	0.041	0.077	0.057	0.053	0.027	0.102	0.040	0.041
Wichí	0.085	0.094	0.071	0.041	0.092	0.096	0.086	0.026
Yaqui	0.073	0.104	0.061	0.030	0.064	0.072	0.052	0.028
Zinacantán Tzotzil	0.053	0.098	0.058	0.038	0.049	0.063	0.056	0.030
Mean	0.049	0.098	0.055	0.039	0.048	0.093	0.054	0.039

TABLE A.8: Borrowing and phonological characteristics by language.

				Fraction			Primary donor		
Language	**Words**	**Borrowed**	**Inherited**	**Bor.**	**Bor. sounds**	**Inh. sounds**	**Language**	**Bor.**	**Bor. fraction**
Archi	1254	268	986	0.21	0.02	0.36	Avar	77	0.29
Bezhta	1473	441	1032	0.3	0.03	0.29	Avar	268	0.61
Ceq Wong	956	318	638	0.33	0.05	0.16	Malay	318	1.00
Dutch	1588	295	1293	0.19	0.05	0.02	French	106	0.36
English	1516	585	931	0.39	0.02	0	French	373	0.64
Gawwada	1163	106	1057	0.09	0.05	0.2	Amharic	96	0.91
Gurindji	1028	129	899	0.13	0	0.06	Jaminjung	37	0.29
Hausa	1668	330	1338	0.2	0.07	0.19	Arabic	123	0.37
Hawaiian	1544	248	1296	0.16	0.29	0.14	English	199	0.80
Hup	1179	119	1060	0.1	0.12	0.31	Portuguese	84	0.71
Imbabura Quechua	1319	369	950	0.28	0.06	0.11	Spanish	359	0.97
Indonesian	2049	619	1430	0.3	0.08	0.03	Sanskrit	132	0.21
Iraqw	1262	164	1098	0.13	0.06	0.21	Swahili	146	0.89
Japanese	2131	631	1500	0.3	0.03	0.03	Chinese	494	0.78
Kali'na	1373	199	1174	0.14	0.29	0.15	Sranan	79	0.40
Kanuri	1591	281	1310	0.18	0	0.06	Arabic	142	0.51
Ket	1262	106	1156	0.08	0.05	0.47	Russian	100	0.94
Kildin Saami	1473	280	1193	0.19	0.02	0.29	Russian	176	0.63
Lower Sorbian	1765	348	1417	0.2	0.03	0.05	New High German	217	0.62
Malagasy	1680	210	1470	0.12	0	0.05	French	90	0.43
Manange	1124	73	1051	0.06	0.11	0.33	Nepali	66	0.90
Mandarin Chinese	2130	15	2115	0.01	0	0.38	English	2	0.13
Mapudungun	1412	282	1130	0.2	0.09	0.06	Spanish	261	0.93
Old High German	1258	67	1191	0.05	0	0.25	Latin	61	0.91
Oroqen	1205	94	1111	0.08	0.08	0.18	Chinese	54	0.57
Otomi	2558	251	2307	0.1	0	0.29	Spanish	251	1.00
Q'eqchi'	1995	209	1786	0.1	0.1	0.18	Spanish	201	0.96
Romanian	2270	908	1362	0.4	0.04	0.08	French	283	0.31
Sakha	1588	395	1193	0.25	0.08	0.08	Russian	253	0.64
Saramaccan	1303	462	841	0.35	0	0.02	Suriname Portuguese	242	0.52
Selice Romani	1732	993	739	0.57	0.22	0.1	Hungarian	792	0.80
Seychelles Creole	2089	185	1904	0.09	0.11	0.06	English	83	0.45
Swahili	1830	443	1387	0.24	0.02	0.1	Arabic	331	0.75
Takia	1329	308	1021	0.23	0.07	0.33	Tok Pisin	248	0.81
Tarifiyt Berber	1688	826	862	0.49	0.14	0.02	Arabic (Moroccan)	665	0.81
Thai	2107	454	1653	0.22	0	0.06	Sanskrit	260	0.57
Vietnamese	1534	281	1253	0.18	0	0.1	Chinese	254	0.90
White Hmong	1474	229	1245	0.16	0.01	0.21	Chinese	158	0.69
Wichí	1361	194	1167	0.14	0.14	0.25	Spanish	192	0.99
Yaqui	1615	387	1228	0.24	0.07	0.07	Spanish	386	1.00
Zinacantán Tzotzil	1413	202	1211	0.14	0.15	0.22	Spanish	201	1.00
Mean	1568.0	324.5	1243.5	0.203	0.067	0.160		216.1	0.683

TABLE A.9: Real world borrowing from the World Online Loan Database (WOLD) – 10-fold cross-validation [neural network Transformer module].

