Nahuatl contemporary writing: studying convergence in the absence of a written norm

by

Jonathan Israel Escobar Farfán

Thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

The University of Sheffield
Faculty of Arts and Humanities
School of Languages and Cultures

March 2019

Abstract

Language revitalisation (LR) is influenced by a concern with authenticity around which written conventions are largely seen as the result of careful designs based on authentic spoken usage. This dissertation proposes to see written conventions as the result of an authentication process carried on by a community of practice of writers, and to explore the points of convergence in this practice as a set of examples which could eventually become shared conventions in most varieties of a linguistic continuum. I focus on the revitalisation of Nahuatl, a linguistic continuum spoken in Mexico. The study of convergence in the written practice of Nahuatl must be carried out in a context of ideological, linguistic, and orthographical heterogeneity. I have tested a methodology to extensively investigate points of convergence between eight contemporary Nahuatl texts from eight contemporary varieties, comparing them with Classical Nahuatl (CN), an old Nahuatl variety codified in prescriptive sources. I have attempted to locate commonalities in these texts by identifying nuclear clauses (NCs): morphosyntactic structures which are a common feature across the Nahuatl continuum. I have used a Finite State (FS) model of CN to attempt a morphological analysis of the word types found in the contemporary texts. The word types that could be plausibly analysed as CN NCs by our FS model, were proposed as plausible points of convergence between the texts and CN. Using a force atlas diagram, each text was represented as a node in a network, with the distance between them being proportional to the number of plausible points of convergence between them. Findings are that the ambiguity of analyses proposed by the developing FS model are currently a pitfall of our approach, but that the plausible points of convergence could be used to locate texts occupying a 'central' position in an expanding network.

Acknowledgements

I want to thank all my family, my clan, for their work and teachings, which have made possible each and every one of my achievements. My deepest gratitude and love to Irina. She has enriched my life with her wisdom, kindness and bravery, and has accompanied and supported me in this adventure.

No words are enough to thank the Wolfson Foundation for their most generous support for my PhD studies. The Foundation made possible projects I would have only dreamed of otherwise.

I would like to thank my supervisors, Neil, Rob and Jane for his guidance and endless support during my studies. I also thank Marta and Dagmar, who always pointed me in the right direction during the early, foggy days of this research. I particularly want to thank Jane and Marta, whose invaluable support helped me begin this adventure.

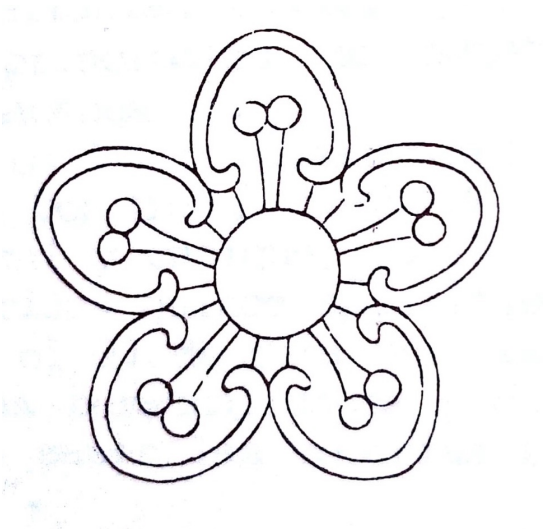
I am also especially grateful to Neil for all the patience with which he received my ideas and discussed them with me; for the generosity with which he shared his knowledge; and for never using the enormous expertise gap between us to make me feel any less worthy as scholar. Every meeting with him helped me see the weakness and strengths of my work, and encouraged both my academic humbleness and boldness.

I am grateful to all the Jessop West community, particularly to Caroline, Claire and Sandra. Their invaluable work always helped me survive in the administrative jungle. From the very first day of my arrival, Caroline's kindness has always been the perfect example of the warmth with which Sheffield welcomed me.

I would like to thank all my friends in Sheffield, who have been my family all these years. I thank all the gang in the 2nd floor PGR work space (Ehsan,

Kathy, Tom, Debbie, and a long etc.), for listening and contributing to my frequent digressions, and for sharing with me our days of pain and glory. Jarek, Ángela, Valentina and Pete, *los quiero mil ocho mil.*

Finally, I am grateful to the city of Sheffield and its people, and to the UK that have so warmly welcomed me. Alongside my beloved Xochimilco, Sheffield occupies now a very special place in my heart.



Contents

List of Figures

List of Tables

Introduction

The use of literacy in language revitalisation (LR) is a source of challenges and contentions, because assumptions about authenticity tend to accentuate the differences found in the spoken varieties of a linguistic continuum and make them a departing point for designing written conventions. This thesis argues that, if the users of a linguistic continuum write texts despite the lack of written conventions, LR actors could also explore the potential commonalities found in written samples from across the continuum, and use them to encourage a supra-local, literate usage. This work proposes a methodology to explore the overlaps between texts written in varieties of a linguistic continuum, Nahuatl, using an available description of a variety of the language (Classical Nahuatl, CN) as pivot for the analysis.

This work attempts to find points of convergence between a number of Nahuatl written varieties, examining the question from the point of view of a central structure: the nuclear clause. As part of this major objective, this work will investigate whether a formal model can point us towards commonalities in written texts that could be used as examples to support and not only regulate the written practice across a number of Nahuatl varieties.

The influence of assumptions about authenticity on LR enterprises presents barriers to moving forward to the strong side of LR, i.e. a stage of LR in which the revitalised language is used in supra-local, literate contexts. An exclusive focus on authenticity and the spoken language as authentic language can fragment LR enterprises. The role of a standard language in the delimitation of 'a language' suggests that a given linguistic continuum could be deconstructed into a collection of 'different languages', each with potentially few users, if written standards are designed to accommodate to as many recognisable 'authentic' spoken varieties within the continuum. Additionally, an exclusive focus on authenticity can confront LR actors by drawing a line between the 'common man'

and the elites, and between their respective 'authentic' and 'artificial' language usage.

LR endeavours rarely occur in a context of absolute illiteracy, and there could be an ongoing written practice despite the lack of widely acknowledged written conventions. Moreover, there could be cases in which texts from older written related varieties are available. It will be argued that written conventions are not only nor necessarily the product of careful designs based on the spoken language, but largely achieved by the practice of writing itself. The practice of writing can be seen as a community of practice, and as a process of authentication of conventions shared by the writers involved. Therefore, an exploration of commonalities between texts produced by writers from a linguistic continuum, if available, should complement the study of spoken local varieties: the points of convergence in the practice of writers could constitute imitable/acceptable examples for literate users across a continuum X, and be the departure for eventual, widespread conventions.

There are antecedents of studies exploring written texts in order to develop written standards. The Euskera Batua, for example, was developed by Koldo Mitxelena in the 1960s based on a selection of forms attested in contemporary literary Basque texts (Hualde & Zuazo, 2007). To the best of my knowledge, however, the extensive, systematic exploration of convergence between written samples from a continuum is not an approach that has been attempted so far to support LR. In addition, this work does not aim to select features, nor to sketch a proposal for a written standard but, following the idea of writing as an authentication practice, to draw attention to plausible commonalities that could be taking advantage of by the writers of the Nahuatl continuum.

There are significant challenges for studying a heterogeneous written practice in a linguistic continuum. Apart from the linguistic differences, there might be considerable orthographical divergence between texts, and a scarcity of resources like grammatical descriptions and extensive dictionaries. In the case of Nahuatl, the lack of widely shared written conventions contrasts with the availability of grammatical descriptions and dictionaries resulting from the study of the 'Classical Nahuatl'. In addition, the morphosyntactic structures called nuclear clauses (NCs) seem a common feature of all Nahuatl varieties.

This work attempted to take advantage of CN and the NCs for an exploration of overlaps between contemporary written texts. As part of this study, a model

of CN using the Finite State (FS) formalism was devekoped. The FS model was used to propose plausible morphological analyses of word types in contemporary texts as CN NCs. In this initial prototype, CN is represented as a set of strings whose formation is determined by a valid concatenation of affixes around a stem. The rules of valid concatenations were coded following a grammatical description of CN (Launey, 2011), and the stems were compiled from a CN dictionary (Wimmer, 2006). One advantage of a FS model for this research is that a combination of alternation rules can be used to map alternative spellings of one same NC to plausible morphological analyses in our CN model. Drawing upon plausible points of convergence between the texts and CN, this work has tried to look at written Nahuatl not as a reflection of neatly delimited varieties, but as an expanding network of texts. In this network, certain texts might have many points in common with most other texts, and thus offer a set of written imitable/acceptable examples for most users across the Nahuatl continuum.

Chapter 1 introduces terminology and key concepts for the discussion of linguistic variation, language planning (LP) and language revitalisation (LR). The role of writing and a standard language in the construction of 'a language' is discussed using the example of the standardisation of Norwegian in the nineteenth century. This chapter explores the influence of4 assumptions about authenticity on LR enterprises and how this focus might hinder a moving forward to the strong side of LR.

Chapter 2 discusses some of the LR efforts focused on Nahuatl. Chapter 2 will discuss the complexity involved in the delimitation of Nahuatl varieties, and in the definition of the concept *Classical Nahuatl*. This chapter also introduces the concept of *nuclear clause*. Afterwards, some perspectives regarding the role of literacy in the revitalisation of Nahuatl will be described. The chapter will present the official policy, represented by the Instituto Nacional de Lenguas Indígenas (INALI), which is to design a number of written standards accommodating to each of the varieties it has recognised by drawing upon studies of the spoken language. The chapter also presents the approach to LR of José Antonio Flores Farfán, an influential sociolinguist and LR activist, and will contrast it with the strategy envisaged by Justyna Olko and John Sullivan. Both projects largely differ in the importance for LR that they give to literacy, written standardisation, CN, and a Nahuatl intelligentsia. It will be argue that Nahuatl users write despite the lack of widely recognised conventions, and the chapter will finally discuss the significance of academic spaces that encourage writing in Nahuatl.

Chapter 3 gives an overview of the resources currently available to attempt a systematic, largescale exploration of the overlaps between contemporary texts using CN as pivot. It is argued that the current orthographic divergence hinders the compilation of large data sets that could be used in a probabilistic approach to analyse NCs, but that a description of CN can be formalised in a FS model.

Chapter 4 describes the methodology followed to explore the overlaps of eight contemporary Nahuatl texts. The test corpus is described, and it is explained how the comprising modules of our CN FS model were created. Finally, it is explained what the different outputs of our FS model might indicate regarding the points of convergence between a contemporary text and CN.

Chapter 5 presents the results obtained after analysing word types from our test texts with the CN model. It firs reports on the results on the performance of our model against CN texts. Afterwards, the points of convergence found between our texts and CN are discussed. Finally, the points of convergence are used to represent graphically the test texts as part of a network where the distance between texts depends on the number of connections they share.

Finally, Chapter 6 presents some conclusions, a summary of current limitations of our approach, and possibilities for future work.

Chapter 1

Authenticity and literacy in LR

1.1 LR, LP and language ideologies

1.1.1 Dialect, language and standardisation

Some key concepts are needed for a discussion involving linguistic variation and the relationship between different sets of linguistic forms. Following Penny (2000), *dialect* will be from now on loosely understood in geographical terms as 'the speech of a particular area'. The term *dialect* will occasionally be used to mean writing that incorporates features common to speakers of a roughly defined geographical area, and arguably specific to them. Sebba (1997) noted how 'dialect' is nearly always used with negative connotations to refer to the communication systems of groups which do not have the status of 'nations'. The term *variety* will be preferred to *dialect*. Wherever the term *dialect* is used in this work it is without any evaluative judgement regarding correctness, or social/political status, and it is because this term is commonly used in the literature in a geographical sense.

This work accepts that in reality, linguistic variation occurs in a *continuum*, i.e. elements change gradually as you move across a territory without any clear dividing points, in such a way that adjacent elements may not be perceptibly different from each other, whereas the extremes may be considerably different from each other. A *dialect continuum*, for example, is a series of locally spoken varieties (see below) in different parts of a territory, in which the speech of each locality differs from its neighbours in some feature or features, until the

accumulated differences hinders the mutual intelligibility of the varieties at the extremes (Penny, 2000:1). It is therefore necessary to bear in mind that a dialect is not a discrete, delimited entity, whether in geographical or social terms (Penny, 2000:10-2). Likewise, Sebba (1997:5) highlights how linguistic variation does not observe regional or state boundaries, and how even today dialect boundaries do not necessarily coincide with national borders.

Linguistic variation does not occur exclusively in a geographical dimension, and a broader conceptualisation is necessary to reflect linguistic variation alongside other correlated social parameters like age, socio-economic status, education, etc. In this dissertation the term *variety* will be used with preference over *dialect* to refer to the set of linguistic items used in a specific set of social circumstances, not only geographical. Besides its negative connotations, *dialect* refers mainly to geographical variation. In contrast, a variety occupies a segment [whose delimitation is largely arbitrary] of a multidimensional 'area' defined along the totality of parameters governing linguistic variation, e.g. spatial, social, temporal, register, etc. (Penny, 2000:19). Following this multidimensional conceptualisation of linguistic variation, the term *language* will be used mainly to refer to a multidimensional cluster of varieties comprising, for example, a dialect continuum, and its related historical and sociolinguistic varieties, which in turn constitute continua in temporal and social dimensions. What must be emphasised is that "each variety, except perhaps the last, shades almost imperceptibly into all neighbouring varieties" (Penny, 2000:19).

There is no satisfactory criterion to distinguish between a 'language' and a 'dialect', nor to delimit a variety as 'a language' from within a continuum. If one variety is differentiated and separated as a 'language' from other dialects in the continuum, is largely as a result of an arbitrary spatial, political delimitation and a fixation in writing of a particular version of the variety in time.

Sebba (1997:1-3) concludes that the real difference between *language* and *dialect* is social: a language is often a dialect with a high status, associated with nationhood and to some extent with European ideas of civilisation. The difference is not linguistic, but the result of people organising in political entities with boundaries and the mechanism to defend them; in such organisation, a 'language' is often just another national asset. As Penny (2000:11-3) illustrates with Spanish, the association of the Castilian variety with a political entity (first the Kingdom of Castile; later on, the Kingdom of Spain and its colonial

domains) was one main force behind the delimitation of the Castilian variety as a separate entity in the *Romance* continuum (the continuum of all oral descendants of Latin) with a given name (*Spanish*).

Anderson's (2006) observations on the development of imagined communities, show that writing, in the specific form that he calls *print-capitalism*, is capable of assembling diverse spoken varieties into 'a language'. He argues that capitalism served and sought to 'assemble' numerous, related vernaculars and idiolects through mechanically reproduced *print-languages*, capable of dissemination trough a vast market. According to Anderson, such print-languages in Europe created unified fields of exchange and communication below Latin, but above the considerable varieties of 'Frenches', 'Englishes' or 'Spanishes'. Print-languages allowed for the development of 'fellowships' of readers, connected via print and paper, who gradually became aware of people in their particular language-field, and of the difference between them and those who did not belong to this particular language-field. Anderson believes that the communities of fellow-readers connected through print-languages created the embryo of the nationally imagined communities (Anderson, 2006:42-4). Thus, print-languages seem very significant motors behind the idea of 'national languages' and the idea of 'nation'. Sebba (1997:8) identifies the 'print-languages' of Anderson with 'standard languages'.