	Neural transf. mean				Neural transf. st. dev.				Prop.
Language	**Prec.**	**Recall**	**F1**	**Acc**	**Prec.**	**Recall**	**F1**	**Acc**	**Inher.**
Archi	0.615	0.755	0.673	0.846	0.100	0.065	0.071	0.025	0.786
Bezhta	0.766	0.832	0.796	0.873	0.022	0.071	0.033	0.019	0.701
Ceq Wong	0.729	0.797	0.758	0.832	0.100	0.065	0.066	0.047	0.667
Dutch	0.444	0.622	0.512	0.785	0.069	0.118	0.067	0.025	0.814
English	0.656	0.714	0.682	0.745	0.055	0.053	0.039	0.029	0.614
Gawwada	0.385	0.648	0.468	0.879	0.101	0.207	0.098	0.023	0.909
Gurindji	0.199	0.339	0.244	0.748	0.067	0.142	0.076	0.046	0.875
Hausa	0.533	0.734	0.615	0.820	0.054	0.053	0.040	0.021	0.802
Hawaiian	0.425	0.769	0.542	0.797	0.059	0.086	0.050	0.012	0.839
Hup	0.698	0.891	0.777	0.946	0.107	0.112	0.093	0.030	0.899
Imbabura Quechua	0.776	0.851	0.811	0.889	0.042	0.065	0.049	0.030	0.720
Indonesian	0.636	0.681	0.655	0.784	0.087	0.059	0.064	0.042	0.698
Iraqw	0.573	0.761	0.648	0.896	0.121	0.069	0.098	0.022	0.870
Japanese	0.694	0.824	0.753	0.841	0.029	0.039	0.027	0.015	0.704
Kalina	0.462	0.747	0.567	0.838	0.104	0.096	0.100	0.034	0.855
Kanuri	0.436	0.654	0.513	0.788	0.124	0.088	0.096	0.033	0.823
Ket	0.514	0.768	0.603	0.920	0.127	0.118	0.094	0.016	0.916
Kildin Saami	0.493	0.578	0.530	0.809	0.093	0.136	0.109	0.040	0.810
Lower Sorbian	0.596	0.747	0.658	0.851	0.066	0.101	0.056	0.017	0.803
Malagasy	0.353	0.648	0.456	0.808	0.057	0.117	0.071	0.028	0.875
Manange	0.402	0.650	0.464	0.899	0.204	0.117	0.137	0.043	0.935
Mandarin Chinese	0.050	0.111	0.072	0.973	0.107	0.208	0.140	0.011	0.993
Mapudungun	0.673	0.799	0.726	0.884	0.100	0.073	0.074	0.031	0.800
Old High German	0.359	0.265	0.284	0.927	0.284	0.164	0.172	0.027	0.947
Oroqen	0.330	0.573	0.416	0.876	0.108	0.185	0.133	0.033	0.922
Otomi	0.692	0.899	0.779	0.950	0.060	0.046	0.031	0.010	0.902
Qeqchi	0.679	0.869	0.754	0.941	0.116	0.055	0.067	0.017	0.895
Romanian	0.664	0.733	0.694	0.744	0.051	0.037	0.027	0.017	0.600
Sakha	0.592	0.681	0.631	0.803	0.079	0.063	0.062	0.044	0.751
Saramaccan	0.591	0.642	0.613	0.716	0.053	0.090	0.059	0.034	0.645
Selice Romani	0.900	0.873	0.886	0.871	0.018	0.028	0.019	0.024	0.427
Seychelles Creole	0.280	0.599	0.378	0.831	0.063	0.096	0.071	0.019	0.911
Swahili	0.664	0.799	0.723	0.852	0.058	0.061	0.047	0.031	0.758
Takia	0.580	0.789	0.665	0.814	0.082	0.083	0.070	0.051	0.768
Tarifiyt Berber	0.786	0.826	0.805	0.805	0.033	0.032	0.029	0.028	0.511
Thai	0.544	0.689	0.605	0.806	0.071	0.048	0.051	0.035	0.785
Vietnamese	0.451	0.659	0.532	0.790	0.098	0.082	0.092	0.033	0.817
White Hmong	0.385	0.597	0.466	0.793	0.083	0.112	0.088	0.018	0.845
Wichi	0.703	0.888	0.783	0.934	0.085	0.047	0.067	0.015	0.857
Yaqui	0.719	0.850	0.775	0.885	0.084	0.031	0.051	0.016	0.760
Zinacantan Tzotzil	0.789	0.899	0.836	0.950	0.060	0.098	0.051	0.018	0.857
Main	0.556	0.709	0.613	0.847	0.085	0.088	0.072	0.027	0.797