Standardisation is a process ultimately intended to reduce linguistic variation within certain high-prestige varieties. A *standard language* is an ideal, prestigious set of abstract rules, which is conceivable mainly in the written language, and is characterised by a reduced variation of linguistic forms. Unless it is stated otherwise, this work will use the term *standard language* in relation to *written standard languages*. Standardisation takes place mainly within the written language as it cannot in principle directly affect the phonetic and phonological levels of language, and is thus inseparable of the written language (Haugen, 1966; Penny, 2000:193-6).

According to Milroy & Milroy (1999:54-6), although intolerance to optional variability is characteristic of the written medium, this intolerance has been to some extent applied to speech, and the norms of written prose are often held up as models of 'correct' speech, and into prescriptive pronouncements on usage. Thus, Gammelgaard (2002:623) proposes to consider a *spoken standard* "a spoken reflection of written standard language referred to by all members of a mod-

ernised linguistic community, regardless of geographical divisions, in all formal spoken functions and differing from other spoken varieties especially in morphological terms". Gammelgaard accepts that this definition is very rigid, but argues that it can serve as definitional basis for the cases she analysed (Polish, Czech and Slovak), and for other future cases. Standard English is used for Trudgill & Chambers (1991) to make a distinction between mainstream dialects and traditional dialects. They call *mainstream dialects* the varieties of English that are grammatically very close to Standard English, and are reasonably readily mutually intelligible. *Traditional dialects* denote the varieties that differ considerably from Standard English in unsystematic and unpredictable ways. Significantly, some traditional dialects of English may not be readily intelligible to speakers of other dialects. Kerswill (2007) distinguishes in English two sets of norms which he labels as 'mainly written/standard' and 'mainly spoken/non-standard'. The latter incorporates both "informal and dialectal features" (Kerswill, 2007:36). It seems in any case that a standard language is inherently associated with writing, reduced variation and high prestige, all of which are characteristics of major, national languages. In the context of a multilingual state, therefore, standardisation usually involves imposing "majority rules" for the languages of minorities, whose functions, spheres of usage, economic value, etc. might be quite different.

A standard language is more an idea than a reality, a set of abstract rules to which actual usage conforms to a greater or lesser extent; the only fully standardised languages are dead languages as they will not present further variation (Milroy, 2001, 2007, 2012). The idea of standard languages is seen as a practical reality because it is reinforced by official institutions, whether these be language planning bodies like language academies or official language repositories like dictionaries and grammars. The practical reality of a standard language, however, is not only enshrined in formal books and pronouncements, but also looms large in the popular imagination (Linn, 2013:370-3).

Haugen (1966) distinguished four stages in a standardisation process: selection of a norm, codification of form, elaboration of function and acceptance by the community. The reduction of variation expected for a standard language is achieved through *codification*, the prescription of a set of largely unvarying orthographical, grammatical and lexical rules to which writers should conform, if their writing is to carry the highest prestige (Penny, 2000:200). As Penny illustrates with Spanish, a standard language is granted high-prestige largely because its development is typically linked to groups that enjoy enough social

prestige and have enough resources to influence the enterprise. Every standard language grows out of a spoken variety; but the varieties spoken by economically and socially powerful groups[1] are most likely to be selected as the basis for standards (selection). Moreover, these powerful groups are more capable of imposing particular prescriptions (codification) and of ensuring that such codification is used in an increased number of domains (elaboration of function). Finally, the linguistic preferences of high-prestige groups are more likely to be followed by other groups (acceptance).

Being a screenshot of a high prestige spoken variety at one point in time, a standard language will always be out of sync with change in the spoken varieties (Linn, 2013:373). Besides, the reduced range of variants deemed 'acceptable' within a standard language contrasts with the colourful vitality of the numerous choices available in spoken varieties. Standard languages have thus been compared with zombies (Deumert, 2010), as soul-less shells which are essentially dead, but continue to structure our actions and experiences because we treat them as if they were real. Like living dead, standard languages are no longer 'real' but still haunt those who believe in languages as unitary, well defined and countable objects. On a more positive view, whereas standard languages might be zombies in their totality, standard forms are resources available to language users to be combined creatively with all the other words and forms they know and need when interacting (Deumert, 2010:260).

Standardisation is a decisive factor to distinguish a variety in a continuum as a delimited entity, with a given name typically associated with a polity. 'Spanish', for example, came to be associated with the Kingdom of Spain and its dominions. Among other social, geographical, ideological or cultural factors, "language standardisation stands out as crucially important in defining what constitutes a language" (Milroy, 2001:541). According to Milroy, without a process of standardisation, languages are fluid and highly variable; it is not clear where one language ends and where the other begins. Standardisation created the unified fields of exchange and communication between speakers of different varieties, highlighted by Anderson, and at the same time encouraged an idea of the existence of separated, delimited 'language-fields'. The promotion of uniformity in the written usage made speakers of numerous varieties, who might find it

[1]Exceptions to this statement are the standard Slovak developed by L'udovit Štúr in the 1840s, largely based on the dialects of Central Slovakia (Gammelgaard, 2002:615), and thus far from the largest cities to the West, e.g. Bratislava and Trnava; and the Norwegian standard of Ivar Aasen (see section 1.1.2.1)

difficult or even impossible to communicate in conversation, capable to communicate in writing. This 'assembling' of related vernaculars via a print-language, Anderson argues, laid not only the basis of national consciousness, but at the same time made speakers conscious of belonging to a particular 'language-field' (Anderson, 2006:44).

Thus, retaking the example of Penny, another force behind the delimitation of Castilian as 'a language' within the Romance continuum was the development and standardisation of a written code. The fact that at a certain point Castilian achieved written status and underwent a process of standardisation, sharpened speakers' awareness of Castilian as a separate entity requiring a separate name to distinguish it from other written linguistic codes like Latin, or those that eventually got the names Catalan, Portuguese, etc. (Penny, 2000:11-3). Likewise, Sebba (1997:7-8) exemplifies with Dutch and German the influence of standards, compulsory education and universal literacy, on the delimitation of close varieties from a continuum as separated languages. Sebba theorises that, due to the written standards for German and Dutch, speakers in the border regions between the Netherlands and Germany gradually became conscious of speaking either 'German' or 'Dutch', despite the fact that the varieties they spoke were, up to a few generations ago, as similar as two varieties of the same continuum separated by few miles could be.

1.1.2 Language Planning

Standardisation is an example of a process carried out by people intervening in the development of language. This work will refer to *language planning* (LP) as the process of intervention in and nurturing of language according to political and cultural ideology (Linn, 2013:361). LP consists of ideas, policies and practices intended to achieve or stop change in the language used in one or more communities; changing the public's view of a language and its usage range are also main objectives of LP (Kaplan & Baldauf, 1997:3). LP comprises the interrelated activities of corpus planning and status planning. *Corpus planning* consists of "planned efforts to change the lexicon, grammar, phonology and orthography or writing system of a language" (Fishman, 2004:79), and can be therefore considered as comprising those aspects of planning which are primarily linguistic and hence internal to language (Kaplan & Baldauf, 1997:38). *Status planning* involves the planned efforts to change the societal functions of

the language, e.g. governmental, educational, mass media, legal, etc. (Fishman, 2004:79), which are in turn more related to the extra-linguistic context of LP. Besides the standardisation of Norwegian, some examples of LP are the orthographic reforms in Czech (Bermel, 2007) and German (Johnson, 2005).

Standardisation implies the reinforcement of a cultural idea, and can be seen as the story of the triumph of certain values over others, of certain classes of peoples over others (Linn, 2013:370-3). The written standardisation of Norwegian is an example of an enterprise aiming to delimit a variety of the Scandinavian continuum as a separate language through the development of a written standard. The standardisation of Norwegian will be discussed in the next section, as relevant for other LP endeavours. It exemplifies how the values and assumptions –the cult of the uncorrupted common people and of the language they spoke– which were used to differentiate a group of varieties within a linguistic continuum as 'a language' of a community, also triggered an internal confrontation within this community.

1.1.2.1 Norwegian LP: distinguished from the 'exterior', divided at the 'interior'

Perhaps the best known example of LP relates to Norwegian, and the term LP was actually coined by Haugen (1966) in his study of the development of a Norwegian written standard. Jahr (2014) discusses extensively what is resumed in the following paragraphs. During the nineteenth century Norway underwent a series of socio-political changes including a change of political dependence from Denmark to Sweden, the increased influence of National Romanticism on Norway, and the rise of literacy rates by the end of the century. Due to the years of the political union with Denmark, there was a written tradition in Danish, and the upper-middle class spoke a Dano-Norwegian, high-status variety developed by language contact (or variety contact, if one considers the Scandinavian continuum as a whole). A question arose in this context: what was the distinctive language that could distinguish Norwegians as a nation different to the Danes and the Swedes, who spoke nonetheless very close varieties of the same continuum?

The importance given to language in the definition of Norway as a nation was a result of National Romanticism, an influential ideology at the time based

on the works of the German philosopher Johann Gottfried Herder (1744–1803). Herder claimed that it was of the utmost importance for a nation to have its own distinctive language. According to these ideas, language was the most prominent defining feature of a separate and independent country. "National Romanticism held that a nation expresses itself, and its unique and distinctive qualities, through its own language." (Jahr, 2014:24, 44).

Hroch (2007) designs as *National Romanticism* the branch of romantic approaches promoting affiliation to a new community, 'the nation', as a solution for the sense of social alienation and loss of security caused by cultural, social and political convulsions. According to Hroch, although the idea of 'nation' was easily endowed with emotional attractiveness, in the national movements both Enlightenment rationalism and Romantic emotionality were present. The individual features that could potentially distinguish a nation from other, e.g. history, customs, or language, were defined and scholarly studied, and became a subject of academic interest (Hroch, 2007). The Romantic emotional content of National Romanticism includes the cult of language, the idealisation of the past, and the cult of the common people. The search for new stability of Romanticism led to the common people, and to "the idealization of the common man, usually a peasant or countryman, as the vehicle of elementary, universally human, national values" (Hroch, 2007). Hobsbawm agrees with Hroch in that although the folkloric rediscovery of 'the people' did not imply political aspirations like independence, it can be counted as the 'phase A' of the development of many subsequent nationalist movements. Significantly, in this rediscovery of 'the people', the romantic passion for the pure, simple and uncorrupted peasantry gave new importance to the vernaculars they spoke (Hobsbawm, 2002:102-4).

In Norway the goal of having a national language brought along a change of status of the language varieties spoken in Norway. The constitution of 1814, in which Norway was united to Sweden, included a clause stating that the business of the states would be conducted in the 'Norwegian language'. Jahr notices that, the phrasing 'Norwegian language' meant in practice 'not Swedish', since: one, the high-variety used by the upper-middle class for government business was closely connected to Danish, and the dialects of Norway had so far been referred to as Danish dialects, both because of their linguistic closeness to Danish, and of the long union with Denmark; and two, the written standard language used in Norway was essentially Danish, i.e. the same standard used in Denmark,

with features closer to a variety spoken in Denmark than to a variety spoken in Norway.

National romanticism, however, called for a 'Norwegianisation' of the Danish written standard, an initiative advocated by the nationalist poet Henrik Wergeland. National romanticism turned to the spoken Norwegian varieties in order to build a standard expected to fulfil the societal functions that written Danish had covered so far, and Swedish should not cover. Wergeland set the example by using words, idioms and forms taken from spoken Norwegian dialects in his texts. Although opposition to the Norwegianisation of the Danish standard existed (Jahr, 2014:31), the further development of Norwegian LP shows that, at least regarding status planning, there was a fair degree of agreement: one Norwegian (or Norwegianised) variety should occupy the important societal functions in Norway. In this way, a spatial delimitation of a variety within the Scandinavian continuum was to be drawn alongside the borders of the political entity known as Norway.

Regarding corpus planning, specifically the choice of the Norwegian variety to be used as basis for the written standard, the agreement was less clear. National romanticism influenced a sense of opposition between the rural and urban varieties spoken in Norway. A written standard is one main focus of LP, as important societal functions like government and education are largely related to literacy. In the nineteenth century there were two contrasting positions regarding the Norwegian variety on which to develop a Norwegian written standard. On one hand, Knud Knudsen (1812-1895) proposed to expand on the Dano-Norwegian spoken by the upper-middle class; on the other hand, Ivar Aasen (1813-1896) chose the popular dialects of Norwegian peasants as basis for his proposal for a written standard, trying to forge a link to the Norse past. According to Jahr (2014:54), the conflictive LP in Norway during the nineteenth and twentieth century was largely a result of the sociolinguistic differences between the two standards. Bull (2005) distinguishes Aasen's approach to Norwegianisation as abrupt and radical, in comparison to Kundsen's idea of a more gradual Norwegianisation. Since a parliamentary decision in 1929, the standard that emerged from the proposals of Knudsen has been called *Bokmål*, 'book tongue', whereas the standard developed upon the proposals of Aasen has been called *Nynorsk*, 'new Norse'. Both standards have official status, and have undergone reforms in 1917 (both), 1938 (both), 1981 and 2005 (Bokmål), and 2012 (Nynorsk).

National romanticism had thus one significant consequence for Norwegian LP in the nineteenth century: besides creating a sense of Norwegian community, different from the Danish one, it contributed to create an internal division by confronting the Dano-Norwegian creoloid spoken by the upper-middle classes to the dialects of the rural peasants. Aasen considered that to build a Norwegian national awareness, it was necessary to base the written standard in the dialects that, he argued, had evolved directly from Old Norse, without the influence of Danish. Aasen had also the idea of giving all Norwegian dialects the same importance for the development of the written form. However, in the light of his romantic attempt to connect with an idealised past through the language of the 'common' people, Aasen excluded the Dano-Norwegian variety from his programme, even though it was the mother tongue of the upper-middle classes and had undoubtedly developed within the borders of Norway (Jahr, 2014:44-5).

The division caused around the two standards, however, was not necessarily coupled with one of class. An emerging urban working class did not readily accept the idea of Aasen's popular yet 'rural' standard. Likewise, upper-middle classes were not necessarily inclined to the standard of Knudsen, despite it being based on their own sociolect, partly because most of them did not see reasons to reform the current written standard. Early in the century, university academics, for example, had simply suggested calling the written standard shared with Denmark 'Norwegian' whenever it was applied in Norway.

National romanticism, however, seems to have been a major reason behind the general acceptance of a standard. In 1832, the historian P. A. Munch, for example, had rejected Wergeland's idea of altering written Danish. Munch's opinion was that "it would be far better to select an archaic peasant dialect and adapt it in the direction of Old Norse than implement what he regarded as Wergeland's chaotic linguistic programme" (Jahr, 2014:32). This kind of opinion might very well explain why the works of Aasen won the approval of upper-middle classes, which were much more influenced by national romanticism, than they were inclined to the Norwegianisation of the Danish standard.