Appendix B

Multilingual detail results

Multilingual methods consider all languages simultaneously, especially when using cognate based methods. So the norm for performing an analysis, such as a 10-fold cross-validation, is to report on borrowing detection performance over all languages, and not individually. However, to offer a detail comparison of results versus monolingual methods, albeit still with the measurement difference between general lexical borrowing detection versus dominant donor borrowing detection, languages can be analyzed and reported separately. Cognate based methods are not included in these analyses.

Overall results for method where each recipient language is analyzed separately are reported in §3.2.1 and §3.2.3. Detail results for this analysis are reported in this appendix as Tab. B.1. Previously it was noted that performing analyses individually by language resulted in slightly reduced detection performance with F1 score reductions of ≈ 1 point. This is seen here in the mean results by method. Even with the small number of parameters estimated by multilingual method, there seems to be an added cost to performing analysis by language instead of one overall analysis.

TABLE B.1: 10-fold cross-validation by language for detection method. Each language target analyzed separately.

	Method mean			Method st. dev.			Prop.
Language	**Prec.**	**Recall**	**F1**	**Prec.**	**Recall**	**F1**	**Sp. bor.**
Closest Match - SCA							
Imbabura Quechua	0.898	0.779	0.833	0.059	0.054	0.042	0.26
Mapudungun	0.855	0.687	0.757	0.156	0.083	0.103	0.15
Otomi	0.788	0.691	0.730	0.151	0.110	0.112	0.09
Q'eqchi'	0.908	0.651	0.755	0.101	0.081	0.072	0.09
Wichí	0.915	0.710	0.796	0.041	0.116	0.083	0.12
Yaqui	0.859	0.778	0.813	0.083	0.062	0.040	0.22
Zinacantán Tzotzil	0.796	0.654	0.709	0.086	0.141	0.097	0.13
Mean	0.860	0.707	0.770	0.111	0.105	0.090	0.15
Least cross-entropy - LCE							
Imbabura Quechua	0.826	0.800	0.812	0.058	0.073	0.060	0.26
Mapudungun	0.691	0.607	0.639	0.146	0.134	0.120	0.15
Otomi	0.836	0.723	0.772	0.097	0.117	0.098	0.09
Q'eqchi'	0.797	0.742	0.755	0.087	0.162	0.081	0.09
Wichí	0.841	0.740	0.777	0.074	0.148	0.105	0.12
Yaqui	0.811	0.747	0.775	0.081	0.086	0.067	0.22
Zinacantán Tzotzil	0.824	0.769	0.789	0.109	0.097	0.077	0.13
Mean	0.804	0.733	0.760	0.104	0.128	0.100	0.15
Classifier - Linear SVM - NED, SCA							
Imbabura Quechua	0.916	0.769	0.835	0.046	0.067	0.054	0.26
Mapudungun	0.939	0.717	0.812	0.046	0.063	0.047	0.15
Otomi	0.945	0.655	0.765	0.056	0.137	0.095	0.09
Q'eqchi'	0.930	0.643	0.758	0.097	0.088	0.082	0.09
Wichí	0.957	0.660	0.775	0.062	0.134	0.111	0.12
Yaqui	0.933	0.781	0.848	0.057	0.066	0.040	0.22
Zinacantán Tzotzil	0.934	0.653	0.760	0.068	0.139	0.100	0.13
Mean	0.936	0.697	0.793	0.062	0.114	0.084	0.15
Classifier - Linear SVM - NED, SCA, LCE							
Imbabura Quechua	0.884	0.871	0.876	0.051	0.074	0.051	0.26
Mapudungun	0.768	0.719	0.738	0.109	0.103	0.087	0.15
Otomi	0.891	0.824	0.853	0.062	0.104	0.069	0.09
Q'eqchi'	0.887	0.798	0.835	0.103	0.112	0.085	0.09
Wichí	0.883	0.839	0.855	0.082	0.108	0.068	0.12
Yaqui	0.871	0.845	0.855	0.085	0.060	0.047	0.22
Zinacantán Tzotzil	0.897	0.838	0.865	0.086	0.075	0.065	0.13
Mean	0.869	0.819	0.839	0.091	0.100	0.079	0.15