Accepting the view that only peasants were bearers and saviours of the Norwegian culture and traditions also resonated in later political developments: an agrarian movement in the 1870s and 1880s turned Aasen's programme into a political project directed against the upper-middle class culture, and against the ruling class. It was after this political turmoil that Knudsen's proposal

became more acceptable to the upper-middle classes. By then, however, the standard of Aasen had gained enough support as to be officially accepted alongside Knudsen's by the parliament in the 'Language Equality Resolution' of 1885, a significant landmark of a debate that continues today (see, for example Linn, 2010, 2014; Özerk & Todal, 2013).

The national romanticism aiming to unite a Norwegian nation around a Norwegian language, created at the same time an internal division reflected in the form of two written standards. LP in Norway succeeded in highlighting Norwegian as a separate language from Danish; but the same national romanticism that triggered Norwegian LP, eventually created confronting views respect internal varieties: the 'pure' Norwegian varieties spoken by the Norwegian peasantry, the 'common' people, and the Danish-influenced variety spoken by the urban middle-classes.

1.1.3 Language Revitalisation

It often happens that within a community where two languages coexist, one of these languages begins replacing the other as the main means of communication. Language X is gradually used in fewer domains and learned by fewer people, who instead adopt language Y[2] . The most dramatic outcome of this replacement can be the extinction of the endangered language X in one or all the communities that once used it.

The efforts ultimately intended to counter the extinction of X have been discussed under different names in the literature. Fishman (1990, 1991, 2001a) called the process supported by such efforts Reversing Language Shift (RLS), emphasising that users of a given language X gradually shift to language Y as main language of communication. Among others, Kroskrity (2009) and Crowley (2005), use the term *language renewal*. *Language revival* is a term used, for example, by Bell (2013) when the language focus of the enterprise has no native speakers. Zuckermann & Walsh (2011) , using Hebrew as example, call this last particular circumstance 'language reclamation'. Zuckermann & Walsh follow Amery (2001), who proposed language revival as a superordinate term comprising three RE-terms: *reclamation* (to relearn a language on the basis of

[2]This work will follow the notation employed by Fishman (1991, 2001a). Although it might be true that this algebraic notation unrealistically reduces the complexity of LR, it has the advantage of being brief and avoids judgemental descriptions

historical records), *renewal* (to reinvigorate and extend the remaining body of language in the absence of fluent speakers), and *revitalisation* (to reintroduce the language to younger generations of speakers). Tsunoda (2005), in contrast, uses *revitalisation* as a general term (the activities aimed to restore vitality to a language that has lost or is losing it), which comprises language maintenance (concerning languages that are still alive), and language revival (concerning extinct languages). Grenoble & Whaley (2006), Hinton (2001) and Reyhner & Lockard (2009) also use *language revitalisation* as general term. The term *revitalisation* seems to have gained more exposure in and out of academic circles, as shown by events like the First International Conference on Revitalization of Indigenous and Minoritized Languages (Barcelona/Vic, 2017).

This work will use the term *language revitalisation* (LR) to name the process ultimately intended to counter the extinction of language X by means of the intermediate goals of increasing its number of users and extending its domains of usage. These two intermediate goals make LR part of LP, and more specifically an example of status planning, since status planning seeks to allocate societal resources to encourage the use of a language in more (and more important) societal functions among larger numbers of individuals (Amery, 2001:211; Fishman, 1991:338).

Despite the divergent terminology, all the different efforts to counter the extinction of a language have in common a concern for increasing the number of users of X. Fishman (1990, 1991, 2001a) emphasises the importance of intergenerational transmission of X as a first language at home; but X can also gain users by teaching it as a second language outside the intimate space of the family. This is the case, for example, of the *ikastolas* that have been a significant force behind the recovery of Basque. *Ikastolas* began as schools where Basque-speaking children were taught in Basque, and Spanish-speaking children have an early immersion in Basque (López-Goñi, 2003).

The extension of domains of usage of X is the other significant concern of LR efforts, although this goal can of course be proportionate to specific ambitions and circumstances. Individuals acquiring X need opportunities to use it, whether they learn it at home or not. It is therefore of the highest importance to extend the usage of X to as many activities as possible[3], and thus create an expanding (or at least stable) community of users. It can be said in general that extending

[3] *Ikastolas* again are notable for not only teaching Basque, but instead teaching in Basque (López-Goñi, 2003:674). This strategy did not only provided students an opportunity of using

STAGES OF REVERSING LANGUAGE SHIFT
SEVERITY OF INTERGENERATIONAL DISLOCATION
(read from bottom up)

1. Education, work sphere, mass media and governmental operations at higher and nationwide levels.
2. Local/regional mass media and governmental services.
3. The local/regional (i.e. non-neighbourhood) work sphere, both among Xmen and among Ymen.

4b. Public schools for Xish children, offering some instruction via Xish, but substantially under Yish curricular and staffing control.

4a. Schools in lieu of compulsory education and substantially under Xish curricular and staffing control.

II. RLS to transcend diglossia, subsequent to its attainment

5. Schools for literacy acquisition, for the old and for the young, and not in lieu of compulsory education.
6. **The intergenerational and demographically concentrated home-family-neighbourhood-community: the basis of mother-tongue transmission.**
7. Cultural interaction in Xish primarily involving the community-based older generation
8. Reconstructing Xish and adult acquisition of XSL

I. RLS to attain diglossia, assuming (prior ideological clarification)

Figure 1.1 The GIDS scale. The Graded Intergenerational Disruption Scale (GIDS) was proposed by Fishman (2001a)

the usage of X across an ideally increasing community is one key concern of LR. The GIDS scale of *Fishman* (Figure 1.1) exemplifies this concern, and shows how LR efforts could ideally advance towards a 'strong side', i.e. the stages of LR where X is used in important societal functions (formal education, government, etc.) in larger communities (e.g. at regional or national levels).

1.1.3.1 Towards the strong side of LR

Fishman (1991) conceptualised LR as a process that is better approached in stages, which he arranged in a Graded Intergenerational Disruption Scale (GIDS). The eight stages of the GIDS are intended to help LR actors to locate the functional disruption of a language in the social space, and to evaluate suitable departing points and priorities. In Fishman's idea, LR enterprises must not necessarily proceed in a step-by-step progression, and multistage efforts are not rejected. However, the main priority of LR, according to Fishman, must always by securing the intergenerational transmission that occurs in face-to-face interaction at home, with the family. The fulcrum of the GIDS scale is thus usually stage 6, although multistage efforts are not discouraged by Fishman (2001b:467) as long as the intergenerational transmission stays the main concern of LR.

the language, but at the same time extended the usage of Basque to a significant societal activity like education.

As the GIDS model shows, LR enterprises could eventually aim to transit from a family-community scope to the sphere of the regional, national and so on. In GIDS terms, LR enterprises might want to move from the weak side (stages 8-5) to the strong (stages 4-1). As Fishman argues, the intimate level of usage (stage 6) is of great importance. Nevertheless, the strong side of the GIDS scale also recognises that LR efforts might benefit by gradually widening its scope to regional, literate levels.

As Fishman's GIDS scale shows the extension of domains of usage can be understood both in functional and geographical terms. Grenoble & Whaley, for example, call Y – the language gaining speakers – the *language of wider communication*, since it is typically used by a larger number of speakers and in a broader range of domains. For X – the language losing speakers – they use the term *local language*. With this terminology they seek to highlight that "revitalisation is tied to a particular geography, and that the people involved in revitalisation desire that the language be more widely used in this particular location" (Grenoble & Whaley, 2006:14). They do not elaborate on the geographical extension of the particular location, but this work will consider that revitalisation, understood in terms of its goals of extending domains of usage and number of speakers, is tied to small localities only in principle. As Fishman's GIDS scale hints, a revitalised language should ideally become a language of *geographically-wider* communication soon or later, for example, by gradually extending over a number of small localities speaking varieties of a linguistic continuum.

Some continua like Quechua, whose speakers inhabit a large geographical area, offer LR projects the chance to aim for a geographically wide usage of the language. Varieties of the Quechua continuum, for example, are spoken in five Andean countries, and in combination have an estimated 7,734,620 speakers (Simons & Fennig, 2018). LR projects focusing on such a continuum as a whole might have an increased opportunity for using the language within an extended community.

Broad regional LR efforts might offer the opportunity of working at multiple GIDS stages simultaneously. Hornberger & King (2001), for example, argue that Quechua LR should see beyond the local reality of specific Quechua communities; Quechua LR should also take into account the multinational character of Quechua, and the evidence of its worldwide extension thanks to the electronic media and courses of study at higher education institutions. Due to

this regional, multinational potential, it is possible to carry on programmes like the *Programa de Formación en Educación Intercultural Bilingüe para los Países Andinos* 'Andean Programme in Bilingual Intercultural Education' (PROEIB-Andes). According to Hornberger & King, programmes like PROEIB-Andes are significant for LR because they address top and bottom GIDS stages concurrently. For example, PROEIB-Andes promotes the use of Quechua in higher education (stage 1), facilitates the cultural interaction in Quechua (stage 7), and supports the process of reconstruction of the language and adult acquisition (stage 8). (Hornberger & King, 2001:185). Thus, this kind of supra-local programmes complement local efforts focused on stage 6, which might concentrate on few specific Quechua communities only.

One important reason to support supra-local programmes like PROEIB-Andes is that they are an opportunity of contact between speakers from different varieties, and help them engage in a context – academic in this case – that promotes usages of the language different to the day-to-day interactions that occur at local level. This formal linguistic socialisation (e.g. in stage 5) is important for LR because it adds additional varieties to the learner's repertoire, and gives X a range of functions that are normally available only in Y (Fishman, 1990:21).

Focusing on a whole continuum offers LR an opportunity of adding up the efforts and resources of local LR enterprises; it entails at the same time managing accumulated linguistic differences between varieties, and facing the ideological interpretations of these differences. How many localities can or should be grouped to revitalise their language together? The answer seems obvious: all the localities who speak one same language. The example of Norwegian, however, shows that what constitutes one same language can be open to interpretation in LP. Closely related 'languages' form part of a continuum, and any dividing line cutting through this continuum is the result of political rather than purely linguistic reasons (Penny, 2000:19). Moreover, two varieties from the same continuum can be delimited ideologically as two languages: national romanticism guided the establishment of Norwegian as a separate language from Danish, despite the close relationship between them.

1.1.4 Language ideologies: the representation of differences and making of groups

Broadly speaking, ideology is the set of beliefs or opinions of a group or an individual. A language ideology can be understood as "sets of beliefs about language articulated by users as rationalisations or justifications of perceived language structure and use" (Silverstein, 1979). For Woolard (1998) language ideologies are representations, explicit or implicit, that construe the interaction of language and human beings in a social world.

Ideologies are important for LR because they might draw upon differences to make and unmake linguistic groups: can/should a locality be part of the LR of X, or the language in this locality is rather a different language that should have its own LR programme? The answer might depend on the way linguistic ideologies focus on linguistic differentiation, i.e. the formation of languages and dialects, as this focus determines the creation of social and linguistic boundaries and hierarchies (Gal & Irvine, 1995). Description of languages can be motivated by notions of their distinctness, and thus their differences be highlighted and variation and overlap ignored to differentiate one from the other (Gal & Irvine, 1995:973). Linguistic descriptions can also minimise or ignore differences to argue for a linguistic kinship that can justify political unity and territorial expansion. This was the case with the linguistic descriptions of Macedonian dialects by Serbian and Bulgarian linguists by the end of the nineteenth century. In Serbian and Bulgarian arguments, linguistic kinship of Macedonian dialects with either Serbian or Bulgarian was identified by selecting some linguistic features and ignoring others (Gal & Irvine, 1995:984).

Ideologies are also important for LR because they might encourage the assumption that two varieties from a cluster are essentially different and in confrontation. The ideology of the standard (IOS) and the ideology of the dialect (IOD) are two examples of contrasting ideologies that use differences to represent varieties of the same linguistic cluster as essentially different and in a hierarchy. The IOS and IOD draw upon and foster contrasting perceptions on the importance of either supra-regional standards or local dialects. The ideology of the standard (Milroy, 2001, 2007)(Milroy, 2001, 2007) includes the belief in the existence of a canonical form of language against which all linguistic behaviour is compared for correctness. The ideology of the dialect (Watts, 1999) includes the conviction that a local variety, e.g. a Swiss German dialect, is the

'mother tongue' of a speaker, and has a higher value than a standardised written variety; besides, the IOD promotes the idea that a written standard, in this case Standard German, is the first 'foreign language', distant from the local dialect which is the language of proximity, everyday organisation and emotional affect. This perception of distance is reproduced in the education system, and promoted in the classroom through practice materials and exercises often framed to raise awareness about, and highlight the differences between the dialect and the standard language (Watts, 1999:91). The IOS and the IOD show how linguistic ideologies can use differences to reinforce delimitations of 'languages' and social groups. Whereas the IOS assumes uniformity and negatively judges variation, the IOD takes variation to highlight the closeness of the local spoken language against the foreignness of the standard.

The IOS and the IOD exemplify how through ideological points of view the varieties of a cluster are valued in relation to others. Varieties are labelled as either correct or incorrect by the IOS, or as either native or foreign by the IOD. In the case of the IOD, the evaluation of varieties is made in terms of foreignness, in a way that shows preference for the local 'mother tongue'. The IOD thus creates those images of "self/other" or "us/them" that are likely to be found in language ideologies (Schieffelin & Charlier Doucet, 1998:286). Ideologies like the IOS and the IOD can motivate different answers to the question: could/should the standard variety (i.e. the written, prestigious, fairly uniform set of linguistic forms) of locality X_a be used at locality X_b?

Finally, ideologies also represent language varieties as static, closed entities. Thus the IOS and the IOD overlook, for example, that variants now considered correct or native could have once been seen as incorrect or foreign (see e.g. Milroy & Milroy, 1999:17), or that low-prestige features used by 'uneducated people' can be gradually adapted by those at 'higher' levels in the socio-economic continuum and vice versa (Penny, 2000:68-9).

1.1.4.1 Language ideologies and LR: groups and confrontations

In the context of LR, ideologies and their representation of differences can guide a regrouping of the varieties of a continuum. Costa & Gasquet-Cyrus (2013) argue that the LR is after all a struggle over classifications, i.e. over the making and unmaking of groups, and over the power to impose these divisions on the social

world. They support this observation with the example of Provençal, a variety of the Southern-France continuum known as Occitan. LR actors have elaborated the distinctive features of Provençal to establish it as a separate language, and have thus created an internal competition of LR within the Occitan continuum (Costa & Gasquet-Cyrus, 2013). The most hope-giving example of LR is a coordinated, unified enterprise encompassing a continuum as 'a language', albeit recognising the internal variation within it, e.g. in Basque (Azurmendi et al., 2001). In extreme cases, LR actors can choose to elaborate the differences and begin a competition to establish one variety from a continuum as a separate language, as is the case of Provençal (Costa & Gasquet-Cyrus, 2013), thus dividing an already small community of speakers.

Whereas Fishman (2004) delineated the importance of recognising language ideologies underlying corpus planning, he conceptualised his stages for LR assuming a 'prior ideological clarification' (Fishman, 1990, 1991). In practice, however, LR can be seen not only as a site of ideological struggle (Kroskrity, 2009), but also as a competition over symbolic power, resources and even material benefit (Coronel-Molina, 2008; Costa & Gasquet-Cyrus, 2013; Flores Farfán, 2017; Lagos et al., 2013).