Bibliography

Abadi, Martín et al. (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. URL: https://www.tensorflow.org/.

Anderson, Cormac et al. (2019). "A Cross-Linguistic Database of Phonetic Transcription Systems". In: *Yearbook of the Poznań Linguistic Meeting*, pp. 1–27.

Baayen, R. H., R. Piepenbrock, and L. Gulikers, comp. (1995). *The CELEX Lexical Database*. Philadelphia: Linguistic Data Consortium.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (2015). "Neural Machine Translation by Jointly Learning to Align and Translate". In: *ICLR*. arXiv: 1409.0473. URL: http://arxiv.org/abs/1409.0473.

Baytukalov, Timur (2019). *EasyPronunciation.com*. website: easypronunciation.com. URL: https://easypronunciation.com.

Bender, Emily M. and Alexander Koller (July 2020). "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 5185–5198. DOI: 10.18653/v1/2020.acl-main.463. URL: https://aclanthology.org/2020.acl-main.463.

Bengio, Yoshua et al. (Mar. 2003). "A Neural Probabilistic Language Model". In: *J. Mach. Learn. Res.* 3, pp. 1137–1155. ISSN: 1532-4435. URL: http://dl.acm.org/citation.cfm?id=944919.944966.

Campbell, L. (2013). *Historical Linguistics: An Introduction*. 3rd. Edinburgh University Press.

Carling, Gerd et al. (2019). "The causality of borrowing: Lexical loans in Eurasian languages". In: *PLOS ONE* 14.10, pp. 1–33. DOI: https://doi.org/10.1371/journal.pone.0223588. URL: https://doi.org/10.1371/journal.pone.0223588.

Cristea, Alina Maria et al. (Nov. 2021). "Automatic Discrimination between Inherited and Borrowed Latin Words in Romance Languages". In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 2845–2855.

Cristianini, Nello and John Shawe-Taylor (2000). *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press. DOI: 10.1017/CBO9780511801389.

Delz, Marisa (2014). "Mismatches between phylogenetic trees in Historical Linguistics". MA thesis. Tübingen, Germany: Eberhard-Karls-Universität.

Dessimoz, Christophe, Daniel Margadant, and Gaston H. Gonnet (2008). "DLIGHT – Lateral gene transfer detection using pairwise evolutionary distances in a statistical framework". In: *Research in Computational Molecular Biology*. Ed. by M. Vingron and L. Won. Berlin and Heidelberg: Springer, pp. 315–330.

Devlin, Jacob et al. (2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *CoRR* abs/1810.04805. arXiv: 1810.04805. URL: http://arxiv.org/abs/1810.04805.

Dixon, Wilfrid and Frank Massey, Jr (1983). *Introduction to Statistical Analysis*. 4th. Boston, US: McGraw-Hill.

Dror, Rotem et al. (July 2018). "The Hitchhiker's Guide to Testing Statistical Significance in Natural Language Processing". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 1383–1392. DOI: 10.18653/v1/P18-1128. URL: https://aclanthology.org/P18-1128.

Firth, John Rupert (1957). *A synopsis of linguistic theory 1930-1955*. Special Volume of the Philological Society. Oxford: Oxford University Press.

Forkel, Robert and Johann-Mattis List (2020). "CLDFBench. Give your Cross-Linguistic data a lift". In: *Proceedings of the Twelfth International Conference on Language Resources and Evaluation*. LREC 2020 (Marseille, May 11, 2020). Luxembourg: European Language Resources Association (ELRA), pp. 6997–7004.

Forkel, Robert et al. (Oct. 2018). "Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics". In: *Scientific Data* 5.