Fishman argued that most LR efforts are not confrontational in any real sense; at least not regarding a confrontation between the threatened X and the unthreatened Y, as LR efforts seek support in both the X and Y community (Fishman, 2001b:454-5). Likewise, Tulloch (2006) considers that macro goals like developing a written standard usable all along the continuum X, and micro goals like the preservation of dialectal features can be pursued as concurrent and complementary goals. However, examples from Quechua (Coronel-Molina, 2008), Mapuche (Lagos et al., 2013) and Nahuatl (Flores Farfán, 2017) show that there are often contrasting visions around the idea of a standard for a continuum. In these discussions the potential confrontation is not only between geographical varieties of the continuum, but, as in the Norwegian case, between social varieties. The promotion of literacy might confront one idealised higher variety cultivated by an indigenous intellectuality, with the 'real' language of rural speakers. Flores Farfán, for example, criticises the texts written by an emerging Nahuatl intelligentsia. In these texts, he argues, the combination of 'Classical' Nahuatl (see section 2.1.2) lexical forms, e.g. the numbers, with contemporary dialectal forms creates "*una variedad* ***únicamente escrita*** *del náhuatl que resulta altamente artificial e ininteligible*", 'a **written-only** Nahu-

atl variety which is highly artificial and unintelligible' (Flores Farfán, 2009:93-4). Although Flores Farfán's criticism is partially justified, such claims that the common Nahuatl speaker cannot access or understand the written forms taken from 'Classical' Nahuatl sources nurtures a discourse that draws a strict, impermeable line between two varieties of a cluster: the artificial language of the intellectual and the language of the lay man.

LR thus often involves not only an effort to promote the usage of X in relation to Y, but also a confrontation between varieties, geographical or social, of X.

1.1.5 Authenticity in LR

As it has been mentioned (section 1.1.2.1), national romanticism was a significant trigger for the development of a Norwegian standard, and a significant influence throughout the process. Romantic ideology was predominant during the 19_{th} century in Europe, and influenced also, for example, the development of a Ukrainian standard (Yavorska, 2010). In a similar way, an ideology of authenticity seems to influence LR of many languages just like national romanticism influenced Norwegian standardisation in the 19_{th} century.

By promoting allegiance to a new community, 'the nation', national romanticism tried to offer a solution for the sense of social alienation and insecurity caused by cultural, social and political convulsions. Likewise, in a context characterised by global interactions, the tension between cultural homogenisation and cultural heterogenisation became a central problem (Appadurai, 1990), and the question of belonging has regained relevance. In the contemporary reconceptualisation of belonging, language can also play a significant role, as it did in national romanticism. Language varieties can nowadays be used to denote authenticity, understood as 'being from somewhere', in opposition to the anonymity of 'being from nowhere' (O'Rourke, 2015). It might therefore be that concerns regarding the homogenising influence of globalisation have played a role in the increased interest on LR, as this triggers reactions reasserting unique cultural identities (Grenoble & Whaley, 2006:2-3). The "postmodern cult of authenticity", can be seen as a protectionist strategy against globalisation processes (Yavorska, 2010:192).

The discussions around ideas of the 'authentic' and 'authenticity' are recurrent in the literature on LR and LP (Adkins, 2013; Hornberger & King, 1998;

Hornsby, 2005; Hornsby & Quentel, 2013; O'Rourke, 2015; Sallabank, 2010; Schieffelin & Charlier Doucet, 1998; Urla et al., 2016; Weinberg & De Korne, 2016). These works report on how in the discourse of LR actors the labels "authenticity" and "authentic" are used not only to classify and evaluate linguistic varieties, but also to question the right of actors to contribute towards political issues, including LR. The urban population, for example, can be scrutinised as insufficiently 'indigenous' and therefore might have to struggle more to be recognised as legitimate actors in LR (Shulist, 2016). Authenticity, understood as a claim to a specific ancestry, can also be flagged as an indication of the credibility and reliability of actors involved in language- or culture-related enterprises (Matras, 2015).

The idea of authenticity in LR was discussed early on by (Fishman, 1990) who considered LR an ethnolinguistic authenticity movement (Fishman, 1991:397). Fishman is perhaps the most influential author in the LR field, and also the one who makes 'authenticity' a central concept of LR theory. Fishman drew upon the view "that humans can be themselves ('be authentic') only if they live without any imposed social structure at all", a central idea to Western social thinkers since the mid-nineteenth century (Fishman, 1990:13). He argued that the pursuit of authenticity stresses an aspiration of either individuals or groups to achieve self-regulatory status, thus promoting authenticity as the liberation of either individuals or groups from an imposed, repressive collectivity.

As with other ideological discussions, issues around authenticity in LR thus relate significantly to the making and unmaking of groups based on the approach to differences. Although Fishman accepted the changeability and intersectionality of all cultures (Fishman, 1991:393), his theory of LR is rooted in the idea of ethnolinguistic authenticity and identity, which is in turn based on emphasising difference: "it is only the conviction that one's own-language-in-culture is crucially different and, therefore, worth sacrificing for ('*vive la difference!*') that makes RLS [Reversing Language Shift] worthwhile" (Fishman, 2001a:17). In Fishman's idea, thus, LR largely implies the preservation of the ethnolinguistic differences that delineate a group as distinct from others: LR is a way for the speakers of X to remain authentic, i.e. as free as possible from an imposed and therefore artificial collectivity like the one constructed via Y.

One important short-coming of putting authenticity at the centre of LR is that the search for authenticity can increase an awareness of the differences

with Y, but also of the differences between the varieties of the continuum X. It might happen that those internal differences are emphasised and used to claim the right of each variety X_n to be considered authentic and separate from other varieties of X, as the example of Provençal show s. The 'ethnolinguistic authenticity movement' can turn from an effort to reinforce X in relation to Y into a competition between 'authentic' varieties of X, e.g. X_a, X_b,... X_n. Competition between LR enterprises promoting different varieties within the same continuum as different languages is not an unlikely outcome: the promoters of competing enterprises can always elaborate on academic arguments rejecting the continuum as 'a language', and presenting it instead as a group of languages 'whose number is yet to be determined', e.g. in the case of Nahuatl (Flores Farfán, 2011a:268); or Quechua (Adelaar & Muysken, 2004:168). Far from questioning the legitimacy or motives for regrouping a linguistic continuum, what must be emphasised is one potential consequence of LR strategies concerned with authenticity: an already small community of users of X can be further divided in order not to disrupt the perceived authenticity of each variety X_n.

Another drawback of authenticity-concerned LR is that it fosters the representation of the communities of users and their linguistic repertoire as closed and unchanging. Concerns with authenticity in LR often result in the rejection of mixed varieties or emerging variations that diverge from the language of 'traditional' speakers. One has to consider, however, that the result of LR can be an amalgam of what new generations learn from the last speakers of a language, who often come from different dialectal backgrounds and have differences in speech (Hinton, 2001); or that new speakers might have appropriated the language not only at home, but also via an amalgamated standard variety that differs from the more informal speech of local dialects, as is the case of Basque new speakers (Urla et al., 2016). In the case of Irish, forms used by 'new speakers' increasingly diverge from traditional local vernaculars. Irish post-traditional forms thus challenge the ideologies that, in the 1950s, promoted the speech of the remaining Irish-speaking communities, the Gaeltacht, as the source of traditional and authentic practices to codify standard Irish (Ó Murchadha, 2016).

It has been suggested that LR could be better conceived as a transformative rather than restorative enterprise (Hornberger & King, 1998) whose results can foster forms of language that are not exactly equal to previous forms of it. This perspective is particularly true in the case of small scattered communities of speakers of close varieties. If the language of such communities as a whole

stands a chance of surviving, it is in the new forms of language that might arise from the collective recreation of the linguistic baggage which each community contributes to a LR effort (see sections 1.3.1 and 1.3.2.2). In addition, it has been noted that 'authenticity' can be achieved also through the expression of traditional values, and not only through linguistic form and vocabulary which might have undergone considerable change, and are therefore not entirely 'traditional' (Hinton & Ahlers, 1999).

Conflicts about authenticity in LR, however, frequently emerge around linguistic forms deviating from the forms promoted as traditional. Based on the cases of Corsican and Guernsey, for example, Sallabank (2010) observes how deep-rooted ideologies regarding 'authenticity' can be. Mixtures of varieties of Corsican or of varieties of Guernsey, or of them with French are not readily accepted (Sallabank, 2010:318-23). Besides, potential changes to the 'authentic' language to develop a standard variety for teaching younger learners, are rejected by Guernsey older speakers, to the point that fear of language change seems often greater than fear of language death (Marquis & Sallabank, 2013:179). The neo-Breton, i.e. the language used in literature, academy and learned in school, is contested as it differs from traditional varieties of spoken Breton (Adkins, 2013; Hornsby, 2005; Hornsby & Quentel, 2013). The notion of authenticity in LR is thus often paired with an idea of language varieties as closed, static repertoires, and it recognises varieties as authentic only when they remain close to traditional local speech, or free from what is perceived as external influences.

1.1.5.1 Spoken language as authentic language in LR

Ideas around authenticity in LR point to a certain pre-eminence of the spoken language over the written language. The paradigm guiding theoretical linguistics for the last eighty years often encourages a perception of the spoken language as the authentic, natural, real language. According to Coulmas (2013:1-6), the priority of the spoken over the written language is generally accepted in modern linguistics, largely as result of the view of Saussure and Bloomfield, who conceptualised writing as a distorted visible image of the abstract inner system of language. In such conceptualisation, the study of writing is neglected, as linguistics should study *natural* language, i.e. the inborn capacity of language.

Sociolinguistics is arguably the most influential field of linguistics on the LR literature, and it valorises the spoken language as source of empirical data. Significantly, Bucholtz (2003) noticed that 'authentic language', 'authentic contexts' and 'authentic speaker' are notions underlying nearly every aspect of much sociolinguistics work, to the point that she identified authenticity-related ideologies in sociolinguistics. *Linguistic isolationism* and *linguistic mundaneness* are two ideologies concerning authentic speakers and authentic language in sociolinguistics. Although these ideologies are not shared by all sociolinguists, she clarifies, they are common enough to have influenced sociolinguistic theory and practice, and to have created the construct of the authentic speaker.

According to *isolationism* the most authentic language is removed and unaffected by other influences, and the most authentic speaker belongs to a well-defined, static, relatively homogeneous, closed social group. According to *mundaneness*, the most authentic language is language that, from its user's point of view, is unremarkable, commonplace, everyday language. Thus isolationism and mundaneness seem to underlie the rejection of mixed varieties in LR, and the choice of rural spoken varieties as models of codification in LR. Language revitalisationists may favour 'authentic' varieties spoken in rural areas, which might be less polluted by the contact with other languages or mixed urban dialects (Sallabank, 2010:314), as it happened with the codification of Irish based on the Gaeltacht. As Quechua (Coronel-Molina, 2008), Mapuche (Lagos et al., 2013) and Nahuatl (Flores Farfán, 2017) examples show, mundaneness seems often to be the counterpart of the purism underlying the practices of an indigenous intelligentsia. Intellectuals' attempts to cultivate in writing a 'high' language (which they, in turn, may consider the 'authentic' or 'pure' one), are confronted with calls to consult 'common' speakers as the best sources of linguistic knowledge for taking decisions (Coronel-Molina, 2008:335), thus valorising the language practices and points of view of indigenous 'laymen'.

The influence of isolationism and mundaneness on LR can be seen in the strategies privileging the 'common man' and its language, in a similar way that national romanticism favoured the cult of the uncorrupted peasantry and the vernaculars they spoke. In Norwegian LP, the supporters of national romanticism challenged the legitimacy of the language of the urban middle class, since it was too influenced by the Danish 'exterior'. Likewise, in LR, activists influenced by isolationism and mundaneness may contend that the vision and usage of language of an urban middle-class or an intelligentsia are not indigenous or local

enough, since these groups may be too influenced by 'external' practices or values. The way in which assumptions about authenticity are used in LR to draw a line between the local layman and other groups might confront rather than coordinate a potential community of language users. As the case of Provencal shows, local endeavours claiming to emanate from 'the people' can compete with the activism of an urban middle-class, that the local activists see as illegitimate and foreign (Costa & Gasquet-Cyrus, 2013:218-20).

According to Bucholtz, the concern with authenticity that often informs studies of language in use restricts the definition of "real language" to increasingly narrow subsets of all language use. Notably, most sociolinguistic traditions share such a strong preference for spoken over written language that *speaker* is synonymous with *language user* in many sociolinguistic studies (Bucholtz, 2003:405-6). Coulmas (2013:8-10) notices that in most of sociolinguistics the condemnation of the tyranny of writing has been replaced by the dominance of the vernacular. This might help to understand why in approaches to LP like *polynomy* (see below), the need of a written standard, i.e. a written model of reduced variation, is questioned and rejected; but the unmixed language of traditional *speakers*, e.g. the Irish Gaeltacht, is seen as a source of authentic linguistic practices.

1.2 Literacy and authenticity in LR: spoken local varieties as model for literacy

1.2.1 Progressing up the GIDS scale: literacy and diglossia

Progressing in the GIDS scale suggests the gradual usage of X in special functions like formal education and government, and thus implies the configuration of an X variety used for these specialised functions. Gaining geographically-broader and more prestigious usage for X soon or later brings about the question of literacy and a high-prestige variety of the language.

Literacy could sometimes be considered as going against the traditional culture LR is expected to support. Mühlhäusler (1996:212-40) considered the introduction of literacy as one first intrusion with the traditional cultural practices of

communities in the islands of the Pacific. In a similar view, in a LR programme in Cochiti, New Mexico, for example, it was initially feared that writing the Keres language could bring unwanted changes to secular and religious traditions (Pecos & Blum-Martinez, 2001). In relation to Mühlhäusler's observations about the languages of the Pacific, Crowley (2000), however, contended that literacy has been incorporated into many local communities, and is indeed now part of their cultural baggage; thus the language usage in the Pacific does not occur in a simple two-way dichotomy between orality and literacy. Crowley disagrees with Mühlhäusler's claim that vernacular literacy inevitably leads to transitional literacy in, and ultimate replacement by a metropolitan language. The sudden imposition of literacy only in a metropolitan language, Crowley argues, combined with the lack of development of a vernacular literacy is what makes people doubt the value of their own language (Crowley, 2000:381-4).

If literacy might not be essential in early stages of LR, it does seem unavoidable to progress in the GIDS scale, mainly because literacy facilitates trans-regional usage. Once local LR enterprises gains support, how can users of varieties of X make contact if they are spread widely across regions? The case of Romani communities shows that, as a complement to the options offered by audio-visual media, literacy is useful in linking up dispersed communities which have no territorial centre or 'home' region. An emerging Romani literacy is becoming a cohesive element among dispersed communities and individuals, who rely on basic literacy skills to expand the usage of Romani into the trans-national domain constituted by the online social media (Matras, 2015).