Geisler, H. and J.-M. List (2013). "Do languages grow on trees? The tree metaphor in the history of linguistics". In: *Classification and evolution in biology, linguistics and the history of science. Concepts – methods – visualization*. Ed. by Heiner Fangerau et al. Stuttgart: Franz Steiner Verlag, pp. 111–124.

Grant, Anthony (2014). "Lexical borrowing". In: ed. by John R. Taylor. United Kingdom: Oxford University Press, pp. 431–444.

Gray, Russell, Simon Greenhill, and Quentin Atkinson (Nov. 2013). "Phylogenetic Models of Language Change: Three New Questions". In: MIT Press, pp. 285–302. ISBN: 9780262019750. DOI: 10.7551/mitpress/9780262019750.003.0015.

Grossman, Eitan et al. (May 2020). "SegBo: A Database of Borrowed Sounds in the World's Languages". English. In: *Proceedings of The 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 5316–5322. ISBN: 979-10-95546-34-4. URL: https://www.aclweb.org/anthology/2020.lrec-1.654.

Hammarström, Harald, Robert Forkel, and Martin Haspelmath (2021). *Glottolog 4.4*. Max Planck Institute for the Science of Human History.

Hantgan, A, H Babiker, and J List (2022). "First steps towards the detection of contact layers in Bangime: a multi-disciplinary, computer-assisted approach [version 2; peer review: 2 approved]". In: *Open Res Europe* 2.10, pp. 1–25.

Haspelmath, Martin and Uri Tadmor, eds. (June 2009). *Loanwords in the World's Languages, A Comparative Handbook*. PB - De Gruyter Mouton CY - Berlin, Boston.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc.

Heggarty, Paul (June 2014). "Prehistory through language and archaeology". In: ed. by Claire Bowern and Bethwyn Evans. Routledge, pp. 598–626.

Jäger, Gerhard (May 2019). "Computational historical linguistics". In: *Theoretical Linguistics* 45.3-4, pp. 151–182. DOI: doi:10.1515/tl-2019-0011. URL: https://doi.org/10.1515/tl-2019-0011.

Jäger, Gerhard and Johann-Mattis List (2016 approval pending). "Statistical and computational elaborations of the classical comparative method". In: *The Oxford Handbook of Diachronic and Historical Linguistics*. Ed. by P. Crisma and G. Longobardi. Oxford, UK: Oxford University Press.

Jäger, Gerhard, Johann-Mattis List, and Pavel Sofroniev (Apr. 2017). "Using support vector machines and state-of-the-art algorithms for phonetic alignment to identify cognates in multi-lingual wordlists". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Valencia, Spain: Association for Computational Linguistics, pp. 1205–1216. URL: https://www.aclweb.org/anthology/E17-1113.

Jauhiainen, Tommi et al. (May 2019). "Automatic Language Identification in Texts: A Survey". In: *J. Artif. Int. Res.* 65.1, pp. 675–682. ISSN: 1076-9757. DOI: 10.1613/jair.1.11675. URL: https://doi.org/10.1613/jair.1.11675.

JMP®, Version 17.0.0 (2022). Software. Cary, NC, USA.

Jurafsky, Daniel and James H. Martin (2009). *Speech and Language Processing (2nd Edition)*. 3rd edition available at website. Upper Saddle River, NJ, USA: Prentice-Hall, Inc. ISBN: 0131873210. URL: https://web.stanford.edu/~jurafsky/slp3/.

Kaiping, Gereon A. and Marian Klamer (2022). "The dialect chain of the Timor-Alor-Pantar language family". In: *Language Dynamics and Change* 0.0.

Key, Mary Ritchie and Bernard Comrie, eds. (2015). *Intercontinental Dictionary Series (IDS)*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Kiparsky, Paul (June 2014). "New perspectives in historical linguistics". In: *The Routledge Handbook of Historical Linguistics*. Ed. by Claire Bowern and Bethwyn Evans. Routledge, pp. 64–102.

Kluge, Friedrich, ed. (2002). *Etymologisches Wörterbuch der deutschen Sprache*. 24th ed. Berlin: de Gruyter.

Lee, Yeon-Ju and Laurent Sagart (2008). "No limits to borrowing: The case of Bai and Chinese". In: *Diachronica* 25.3, pp. 357–385.

Levenshtein, V. I. (1965). "Dvoičnye kody s ispravleniem vypadenij, vstavok i zameščenij simvolov [Binary codes with correction of deletions, insertions and replacements]". In: *Doklady Akademij Nauk SSSR* 163.4, pp. 845–848.