The need for X to fulfil special functions like education might eventually entail a potential *diglossia*. In the original definition of the term (Ferguson, 1959), diglossia is the coexistence of two varieties of a language, distinguished by a specialisation of function for each: one H ('high') variety is used in a set of typically formal situations, e.g. a sermon in church or mosque, a political speech or a university lecture; in contrast, one or more regional varieties, called collectively the L ('low') variety, are used in typically informal situations like a conversation with family, friends or colleagues. In a diglossic situation, the H variety diverges considerably from L, and although it is used for most written and formal spoken purposes, it is not used by any sector of the community for ordinary conversation. Fishman's model considers the attainment of diglossia between X and Y (Fishman, 1991:400) as one desirable early milestone in LR. The diglossia referred to in GIDS is mainly of the type X-Y, where a first goal

is to ensure that X is steadily used as the L variety within X communities, and Y as H. After the attainment of diglossia between X and Y, the GIDS model suggests, the next milestone is to transcend this diglossia by transferring the functions of the H variety from Y to X.

A diglossia closer to the original definition of Ferguson, i.e. one between varieties of the same continuum (of the type X_L-X_H), is a central issue in LR. Ferguson noted three trends whose appearance make members of a community regard diglossia as a 'problem'. One is a trend toward more widespread literacy; another is a trend toward broader communication between regional and social segments of the community; and finally, a growing desire for a standard 'national' language as an attribute of autonomy or of sovereignty (Ferguson, 1959:338). If the varieties of a continuum differ considerably between each other, a diglossia internal to a continuum X is latent once large-scale LR enterprises try to promote trans-regional contact between users. As more X communities get involved in one broad LR project, LR actors might want to consider the possibility of creating a diglossia between related varieties (Hornberger & King, 1998:407) where a unified written variety is used for intra-national written communication, alongside the 'authentic' spoken local varieties of a continuum.

The attainment of diglossia as one LR goal, however, has been criticised by Jaffe (2003) for reproducing language hierarchies that create conflicts around language: "one of the unintended consequences of a revitalisation program built on the idea of 'diglossia' was the internal reproduction of dominant language hierarchies that divided rather than unified Corsicans around language" (Jaffe, 2003). The rejection of language hierarchies underlies the LP approach described as *polynomy* (Marcellesi et al., 2003), which has been applied to the LR of Corsican. The core conviction of polynomy is the validation of regional variation without privileging any variety over the others. Despite having anti-hegemonic principles, the *polynomic* model in Corsica does not readily seem to accept contact, either with French or in the form of a blend of varieties of Corsican: "regional variation is acceptable as 'authentic', but contact-influenced variation is not" (Sallabank, 2010:318). Ó Murchadha (2016) made a similar observation in the context of Irish. He argues that the polynomic model will not have room for the post-traditional forms of new speakers of Irish, as long as the practices of traditional Gaeltacht speakers are considered the roots for Irish LP. This attachment to traditional un-mixed varieties seems problematic to LR: LR might largely need the development of new forms of a language, including

perhaps a mixture of features from different varieties, to fulfil new functions as LR progress in the GIDS scale.

Issues around authenticity are visible in Fishman's approach to an internal diglossia, and complicate the functional and geographical expansion of X. Usage in school, for example, is an opportunity, out of the family sphere, to learn and practice the language, as the example of Basque *ikastolas* shows; but it also puts forward the question of how similar this language usage has to be to the language used outside the school, and especially at home. Fishman was worried that the school variety might weaken and eventually replace the vernacular, because he saw this replacement as undercutting any claim of authenticity (Fishman, 2001b:481). In Fishman's view, thus, if the gap between family-community (X_L) and school language (X_H) increases, the authenticity of the revitalised language could become questionable. A fair similarity between a H and L variety may be arrived at and maintained in small communities. However, when the LR are intended to have an impact over more than one community, it becomes difficult to keep a commitment to authenticity, particularly if the school language reflects practices from a different community. The development of an orthography to support literacy is one example of the complications likely to arise if the language of literacy is kept too close to the local variety.

1.2.1.1 A commitment to authenticity at the cost of separating closed varieties?

The development and consistent usage of an orthography is only one aspect to consider to promote literacy and the extension of domains of usage of a LR. However, the question of a shared orthography is perhaps the best example of how a commitment to authenticity complicates the transition from a family-community usage to a regional usage. A community might reject an imported model if it does not resemble faithfully the local language. Jany (2010) reports on the development of an orthography to support documentation and literacy development in Chuxnabán, a Mixe community of 900 speakers. Chuxnabán Mixe is considered part of a larger community of users, the Midland Mixe community (8,090 speakers), which with North-Eastern Mixe community (13,000) comprises a large Mixe community. All community members in Chuxnabán speak the language and there is intergenerational transmission, which is a promising beginning to work on other stages in the GIDS scale. In addition, the community agreed on

the benefits of writing their language. The community, however, did not wish to use orthographies of neighbouring or other Mixean varieties, but to create their own unique orthography. This decision might keep the language of literacy X_H close to the local language X_L in Chuxnabán, and thus avoid a diglossic situation in relation to other Mixean varieties. Mixean varieties, however, differ primarily in their vowel system, and the desire of each community to keep its distinctive features in the written form has impeded the wide use of an orthography (Jany, 2010:235)(Jany, 2010:235).

A distinctive orthography, in addition, can keep a commitment to authenticity in Chuxnabán Mixe by highlighting their distinctive features, and by not forcing a 'groupness' with the Mixe continuum through an imposed general orthography. This commitment, nevertheless, can also sharpen an awareness between users about Chuxnabán Mixe being a separate entity from the Mixe continuum, requiring therefore a separate name, orthography, and perhaps LR programme. A commitment to authenticity thus also means that the Chuxnabán Mixe LR programme would likely have difficulties, or might even avoid, connecting with other Mixe LR works towards the strong side of GIDS, which gradually comes to involve a literate, supra-local usage among an extensive community of users.

The commitment to authenticity in Chuxnabán can also have as consequence the reification of the local language as a separate language from the neighbouring varieties. Authenticity implies a commitment to preserve the differences that make a local language unique. Besides, as it was mentioned in section 1.1.1, and the Norwegian example shows, the establishment of a distinctive written code is a decisive factor delineating a variety from a continuum as a separated language. Thus, besides the potential consequence of hindering supra-local literate contact, the establishment of a local orthography could in time promote Chuxnabán Mixe as a separate language from the Mixe continuum, with its own LR programme. Thus, a commitment to authenticity, reflected in writing, could construct a language with 900 users, which might stand a worse chance of progressing on the GIDS scale than a community of 21,000 users of related varieties.

1.2.2 Spoken authenticity and the idea of design

When the desire to keep the differences between the spoken local varieties of a continuum meets the need for literacy and trans-local communication, an important authenticity-related issue for LR becomes apparent: the assumption that a written model needs to be designed as a reflection of the spoken language.

If authenticity is believed to reside in the spoken language, then writing is seen as a secondary reflex which can and should be *designed* based on the spoken language. This is a complicated notion, because the authenticity-based design of today can always be contested as artificial in the future. That is the case of the standard for Catalan that the philologist Pompeu Fabra developed at the beginning of the twentieth century. This proposal had the explicit purpose of keeping the written standard close to the spoken language. By the end of the century, however, the standard of Fabra was dismissed by some printed and electronic media as a collection of "stilted formalisms" imprisoning the language. These media followed a new trend promoting contemporary spoken Catalan as the only valid reference for measuring the authenticity of the written and formal registers (DiGiacomo, 1999:111-3).

1.2.2.1 Orthographies and the problems of design

The assumption that written language needs to mirror the spoken language is perhaps best perceived in the discussions about orthography [4]. The selection of an orthography is only one of many aspects to consider for the promotion

[4] An *orthography* is the specific set of rules for applying a script, for example, the Latin script. There is an orthography for English and one for Spanish, and that is why, for example, the same phoneme /f/ can be represented with the grapheme <f>in Spanish and with the combinations <ph>, <f>, <gh>in English, or the names of the months are capitalised in English but not in Spanish. Orthography is a term related to the superordinate terms *writing system* and *script*. A *writing system* is one depicting linguistic units of different structural levels, namely, words, syllables and phonemes. From a purely analytic point of view, it can be said that the difference between writing systems is the linguistic unit to be represented. Coulmas warns us that this analytic point of view necessarily disregards the fact that the represented units are not necessarily given, but can indeed be created by this same representation; thus, for example, phonological theory could be an offshoot of alphabetic writing rather than the other way round (Coulmas, 1989:40). Having this caveat in mind, the represented linguistic unit is a criterion to distinguish, for example, alphabetic writing which aims to represent phonemes, from a syllabary which represents syllables, i.e. combination of phonemes. A *script* is a graphic instantiation of a writing system. Thus, despite both being alphabetic writing systems, one of many differences between the Latin and Greek scripts is the graphic instantiation of the phoneme /a/, respectively represented by the graphemes <a>and <α>

of literacy for LR; however, discussions around orthography are common in LR (e.g. Hornberger, 1995; Jany, 2010; Jones, 1998; Lehmann, 2018; Schwartz, 2017; Sebba, 2000). In these discussions, pedagogical reasons are sometimes raised to call for an orthography that reflects the pronunciation, just as in his time Knudsen considered that an orthophonic spelling of a Norwegian standard would make it much easier for the general public to learn to read and write (Jahr, 2014:38). However, making an orthography a reflection of the spoken language also shows traces of assumptions around authenticity. Reflecting upon the case of Manx, Cornish and Maori, Tsunoda (2005:197-9), for example, suggested that accuracy, i.e. closeness to the pronunciation, and distinctiveness from other languages' orthography might be desirable features of an orthography.

It is notable that two salient themes in the discussions about orthography in LR are precisely whether an orthography is adequate to represent the sounds of the revitalised language, and also whether the orthography is distinct enough from major languages in competition with it, such as English or Spanish. Both themes echo authenticity-related concerns of Fishman (section 1.2.1), namely, that the distinctiveness of each language should be maintained and emphasised; and that authenticity is maintained if the language of the school stays close to the language of a community. Moreover, isolationism and mundaneness (section 1.1.5.1) may present the spoken local varieties as model of authenticity, and thus as basis for the design of writing.

In the light of these assumptions, revitalisationists focusing on different varieties of a continuum may easily accept that the purpose of writing is to directly extend the functions of each individual variety into its own written domain. The drawback of this approach, as it has been argued (section 1.2.1.1), is that the spaces of literate usage for a local variety will remain reduced in comparison to the spaces of usage that could group more communities of speakers of related varieties.

The contemporary design of orthographies to support language documentation and LR, show an underlying assumption that the spoken language, and specifically the sounds of the language, have to guide the design of an orthography. As shown by Jany (2010), a design can respond to the desire of a community. However, Hinton (2014) has noted how one common practice among linguists assisting the development of writing in Native American communities is to avoid "the illogical flaws of English" by designing phonemic orthographies

based on what are believed to be the intuitions of native speakers about the sound system of their language. The linguistic team supporting Mono revitalisation, for example, deemed the folk writing conventions, which followed some English practices, inadequate because these conventions could not consistently represent the distinctive sounds of Western Mono (Kroskrity, 2009:77). Munro (2014), for example, shows a concern to keep writing close to the spoken language and to guide pronunciation, as her two basic rules for a good orthography imply: first, every symbol must represent always the same phoneme; second, every phoneme should always be represented the same way. Karan (2014), reflects on how standardisation cannot be hurried, nor orthography be designed in a scientific laboratory; however, he seems also concerned with examples where writing diverges from the spoken language, as he shows when he laments cases of orthographies where "phonological analysis was done in a hurry" (Karan, 2014:122).

Designing an orthography mirroring the spoken language largely misses the fact that the spoken and written language do not necessarily, clearly, nor permanently map each other. Coulmas (1989:47) has highlighted how widely spread is the assumption that a good writing system is an isomorphic mapping of speech. It is easily taken for granted that the optimal writing system concatenates symbols representing sounds in such a way that the relation between written sign and meaning is always mediated by sound Coulmas (1989:47-8). This might be the case for *shallow orthographies*, where the phoneme-grapheme ratio is closer to the ideal 1:1, i.e. the ideal one-sound-one-letter principle. In a shallow orthography such as Spanish there is a fairly regular grapheme-phoneme correlation, and the spelling of a word guides relatively well its pronunciation. There are, in contrast, *deep orthographies*, like English orthography, where for each phoneme there might be two or more graphemes, and so the relation between the spelling of a word and its pronunciation is not always transparent [5]. Spelling conventions, to a certain degree, reflect phonological structure, but "it is usually the phonological structure of words and morphemes rather than connected speech [...] Alphabetic orthographies, no matter how simple the phoneme-grapheme correspondence are not systems of transcription" (Coulmas, 2003:101). However, a simple phoneme-grapheme correspondence is usually considered ideal,

[5]The "depth" of English orthography can be seen in other ways, e.g. the arbitrary multiple realizations of single phonemes (feet/feat); and the morphophonemic depth shown in spellings that remain consistent along related words despite the phonetic differences in regional pronunciations (e.g. write/writing, ride/riding cf. phonetic [t] vs. [d] in American English)

and shallow orthographies are considered simpler and thus superior to deep orthographies (Coulmas, 2003:102).

Sampson (2015:249-64), taking English as example, distrusts the assumption that a purely phonographic orthography is ideal. A *logographic* system is one based on meaningful units, like words or morphemes; *phonographic* systems are those based on phonological units. Sampson considers English spelling a compromise between phonographic and logographic principles, which sacrifices sound transparency in order to conserve in writing important information to grasp meaning. It is reasonable therefore to consider an orthography not only (or necessarily) a guide for pronunciation, but also a window to other types of information which help its users to grasp meaning.

Rastle (2018), for example, highlights the relevance for English reading acquisition of the morphological information present in spelling. Morphological regularities are often far more salient in written than in spoken English, a principle illustrated by the past tense. Depending on the surrounding context, the past tense is usually denoted by phonemes /əd/, /d/ or /t/ (e.g. in 'busted', 'snored' and 'kicked'), and nevertheless these phonemes are always spelled '-ed'. The significance of this consistency is that it creates a reliable orthographic cue to meaning: word final spelling '-ed' denotes the past. Rastle argues that if English writing was a simple one-to-one transliteration of spoken language – a perfect system for learning to decode the printed word back into spoken language – the information about the past tense would be lost in spellings like 'busted', 'snord', and 'kict'. On another example, while the sequence /əs/ has many potential spellings (e.g., service, nervous, princess, haggis), the spelling '-ous' is reserved to denote adjective status.

In support to the importance of morphology for orthography, Sénéchal & Kearnan (2007) found that morphological information in complex words can facilitate reading and spelling, and that knowledge about the morphemic structure of a language can assist a child in reading, spelling, and deriving the meaning of multimorphemic words. On view of this, they argue that systematic instruction of morphology should be given during the elementary years of schooling. They believe that morphology is not currently taught to elementary school children, because educators are far more familiar with concepts of phonemes than with concepts of morphemes and morphemic awareness.

Designing orthographies with the overall aim of guiding pronunciation is problematic for promoting literacy within a continuum in LR. Writing cannot be an accurate reflex of the spoken language unless its conventions are restricted to small geographical contexts, or regularly updated. Moreover, supposing that the purpose of letters is to represent sounds creates an awareness on the necessity to choose a variety embodying the canonical form of the language in question (Coulmas, 2003:97). A choice between varieties is arguably one that few would dare make due to the conviction to respect linguistic diversity and differences that underlies the LR work. One alternative, of course, is the development of many local orthographies like the one used for Chuxnabáan Mixe, with the potential isolation of a small community of literate users. But even if a choice was made among the varieties of a dialect or sociolect continua, there is no guarantee that writing will forever mirror this spoken variety. Spoken forms change whilst written forms are preserved, and thus a gap between spelling and pronunciation is likely to widen in time in alphabetic orthographies (Coulmas, 2003:97).

Finally, linguist-designed orthographies can be overridden by system-external factors. Unlike a linguist, who needs instruments of precision, a script user is not necessarily interested in an isomorphic mapping of speech (Coulmas, 1989:47). Hinton (2014), for example, points out how linguists focus on descriptive adequacy and documentation of "best speakers" as important goals, whilst community members increasingly see writing systems as a language teaching tool. She highlights that many people involved in revitalisation of Native American languages have been educated in English, and that new writing systems are primarily teaching tools to other English speakers trying to learn their ancestral tongue. These circumstances, she believes, result in a bias toward English-based usage, e.g. the use of various symbols for one sound and vice versa, which goes against linguists' idea of a good orthography. In a similar way, contemporary orthographies for Nahuatl try to avoid, for example, the usage of digraphs introduced by Spanish friars in the 16th century. This trend is lamented by bilingual Nahuatl users, educated in Spanish, who want to connect the contemporary literacy with the older practice that, for good or for bad, resemble Spanish in some ways (Meza Patiño, 2012:e.g.). Hinton also notices the importance of showcasing an orthography to gain acceptance. Through web visibility, for example, an orthography developed and used by a fairly prolific writer can become familiar among language enthusiasts, whereas a linguist-developed system might be relegated to specialists' bookshelves (Hinton, 2014:164).

1.3 Writing and convergence: encourage writing in a context of variation

1.3.1 The importance of convergence: reconsidering authenticity in LR and LP

Grenoble & Whaley (2006) highlight that LR almost always requires changing community attitudes about a language. Arguably, this change of attitude goes beyond a revalorisation of local languages above languages of wider communication; this change of attitudes needs also to reflect the dynamic non-exclusive relations that might exist or be created between different varieties of a linguistic cluster.

One important change of attitude thus largely relates to the assumptions about authenticity. Instead of only pursuing authenticity and emphasising the differences distinguishing small language communities, LR works should look also to the possible commonalities shared across broader communities.

To move towards the strong side of LR, attention should be paid to the points of overlap between the varieties of a cluster, and not only to the differences between them. In the context of conflicts between unified and 'authentic' Quechua varieties, Hornberger & King (1998) highlighted the dynamic, negotiated character of authenticity: authenticity must not be portrayed as either unilateral nor unchanging. Most importantly, in situations where divisions are not only linguistic, but also politico-economic and socio-cultural, they argue, LP and LR must begin by *transforming* rather than *restoring* language, and by challenging these divisions. As it has been argued above, concerns about authenticity can be used to draw divisions between language varieties, as well as between groups of people, e.g. the common man and the intellectual, or the rural inhabitant and the urbanite. One way of conciliating authenticity and unification, they believe, is to ratify "the norms and forms which emerge out of the confluence of multiple streams of use of the language" (Hornberger & King, 1998:407). Focusing on the convergence of varieties thus might be a strategy to attenuate the potential isolationism encouraged by an overall concern with authenticity. Promoting convergence and not only respecting and emphasising divergence between varieties of a continuum should be also a focus around which attitudes and strategies in LR should change.

Departing from Hinskens et al. (2005:1-2), by convergence this work refers initially to both the process of increasing similarity between dialects and the result of this process, i.e the partial similarities increasing at the expense of differences. The study of convergence, as defined in this way, has been of interest to dialectology and variationist sociolinguistics, and has typically entailed a diachronic *comparison* of contemporary data from a single dialect with data from and older stage Hinskens et al. (2005:16). In the present work, however, convergence means mainly the *overlap* between synchronic and diachronic varieties of a linguistic cluster. This is the type of convergence LR efforts should look at as strategy to bridge the varieties of a linguistic cluster.

Seen with extreme suspicion, a call for 'unity' based on convergence can indeed serve questionable agendas; furthermore, unity is an ideological construct as debatable as the separation of varieties within a continuum. However, promoting convergence and emphasising overlap rather than only the differences between varieties is necessary to move LR endeavours beyond the family-community level. To move towards the strong side of LR, an endangered continuum might stand better chances if their users get in contact in large spheres of usage than if they limit this usage to small local communities. Beyond dichotomies, like correct-incorrect, high-low, old-new, intellectual-lay, LR should see different varieties of a linguistic cluster as *complementing* rather than excluding each other, and most importantly, as possibly *overlapping in many points* rather than just adjoining each other.

1.3.2 Revaluating writing in LR

1.3.2.1 Writing as medium of contact

Another attitude which needs to be questioned in LR is the promotion of one form of language over others. In particular, LR focuses on the spoken language as source of authenticity, and needs to be gradually complemented by a revaluation of writing as an authentication practice that allows for contact between communities of users.

Writing in LR is important not only to provide a durable, consultable record as, as it is for example in language documentation; writing gradually acquires importance as LR progress up in the GIDS scale. In the strong side of LR, writing

is important to put in contact users from different varieties of a continuum. The contact between different varieties has happened, for example, in the regrettable case of the reservations of Native Americans, where speakers of different dialects came in contact (Hinton, 2001). It seems, also, that the influence of a model taught in schools can prompt the convergence of younger generations' speech, as has happened in the Basque country (Hualde & Zuazo, 2007:155). But how can varieties used by scattered communities get in touch without the tragedy of a forced relocation, or convergence occur in the absence of a model enforced by a government in a political entity like the Basque country? Written contact, and not necessarily supported by uniform practices, seems an opportunity for both this contact and convergence to happen.

Fishman emphasised the importance of 'real communities' over 'virtual communities', arguing that emails or chat boxes cannot replace the face-to-face interaction which allows for the intergenerational transmission of the mother tongue (Fishman, 2001b:458). Virtual communities, however, are emerging also to support contact between users of diverse varieties, and to encourage the practice of writing. The Romani 'virtual' community of social media users (Matras, 2015), is the best example of contact between users dispersed across regions and nations that have no territorial centre, nor a 'home' region or unified government to enforce a language policy. In the virtual community of Romani users, communicating is more important than a uniform usage of spelling, or 'traditional' grammatical forms, as the actual usage often reflects the influence of other languages with which the members are in contact. Such interactions constitute a bottom-up process in which individuals negotiate language practices by the practice of writing itself. The increasing practice of texting in Africa, in another example, is actually the only situation in which many Africans write in their local languages. Texting has thus emerged as a new literate domain for African languages. Concerns about authenticity, which might limit the acceptable pool of linguistic resources to choose from, do not seem to apply in this context, as writers draw on local as well as global linguistic resources in crafting their messages (Deumert & Lexander, 2013). The virtual communities created via the written language are thus an important complement to the efforts conducted in each 'real' local community, as they may allow for the interaction of writers from different varieties without an overall concern for uniformity or authenticity.

1.3.2.2 Writing as authentication practice

Writing also acquires relevance for LR as it gradually turns from an instrument to record a language into an instrument to allow the communities of users to (re)create language. Writing needs to be used not only to record the 'authentic' language of a community 'just in case' LR fails. LR enterprises might also benefit from encouraging a collective re-creation of language which might bring together all the actors involved. A contrast between *authenticity* and *authentication* was proposed by Bucholtz for sociolinguistics. Bucholtz posits that although it is difficult for sociolinguistics to abandon the concept of *authentic speaker*, sociolinguistic analysis must acknowledge that authenticity does not exist prior to the authentication practices that creates it: "authenticity is always achieved rather than given in social life" (Bucholtz, 2003:408). Likewise, a focus on an authentication process, rather than on a concern for locating/preserving authentic varieties might be beneficial for LR. This view largely conforms with the transforming rather than restoring potential of LR envisaged by Hornberger & King (1998).

Writing is important in LR not only as a tool whose value depends on its closeness to the spoken language, but overall as an authentication practice which can eventually produce converging written usage. Weinberg & De Korne (2016) applied to LR the observations of Bucholtz about authenticity and authentication, and the idea of *communities of practice*, a concept adopted from social psychology by sociolinguistics and language acquisition. A community of practice is defined by three characteristics: the mutual engagement of participants, a joint enterprise and a shared repertoire. Significantly, through the mutual engagement in the joint enterprise, participants develop a shared repertoire which might include not only a common language, but also styles and routines that express members' belonging to the community. If one considers the inclusion of literacy in LR as a joint enterprise, one important goal should be encouraging mutual engagement to take place in writing, even in the absence of widely accepted conventions. A community of practice of writers is of the utmost importance for LR because it might develop and consolidate a shared repertoire (converging written practices), not through explicit debates, or accurate designs but largely through the performance of its members.

One significant point to bear in mind is that current established written conventions in general are not only the product of conscious linguistic design, but

that they largely emerged from the practice of a community of writers. Coulmas (1989) notes that written conventions for a significant number of languages were not established as brand new designs but, for example, by a gradual adaptation of the script used by neighbouring written languages. Solutions to adaptation problems in the earliest written languages tended to develop naturally, not by conscious design. Systematic coherence was never in itself important, because it is unlikely that the systematic make-up of early scripts was clearly understood by users, in the way that few people understand the systematic make-up of alphabetic orthographies nowadays (Coulmas, 1989:42-4). Only recently have professional linguists become involved in hurried endeavours to design written norms, largely as part of revitalisation efforts for languages which are rapidly losing speakers. Karan (2014) contrasts these rushed enterprises with the standardisation processes of European languages to show that standardisation cannot be approached as a race against time. He notices how European written conventions developed over time, based on individual, local, and then regional decisions; even the prescriptive influence of the French Academy, for example, remained limited for a long time, and their first two editions of dictionaries reflected the writing practices of the day. Thus, none of the European alphabetic orthographies seems to be the fruit of deliberate linguistic calculation, since linguists, governments or pedagogues were not necessarily the most influential actors in their development (Karan, 2014:120-2).

The case of written Dutch in Flanders during the nineteenth century (Vandenbussche, 2002) shows how the debates between spelling designers, for example, can be sterile theoretical discussions that take place over the heads of actual writers. To the actual writers of Dutch, divergent spellings were not an issue, and they wrote using a variable spelling system up to 1900. Notably, it was the impact of supra-regional communication (i.e. the expansion of a community of practice of writers) that created a certain awareness of the need for more uniform usage among the literate users of both high and low social class. The English deep orthography, so notorious for its deviation from the spoken language, is also noteworthy as a bottom-up development that "evolved ***through usage*** and decision over time by scribes and printers" (Karan, 2014:121).

The English case suggests that relative written uniformity may be achieved gradually, largely served by the practice of writing itself and building upon previous examples, and not only by an impeccable design or the direct intervention of a language academy. Before printing, English spelling echoed the regional spo-

ken dialects, some of them hardly mutually comprehensible (Svartvik & Leech, 2006:43-5). In the early 15th century, the practices in the Chancery Office were a first significant source of examples for professional scribes across the country (Upward & Davidson, 2011:81). Later, the introduction of printing in 1476 increased the availability of previous written examples to imitate, and as time went on writing adhered more and more to previous written models and reflected the spoken language less and less. Notably, the consolidation of printing coincides with a significant change in the pronunciation of English, i.e. the Great Vowel Shift in the 15th and 16th centuries, and in many cases the gradually fixed spelling no longer represented the pronunciation of the day (Upward & Davidson, 2011:174-7). Printers might have initially used their dialect in their production of books, but "printers knew it was in their own interest to adopt forms of the language that would have wide acceptance, and once that printing was established, there was naturally a tendency to use forms of language that had already appeared in earlier printed books" (Upward & Davidson, 2011:84). The history of English spelling also shows that there was never a centralised plan or body like an academy to guide the development of spelling. Early uniformity was more or less achieved through imitation of available examples, and modern spelling emerged from a "slow process of increasing consensus among printers, lexicographers, reinforced by teachers, authors of literacy primers and published writers" (Upward & Davidson, 2011:4-6). In sum, English spelling was shaped by the people involved in the community of practice of writing.

1.4 Conclusions: looking for convergence in the written practice

The emphasis on differences based on ideas of authenticity makes it a challenge to support a regional literate usage along a continuum X. Moreover, the emphasis on differences might form a source of contention and separation between LR actors and encourage the building of small standard written languages with few users. It is therefore necessary to look also at the overlap between the varieties that may bring together rather than separate users of a continuum.

An investigation of convergence in LR should pay especial attention to the written language produced by users of a continuum. The encouragement of literacy for LR should also depart from an idea of authentication rather than

authenticity. Besides, and maybe instead of, 'designing' written usage based on an investigation of authentic varieties, LR efforts could benefit from investigating the authentication practice in which writing is used by the writers of a continuum X. The development of English spelling suggests the significant role a community of practice plays in the gradual achievement of written conventions across a continuum of spoken varieties. The points of convergence in the practice of writers could constitute imitable/acceptable examples for literate users across the continuum X, and be the departure for eventual, widespread conventions.

Investigating the spoken language to guide the development of a written usage (for example in the design of an orthography) might be necessary when no previous or current written practices exist. Nevertheless, when users from different varieties are actually writing, it might be a good idea to investigate this practice, highlighting the points of convergence occurring in it. The promotion of literacy and the progress towards the strong side of LR, is precisely one of the enterprises for which the study of written rather than spoken usage of the language is necessary.

The case of Koldo Michelena is notable for having drawn upon contemporary written examples to design a unified standard to support Basque LR, the *Euskera Batua*, albeit following the approach recognised by Haugen (1966) as stereotypical, namely, selecting a variety and within this a set of variants to be codified. In addition, a central Academy sanctioned and enforced the choices. In the absence of a prestige dialect, the variety selected to codify was the literary Basque created by a few poets and writers, and within this the forms with the widest usage were selected. Notably, the Academy avoided the inclusion of forms not previously attested in the tradition of the language (Hualde & Zuazo, 2007:149-51).

For many linguistic continua the enforcement and promotion of a standard would be difficult to encourage without a central government or Academy. Moreover, for many LR actors the idea of a cross-varietal standard may seem contrary to the spirit of respect for diversity. Nevertheless, the exploration of writing to guide the development of writing is a very relevant strategy that can be recovered from the Basque case. Moreover, the location of points of convergence in an extensive exploration of written samples across a continuum of varieties seems necessary to balance a potential tendency to fragmentation in LR enterprises. To the best of my knowledge, however, such strategy has not been attempted

in other LR enterprises. Of course, one cause for this might be that literacy is very recent in many continua in need of revitalisation, and thus there are not yet enough written examples to draw upon. Another possible cause is the current heterogeneity of the written texts in terms of orthography, grammar and vocabulary, which adds complexity to the study of written samples from across a continuum. However, the assumptions about authenticity which are common among LR actors might largely underlie a disregard for written data in LR as an artificial, or untraditional form of language.

Summing up, to progress towards the strong side of LR, if writers from different varieties of a continuum have created examples to investigate, it is worth exploring convergence in these examples, and emphasising this overlap over the heterogeneity, or ostensible artificiality of these written attempts.