List, Johann-Mattis (2012). "Multiple sequence alignment in historical linguistics". In: *Proceedings of ConSOLE XIX*. Ed. by Enrico Boone, Kathrin Linke, and Maartje Schulpen, pp. 241–260.

– (2015). "Network Perspectives on Chinese Dialect History: Chances and Challenges". In: *Bulletin of Chinese Linguistics* 8, pp. 27–47. URL: brill.com/bc.

– (2019a). "Automated methods for the investigation of language contact, with a focus on lexical borrowing". In: *Language and Linguistics Compass* 12, pp. 1–16.

– (Mar. 2019b). *Automatic detection of borrowing (Open problems in computational diversity linguistics 2)*. Web blog at: http://phylonetworks.blogspot.com/2019/03/automatic-detection-of-borrowing-open.html.

List, Johann-Mattis and Robert Forkel (2021). *LingPy. A Python library for historical linguistics. Version 2.6.9.* Web page. Software library. With contributions

by Greenhill, Simon, Tresoldi, Tiago, Christoph Rzymski, Gereon Kaiping, Steven Moran, Peter Bouda, Johannes Dellert, Taraka Rama, Frank Nagel. Leizpig: Max Planck Institute for Evolutionary Anthropology. DOI: DOI:https://zenodo.org/badge/latestdoi/5137/lingpy/lingpy. URL: http://lingpy.org.

Automated identification of borrowings in multilingual wordlists [version 3; peer review: 4 approved] (2022) 1:79.

List, Johann-Mattis, Simon J. Greenhill, and Russell D. Gray (Jan. 2017). "The Potential of Automatic Word Comparison for Historical Linguistics". In: *PLOS ONE* 12.1, pp. 1–18.

List, Johann-Mattis et al. (2014a). "Networks of lexical borrowing and lateral gene transfer in language and genome evolution". In: *Bioessays* 36.2, pp. 141–150.

List, Johann-Mattis et al. (2014b). "Using phylogenetic networks to model Chinese dialect history". In: *Language Dynamics and Change* 4.2, pp. 222–252.

List, Johann-Mattis et al. (July 2018). "Sequence comparison in computational historical linguistics". In: *Journal of Language Evolution* 3.2, pp. 130–144.

List, Johann-Mattis et al. (Apr. 2021). *CLTS. Cross-Linguistic Transcription Systems*. Zenodo.

List, Johann Mattis et al., eds. (2022a). *Concepticon. A resource for the linking of concept lists (Version 2.6.0)*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

List, Johann-Mattis et al. (2022b). "Lexibank, A public repository of standardized wordlists with computed phonological and lexical features". In: *Scientific Data* 9.356, pp. 1–16. DOI: https://doi.org/10.1038/s41597-022-01432-0.

Lowder, Matthew W. et al. (June 2018). "Lexical Predictability during Natural Reading: Effects of Surprisal and Entropy Reduction". In: *Cognitive Science* 42, pp. 1166–1183.

Maddieson, Ian (1986). "Borrowed sounds". In: *The Fergusonian Impact: In Honor of Charles A. Ferguson on the Occasion of His 65th Birthday*. Ed. by C.A. Ferguson and J.A. Fishman. v. 1. Berlin: Mouton de Gruyter.

Manning, Christopher D. and Hinirch Schütze (2001). *Foundations of Statistical Natural Langage Processing*. Cambridge, MA, USA: MIT Press.

Markov, Andrey A. (2006). "An Example of Statistical Investigation of the Text Eugene Onegin Concerning the Connection of Samples in Chains". In: *Science in Context* 19, pp. 591–600.

McMahon, April et al. (2005). "Swadesh sublists and the benefits of borrowing: An Andean case study". In: *Transactions of the Philological Society* 103, pp. 147–170.

Meisel, Jürgen M. (2018). "Early child second language acquisition: French gender in German children". In: *Bilingualism: Language and Cognition* 21.4, pp. 656–673. DOI: https://doi.org/10.1017/S1366728916000237.

Mennecier, Philippe et al. (Jan. 2016). "A Central Asian Language Survey: Collecting Data, Measuring Relatedness and Detecting Loans". In: *Language Dynamics and Change* 6, pp. 57–98.