Now this work will move to describe the particular case of Nahuatl, a linguistic continuum spoken in Mexico, and the LR efforts focusing on it. It will be shown that an official strategy of designing written conventions accommodating to the spoken diversity is difficult due to the complexity involved in delimiting varieties. It will also be argued that this strategy might deconstruct the Nahuatl continuum into small communities of users, thus hindering the trans-local literate usage that is necessary to take Nahuatl LR to the strong side of the GIDS scale. It will be argued that Nahuatl users write despite the lack of widely recognised conventions, and that the old corpus known as Classical Nahuatl and the morphosyntactic structures called nuclear clauses could support the exploration of convergence in the heterogeneous contemporary written practice.

Chapter 2

The Nahuatl cluster

2.1 The diversity of Nahuatl

2.1.1 The contemporary continuum

The Nahuatl linguistic cluster plays a central role in the study of the Mexican history, since the Mexicah, founders of the Aztec empire, spoke a variety of Nahuatl. In addition to its historical importance, according to figures of the National Institute of Statistics and Informatics (INEGI), Nahuatl is the most widely spoken indigenous language in Mexico with 1,544,968 speakers (INEGI, 2010).

The spoken varieties of contemporary Nahuatl constitute a dialect continuum, although there is significant geographical dispersion between many Nahuatl communities. By the year 2000 (INEGI, 2005), 95.6% of Nahuatl speakers were concentrated in an area across Central Mexico. This area narrows towards the southeast along nine states which, in order of concentration of speakers, are: Puebla (28.7%), Veracruz (23.2%), Hidalgo (15.3%), San Luis Potosí (9.6%), Guerrero (9.4%), Estado de México (3.9%), Distrito Federal (2.6%), Tlaxcala (1.6%) and Morelos (1.3%). However, Nahuatl communities are often separated by communities speaking Spanish or other indigenous languages, and thus geographical dispersion hinders the direct contact between members of many Nahuatl communities. Scarce contact between communities might have had a significant impact on the abandonment of the language, reducing opportunities to use the language outside small circles. Flores Farfán (2002:229) posited the

lack of contact between speakers of different isolated dialects as one main reason for rapid language shift, as the geographical dispersion often implies that the actual linguistic community for any Nahuatl variety is quite small. One has to take into account, however, other factors e.g. a scarce presence in the Mexican educative system and mass media, which reduces the economic appeal of Nahuatl and the chances for developing broad spheres of trans-local, not-necessarily-face-to-face communication.

Despite the geographical dispersion, it has proven difficult to trace clearly the boundaries between varieties. Lastra de Suárez (1986) and Canger (1980, 1988, 2011) studied contemporary variants aiming to define dialectal areas. Both studies are notable because of their comprehensive attempt to study varieties in all the continuum. Significantly, both studies highlighted the difficulty of defining clearly the limits of dialectal areas. Lastra de Suárez coordinated a dialect survey in 92 localities all across the areas where Nahuatl is spoken. A questionnaire with 473 Spanish words was used to collect Nahuatl equivalents, and interviews were recorded on audio tapes. Variations in the data were grouped according to phonological, lexical and grammatical criteria, trying to find shared features to sketch contemporary dialectal areas. Canger was interested in the contemporary dialectal panorama as a basis to achieve a greater understanding of the dialectal situation in pre-Hispanic times, and to sketch the historical evolution of the dialect areas. Canger used data from contemporary varieties and colonial sources, thus aiming to understand the Nahuatl cluster as a whole. She advocated an interpretative dialectology in which "a comparison of forms from different dialects results in an interpretation of these forms in all the dialects, collectively and individually", and wanted to show that "any one Nahuatl dialect is best understood and described in the perspective of other Nahuatl dialects" (Canger, 1980:18-9).

The studies of Canger and Lastra de Suárez provided a tentative classification of Nahuatl varieties. Lastra de Suaréz proposed to distinguish four areas: Central, La Huasteca, and a periphery subdivided in a Western Periphery and an Eastern Periphery. Canger classified the dialects in Central and Peripheral dialects, with a subdivision in Western and Eastern within the Peripheral group. The major groups and subgroups are listed in **Table 2.1**. The dialectal areas sketched by Canger are shown in **Figure 2.1**. Four of the isoglosses proposed by Canger are shown in **Figure 2.2** and include: the use of either /t͡ɬ/, /t/ or /l/, e.g. in *tlacatl, tagat, lacal*, 'man'; the use of the augment /o/ with past

forms; and the formation of the preterite form by either dropping or keeping the final vowel of the present form in one class of verbs, e.g. /kiːs/, /kiːsak/, ‘he went out’, present form /kːisa/, ‘he goes out’.

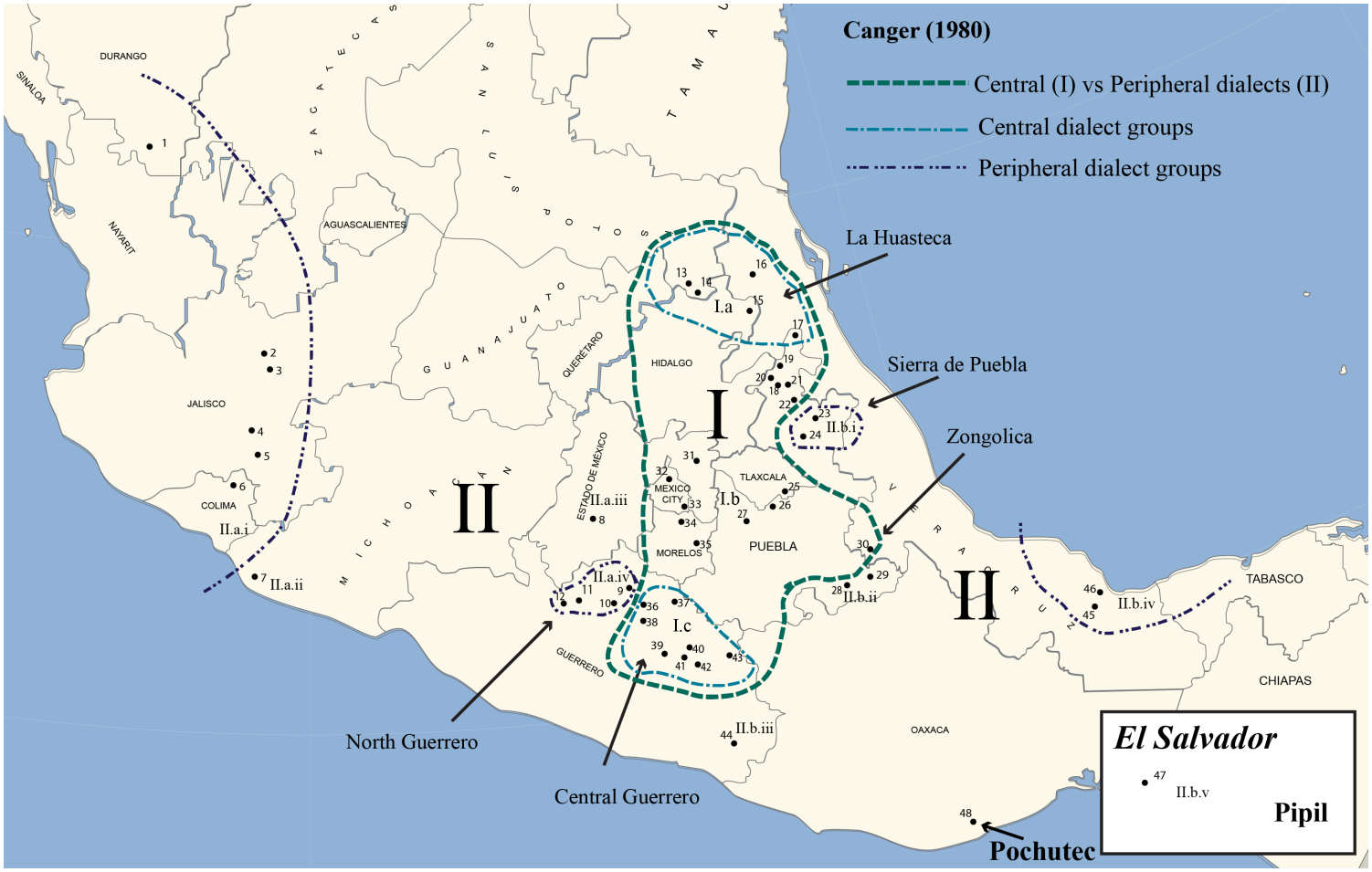


Figure 2.1 A sketch of major dialectal areas based on Canger (1980). The numbered dots represent the villages/varieties for which Canger obtained data, either by collecting them herself, or from other documental sources. Number 32, for example, represents Classical Nahuatl as described in colonial sources (Molina (1571), Carochi (1645), etc.). Number 48, Pochutec is extinct and is considered a branch different to Nahuatl varieties. Pipil (group of varieties of El Salvador) have also later been considered a branch distinct from the one comprising all other varieties in the map (see for example Campbell, 1985)

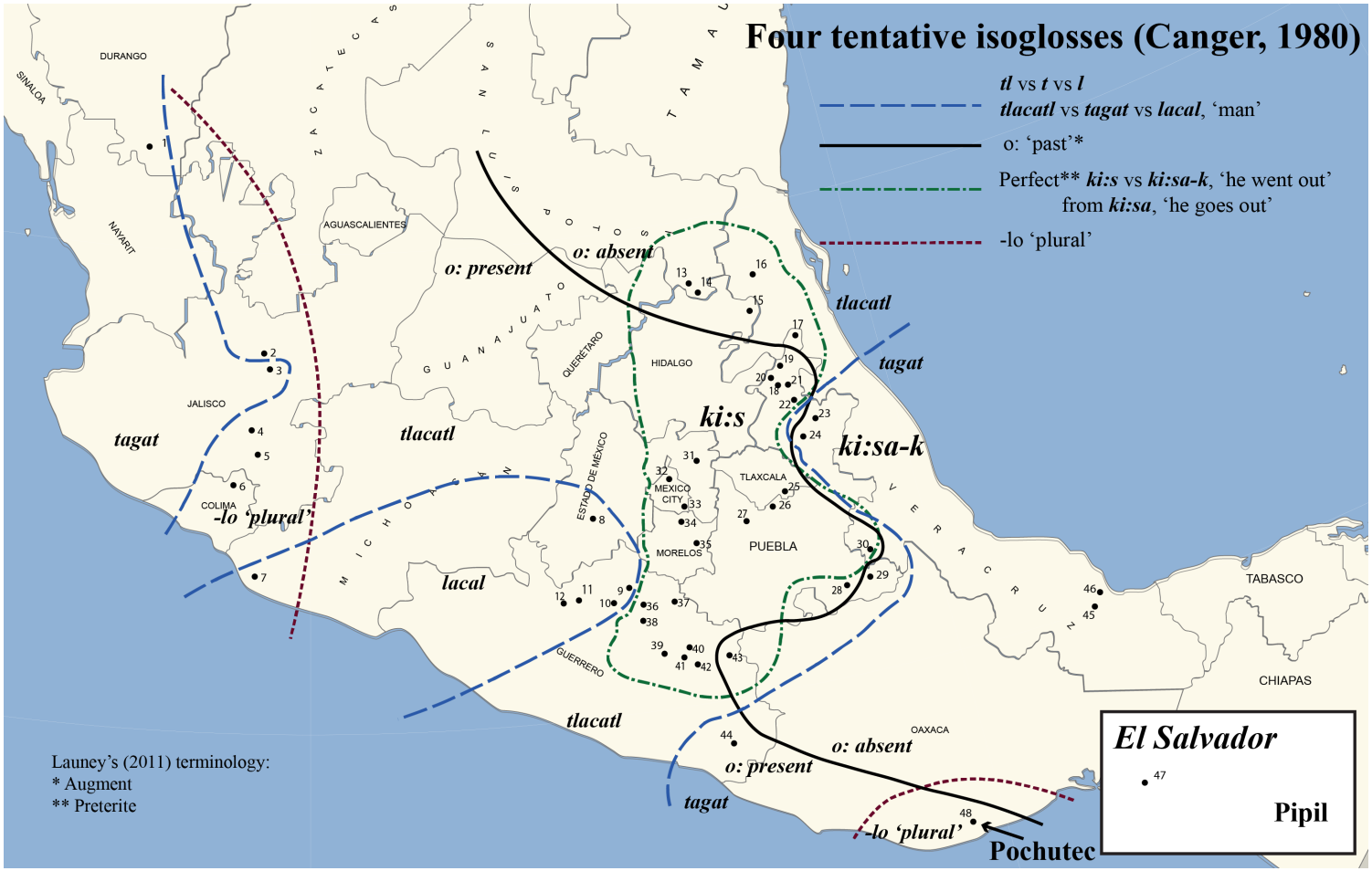


Figure 2.2 Four of the isoglosses proposed by Canger (1980).

I.	CENTRAL		
	a.	La Huasteca	
	b.	North Puebla	
		Valley of Mexico	
		Morelos	
		Tlaxcala	
		Central Puebla	
	c.	Central Guerreo	
II.	PERIPHERAL		
	a.	Western	
			i. Jalisco, Colima, Durango
			ii. Michoacan
			iii. Almomoloa
			iv. North Guerrero
	b.	Eastern	
			i. Sierra de Puebla
			ii. East Puebla
			iii. South Guerrero
			iv. Isthmus
			v. Pipil (El Salvador)

Table 2.1: Nahuatl dialects as proposed by Canger (1980). A division between central and peripheral dialects was proposed by Canger (1980:16), and largely agrees with the classification further sketched by Lastra de Suárez (1986). Pipil (II.b.v) was later considered by Campbell (1985:3) to be a branch different from the rest of the varieties in the above list, which he treats as 'Core Nahua'.

Canger and Lastra de Suaréz highlighted how difficult it was to define dialectal areas, and to show the complex relationship between varieties. The classifications proposed by each largely coincide. They differ mainly in the treatment of La Huasteca and the South East of Puebla (Lastra de Suárez, 1986:189). Canger had lamented the lack of more data to approach the question (Canger, 1980:16), but after the notable effort to obtain more data, Lastra found no significant improvement to aid the definition of dialectal areas: "*Después de examinar los datos recogidos durante siete años hay que reconocer que la respuesta no es fácil*", 'after examining the data collected over seven years, one has to admit that the answer is not an easy one' (Lastra de Suárez, 1986:189). Canger was not satisfied with the sketch proposed in **Table 2.1**. She emphasised that branching charts cannot show the complex relationships that exists between dialects[1], e.g.

[1] It seems more the case that language history consists of a multidimensional rearrangement of variation, for example social or geographical, rather than a branching from one uni-

that La Huasteca (I.a.) shares a number of features with Sierra de Puebla (II.b.i) and Isthmus (II.b.iv) (Canger, 1980:16). One example of a variety difficult to classify is the variety of central Guerrero, geographically situated towards the southwest of Mexico. This variety not only shows influence from the central varieties, but has also traces from the eastern varieties (Canger, 2011). In a similar way, the Zongolica variety, number 30 in **Figure 2.1**, is another variety difficult to classify due to the mixture of features from eastern and central dialectal areas (Hasler, 1996:164).