Mi, Chenggang, Lei Xie, and Yanning Zhang (Feb. 2020). "Loanword Identification in Low-Resource Languages with Minimal Supervision". In: *ACM*

Trans. Asian Low-Resour. Lang. Inf. Process. 19.3. ISSN: 2375-4699. DOI: 10.1145/3374212. URL: https://doi.org/10.1145/3374212.

Mi, Chenggang et al. (Oct. 2016). "Recurrent Neural Network Based Loanwords Identification in Uyghur". In: *Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation: Oral Papers*. Seoul, South Korea, pp. 209–217.

Mi, Chenggang et al. (Aug. 2018). "Toward Better Loanword Identification in Uyghur Using Cross-lingual Word Embeddings". In: *Proceedings of CoLing 2018*, pp. 3027–3037.

Miller, John and Johann-Mattis List (May 2023). "Detecting Lexical Borrowings from Dominant Languages in Multilingual Wordlists". In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Dubrovnik, Croatia: Association for Computational Linguistics, pp. 2591–2597. URL: https://aclanthology.org/2023.eacl-main.190.

Miller, John, Emanuel Pariasca, and Cesar Beltran Castañon (Sept. 2021). "Neural Borrowing Detection with Monolingual Lexical Models". In: *Proceedings of the Student Research Workshop Associated with RANLP 2021*. Online: INCOMA Ltd., pp. 109–117. URL: https://aclanthology.org/2021.ranlp-srw.16.

Miller, John E, Tiago Tresoldi, and Johann-Mattis List (n.d.). *PyBor: A Python library for borrowing detection based on lexical language models*. (Visited on 11/04/2020).

Miller, John E. et al. (Dec. 2020). "Using lexical language models to detect borrowings in monolingual wordlists". In: *PLOS ONE* 15.12, pp. 1–23.

Minett, James, Yue Wang, and Hong Kong (Jan. 2003). "On detecting borrowing: Distance-based and character-based approaches". In: *Diachronica* 20. DOI: 10.1075/dia.20.2.04min.

Minitab, LLC (July 2020). *Minitab® Statistical Software, version 19.* Available from minitab.com. MINITAB® and all other trademarks and logos for the Company's products and services are the exclusive property of Minitab, LLC. All other marks referenced remain the property of their respective owners. See minitab.com for more information.

Moran, Steven and Michael Cysouw (June 2018). *The Unicode Cookbook for Linguists: Managing writing systems using orthography profiles*. This is the Language Science Press publication version. Zenodo. DOI: 10.5281/zenodo.1300528.

Moro, Francesca R., Yunus Sulistyono, and Gereon A. Kaiping (2023). "Detecting Papuan Loanwords in Alorese: Combining Quantitative and Qualitative Methods". In: *Traces of Contact in the Lexicon*. Brill, pp. 213–262. DOI: 10.1163/9789004529458_008. URL: https://doi.org/10.1163/9789004529458_008.

Nakhleh, Luay, Don Ringe, and Tandy Warnow (2005). "Perfect Phylogenetic Networks: A new methodology for reconstructing the evolutionary history of natural languages". In: *Language* 81.2, pp. 382–420. JSTOR: 4489897.

Nath, Abhijnan et al. (Oct. 2022). "A Generalized Method for Automated Multilingual Loanword Detection". In: *Proceedings of the 29th International Conference on Computational Linguistics*. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, pp. 4996–5013. URL: https://aclanthology.org/2022.coling-1.442.

Nelson, P. R., P. S. Wludyka, and K. A. F. Copeland (2005). "The Analysis of Means: A Graphical Method for Comparing Means, Rates, and Proportions." In: *SIAM*.

Nelson-Sathi, Shijulal et al. (2011). "Networks uncover hidden lexical borrowing in Indo-European language evolution". In: *Proceedings of the Royal Society of London B: Biological Sciences* 278.1713, pp. 1794–1803. DOI: https://doi.org/10.1098/rspb.2010.1917. URL: http://rspb.royalsocietypublishing.org/content/278/1713/1794.abstract.

Neri (Feb. 2018). *The 8600 Most Frequently Used Spanish Words*. URL: http://frequencylists.blogspot.com/2016/05/the-8600-most-frequently-used-spanish.html (visited on 05/11/2016).

Neureiter, Nico et al. (June 2022). "Detecting contact in language trees: a Bayesian phylogenetic model with horizontal transfer". In: *Humanities and Social Sciences Communications* 9.1, p. 205. DOI: 10.1057/s41599-022-01211-7. URL: https://doi.org/10.1057/s41599-022-01211-7.

Norman, Jerry (1988). *Chinese*. Cambridge: Cambridge University Press.

Pedregosa, F. et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Plato (1921). *Plato in Twelve Volumes*. Vol. 12. translated by Harold N. Fowler. Cambridge, MA/ London: Harvard University Press/William Heinemann Ltd.

Prochazka, Katharina and Gero Vogl (Mar. 2017). "Quantifying the driving factors for language shift in a bilingual region". In: *Proceedings of the National Academy of Sciences* 114.17, pp. 4365–4369. DOI: 10.1073/pnas.1617252114. URL: https://doi.org/10.1073/pnas.1617252114.

Qiguang, Chen (2013). *Miao and Yao language*. Appendix: Hmong-Mien comparative vocabulary list. Beijing, China: China Minzu University Press. URL: https://en.wiktionary.org/wiki/Appendix:Hmong-Mien_comparative_vocabulary_list.

Rzymski, Christoph et al. (Jan. 2020). "The Database of Cross-Linguistic Colexifications, reproducible analysis of cross-linguistic polysemies". In: *Scientific Data* 7.1, p. 13. DOI: 10.1038/s41597-019-0341-x. URL: https://doi.org/10.1038/s41597-019-0341-x.

Shannon, C. E. (Jan. 2001). "A Mathematical Theory of Communication". In: *SIGMOBILE Mob. Comput. Commun. Rev.* 5.1. Original from: 1948, Bell Systems Journal, pp. 3–55. ISSN: 1559-1662. DOI: 10.1145/584091.584093. URL: http://doi.acm.org/10.1145/584091.584093.

Steven Bird, Ewan Klein and Edward Loper (2019). *Natural Language Processing with Python*. O'Reilly Media. URL: http://www.nltk.org/book/.

Sun, Chaofen (2006). *Chinese: A linguistic introduction*. Cambridge: Cambridge University Press.

Swadesh, M. (1952). "Lexico-statistic dating of prehistoric ethnic contacts". In: *Proceedings of the American Philo- sophical Society* 96.4, pp. 452–463.

Team, Google Brain (Mar. 2021). *Trax — Deep Learning with Clear Code and Speed*. URL: https://github.com/google/trax (visited on 03/2022).

Towards a refined wordlist of German in the Intercontinental Dictionary Series (2020). URL: https://calc.hypotheses.org/2545.

Trask, Robert L, ed. (2000). *The dictionary of historical and comparative linguistics*. Edinburgh, Scotland, UK: Edinburgh University Press.

Tresoldi, Tiago, Robert Forkel, and Natalia Morozova (2019). *CLDF dataset derived from Haspelmath and Tadmor's "World Loanword Database" from 2009*. Geneva: Zenodo.

van der Ark, René et al. (2007). "Preliminary identification of language groups and loan words in Central Asia". In: *Proceedings of the RANLP Workshop on Acquisition and Management of Multilingual Lexicons*, pp. 13–20.

Vaswani, Ashish et al. (2017). "Attention Is All You Need". In: *31st Conference on Neural Information Processing Systems (NIPS 2017)*. Vol. abs/1706.03762. arXiv: 1706.03762. URL: http://arxiv.org/abs/1706.03762.

Wada, Takashi et al. (2020). *Learning Contextualised Cross-lingual Word Embeddings for Extremely Low-Resource Languages Using Parallel Corpora*. arXiv: 2010.14649 [cs.CL].

Willems, Matthieu et al. (Sept. 2016). "Using hybridization networks to retrace the evolution of Indo-European languages". In: *BMC Evolutionary Biology* 16.1, p. 180. ISSN: 1471-2148. DOI: 10.1186/s12862-016-0745-6. URL: https://doi.org/10.1186/s12862-016-0745-6.

Wu, M.-S. et al. (2020). "Computer-Assisted Language Comparison: State of the Art". In: *Journal of Open Humanities Data* 6.1.

Zhang, Liqin et al. (2021). "Detecting loan words computationally". In: *Variation rolls the dice. A worldwide collage in honour of Salikoko Mufwene*. Ed. by Enoch Oladé Aboh and Cécile B. Vigoureux. Amsterdam: Benjamins, pp. 269–288.