More recently, Olko & Sullivan (2013:203) have noted that, leaving aside dialectal classifications for a moment, Nahuatl speaking communities fall roughly in two types: those that experienced early intense contact with Spanish culture, located in the central area of Mexico; and those from peripheral areas that never [or until very recently] underwent intense contact with Europeans. Taking this distinction into account, the varieties spoken by all these communities share similar characteristics of a language that continues to evolve either due the processes of its own inherent structure, or by the influence of Spanish. The contact with Spanish thus adds further complexity to the task of sketching a neat classification of Nahuatl varieties.

The region where the states of Puebla, Oaxaca and Veracruz meet illustrates the sort of complexity involved in classifying Nahuatl varieties. This region has been chosen as example in part because it is relevant to some of the findings regarding the text OaxN **(section 5.4)**. In this area three varieties of Nahuatl are recognised by the National Institute for the Indigenous Languages (INALI) in localities within a radius of approximately 40 km, taking Santa María Teopoxco as centre (**Figure 2.3**). Depending on the source, a '(North) Oaxaca' Nahuatl variety would comprise localities as far as some 60 km northeast from Teopoxco –Apixtepec (Simons & Fennig, 2018)– and 90 km southwest from Teotitlán –Santo Domingo Yolotepec (INALI, 2010).

The variety recognised as *North Oaxaca Nahuatl*, according to the SIL and Ethnologue, comprises a group of localities in the northern part of Oaxaca.

dimensional trunk variety. Penny (2000:23-8) has pointed out that language history is better explained as a change from one state of variation to another state of variation. A branched model of linguistic relationships is inadequate, among other reasons, because it erroneously suggests that varieties springs from a single unitary origin. Penny notices that it is difficult to deal with relationships that are based on gradation, but language is gradated along a number of parameters. One finds it easier to work with models which impose boundaries. However, although a subdividing process can sometimes help, in diachronic and synchronic studies of language it more often distorts.

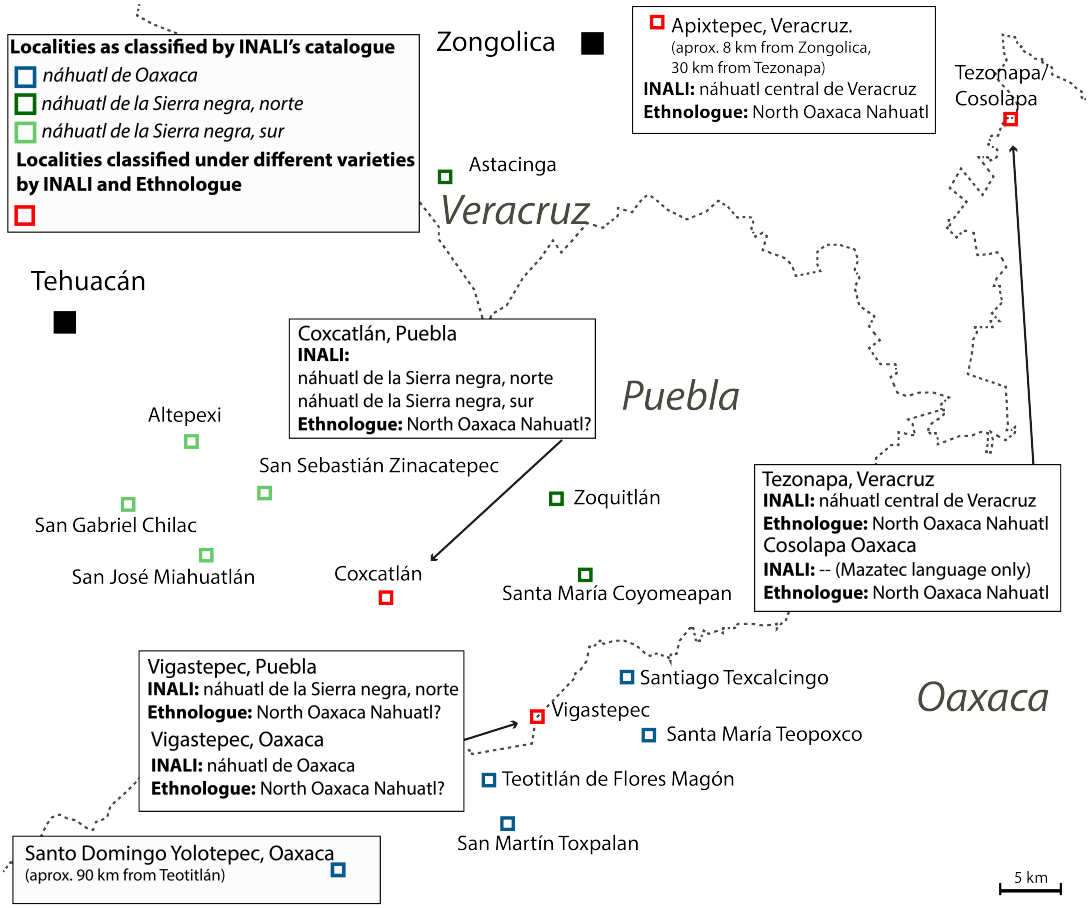


Figure 2.3 The localities in the area roughly related to the text OaxN. Figure based on information from Ethnologue (Simons & Fennig, 2018), the web page dedicated to North Oaxaca Nahuatl at the SIL website for Mexico (Prado Bernardo, 2018), the Catalogue of the National Indigenous Languages of INALI (INALI, 2010), and the website of INEGI.

Mazatec, another indigenous language, is spoken in towns to the south, east and north of Santa María Teopoxco. The localities recognised by SIL in Oaxaca are Apixtepec, Cosolapa, El Manzano de Mazatlán, San Antonio Nanahuatipam, San Gabriel Casa Blanca, San Martín Toxpalan, Santa María Teopoxco, Teotitlán del Camino and Ignacio Zaragoza. Additionally, Ethnologue recognises the localities of Tezonapa in Veracruz, and Coxcatlán in Puebla as pertaining to Northern Oaxaca Nahuatl. The area of these localities is close to the dialectal area labelled by Canger (1980) as II. b. ii. East Puebla, and according to her first sketch would fall into a peripheral dialect (**Figure 2.1** and **Table 2.1**). The closer locality covered by Canger's study is Zoquitlán, marked as number 29 in **Figure 2.1**.

Four localities show difering classifications when comparing the information from Ethnologue with INALI's catalogue: Apixtepec, Tezonapa/Cosolapa, Vigastepec and Coxcatlán. Apixtepec, listed as Ethnologue as located in the state of Oaxaca, is actually in the municipality of Zongolica, Veracruz, and is classified by INALI inside the variety *náhuatl central de Veracruz* (number 6 in **Table 2.2** and **Table 5.4**). Cosolapa and Tezonapa are part of the same town divided between two states: Cosolapa is on the Oaxaca side of the town, whereas Tezonapa is considered part of Veracruz. Cosolapa is listed by INALI only under the variety *Mazateco del noreste*. Tezonapa in turn is classified by INALI under *náhuatl central de Veracruz*.

The discrepancies in the classification of Apixtepec, Tezonapa and Cosolapa might indicate that INALI relies on more accurate or up-to-date information than the SIL. The cases of the localities of Vigastepec and Coxcatlán, however, suggest that the classification criteria of INALI might be highly influenced by political divisions. Vigastepec is politically divided between Oaxaca and Puebla. It has therefore two identification keys in the Catalogue of Municipalities and Localities of INEGI: 205450004 as part of the municipality of Teotitlán de Flores Magón, Oaxaca; 210350014 as part of the municipality of Coxcatlán, Puebla. As part of Oaxaca, Vigastepec is classified by INALI under the variety *náhuatl de Oaxaca*; as part of Puebla, Vigastepec is classified under the variety *náhuatl de la Sierra negra, norte* (number 5 in **Table 2.2** and **Table 5.4**). Coxcatlán is mentioned by Ethnologue as one of the localities in which North Oaxaca Nahuatl is spoken, although it seems to refer to Vigastepec, as part of the municipality of Coxcatlán and not to the actual town of Coxcatlán. INALI, on the other hand, list the town of Coxcatlán under two varieties: *náhuatl de la Sierra negra,*

norte, and *náhuatl de la Sierra negra, sur* (number 10 in **Table 2.2** and **Table 5.4**). The SIL, in contrast, recognises one *Sierra Negra Nahuatl* and does not make a subdivision between north and south[2]. It is reported (Prado Bernardo, 2018) that the people of the zone around Teopoxco relate that the area was repopulated by migrants some 150 years ago, after the majority of inhabitants died in an epidemic.

Canger proposed that the dialect contact caused by migration is a significant historical reason for the complex dialectal situation. After a fundamental early rupture that formed two main groups, one Eastern and one Western, many smaller dialectal groups were formed. However, these groups were not isolated and, through migration and the contact with speakers of different areas, Nahuatl people borrowed forms and shared innovations to a degree that means that, in many cases, the boundaries between dialectal areas are not clear (Canger, 2011:254-6).

As for contemporary migration, there seems to be a gap in the literature regarding the potential effects of dialect contact caused by it. Flores Farfán (2002:229) considered contemporary migration, both internal and to US cities, very relevant for language change, but only as a factor for the abandonment of Nahuatl in favour of Spanish or English. The constant contact with Spanish- and English-speaking populations, he noticed, encourages Nahuatl speakers to develop sociolinguistic competences in those languages. Studies related to language contact focus on the Spanish-Nahuatl contact effects, e.g. Flores Farfán (2003), Francis & Navarrete Gomez (2003), Hill & Hill (2004). The aforementioned isolation of many varieties, also noticed by Flores Farfán, might explain that research on Nahuatl contact focuses first on its relation with Spanish. The census data of 2000[3], however, showed that as result of migration there were Nahuatl speakers in every state of Mexico (INEGI, 2005:4). Furthermore, the census suggests that speakers from different varieties could likely meet in key destinations like Mexico City, in the Distrito Federal. According to the National Institute of Statistics and Geography (INEGI), the figures reported for Nahuatl

[2]Incidentally, Arnulfo Prado Bernardo, the compiler of a dozen short stories labelled as *North Oaxaca Nahuatl*, and author of the webpage for the North Oaxaca variety in the SIL website, is a speaker from San Sebastián Zinacatepec (MacSwan, 1997:96) and, according to the classification of INALI, a speaker of *náhuatl de la Sierra negra, sur*.

[3]The last publication focusing specifically on Nahuatl speakers refers to the general census of 2000. It reports on indicators like migration patterns, literacy rates and territorial distribution. There are updated figures from 2010 on the number of speakers, and rates of monolingualism, but there are no relevant publications covering all the indicators of the report on the census of 2000.

speakers departing from the Distrito Federal to other States can mostly be considered as referring to migrants returning to their home States (INEGI, 2005:25). It follows that speakers from at least 7 states were found in the Distrito Federal between 1995 and 2000.

Contemporary migration, therefore, could not only result in contact of Nahuatl speakers with Spanish; it could also be an opportunity for contact with other Nahuatl varieties. More recently, intercultural universities have allowed for the contact between speakers of different varieties in at least one of their campuses (Pharao Hansen, 2016:358-63). It could be therefore interesting to explore the contact of a number of Nahuatl varieties with a *host* Nahuatl variety, in a similar fashion as Wilson (2011) studied dialect contact in the Czech Republic[4]. Studies like this one in the Nahuatl continuum have not, to the best of my knowledge, been attempted so far, but might in time become possible.

Despite the complexity involved in a dialectal classification of the Nahuatl continuum, the INALI officially considers that there are around thirty Nahuatl varieties, listed in **Table 2.2**. INALI's classification is based on criteria from linguistic genealogy, dialectology, sociolinguistics, and even on the name given by speakers to their variety (INALI, 2010). The number of varieties recognised contrasts with the major dialectal areas posited by Canger (1980, 2011), and thus shows INALI's aim of achieving greater granularity in the analysis of the continuum. It also echoes a tendency to discuss the varieties in the continuum as discrete entities. This trend is frequent in available studies on contemporary Nahuatl, some examples of which are Beller & Beller (1979), Brockway (1979), Sischo (1979), Tuggy (1979), and more recently Hasler Hangert (2001), Tuggy (2001) and AVELI (2010). It is true that many studies have to focus on one small community at a time, being conditioned by the resources and data available. However, the tendency in descriptive studies to present varieties as delimited, named entities contributes to a fragmented perception of the continuum.

A thorough classification of varieties is the departing point of INALI's practical endeavours, and reflects on its work as LP agency and LR promoter. One of INALI's goals, for example, is to create or consolidate written norms to extend the usage of the indigenous languages to literate contexts. As declared by the former general director, Javier López Sánchez (INALI, 2015), INALI intends

[4]Wilson studied the linguistic behaviour of 39 university students (speakers of three Moravian dialects) living in Prague, where a non-standard *koïné*, Common Czech, is normally used in informal communication.

Number	Variety	Speakers
1	Mexicano de la Huasteca Hidalguense	212,300
2	Mexicano de Guerrero	153,773
3	Náhuatl de la Sierra, noreste de Puebla	141,737
4	Náhuatl de la Huasteca potosina	133,343
5	Náhuatl de la Sierra negra, norte	131,015
6	Náhuatl central de Veracruz	130,979
7	Náhuatl de la Huasteca veracruzana	130,364
8	Náhuatl del noroeste central	76,837
9	Náhuatl del centro de Puebla	57,382
10	Náhuatl de la Sierra negra, sur	32,321
11	Náhuatl del Istmo	27,210
12	Náhuatl de la Sierra oeste de Puebla	20,461
13	Mexicano del oriente central	19,252
14	Náhuatl de Oaxaca	8,556
15	Mexicano del centro alto	8,100
16	Náhuatl del Istmo bajo	7,707
17	Mexicano central de occidente	5,999
18	Mexicano de Temixco	4,199
19	Mexicano del oriente de Puebla	3,817
20	Mexicano de Puente de Ixtla	3,069
21	Mexicano de Tetela del Volcán	2,972
22	Mexicano del centro	2,623
23	Mexicano central bajo	1,223
24	Mexicano bajo de occidente	1,108
25	Náhuatl alto del norte de Puebla	1,088
26	Mexicano del centro bajo	841
27	Mexicano alto de occidente	671
28	Mexicano del noroeste	591
29	Mexicano del oriente	286
30	Mexicano de occidente	84
	TOTAL	1,319,547

Table 2.2: The 30 Nahuatl varieties recognised by INALI. The number of speakers were estimated by INALI according to the 2010 census. Table adapted from De la Cruz Cruz (2014).

to create norms **(see 2.3.2)** for at least 68 indigenous languages, although the goal is to create as many written norms as necessary according to the variation within each linguistic group, e.g. the Nahuatl continuum. Taking into account the variability and complex dialectal situation of Nahuatl, to progress on a regional or national scale seems difficult for a strategy pretending to accommodate literate usage to the spoken variations of each community. A thorough classification might help INALI to deal with internal variation by working with smaller groups towards the development of a literacy model acceptable for each com-

munity. It is notable, however, that seventeen of the thirty varieties have less than 9,000 speakers. The most optimistic outcome of this strategy could thus be the splitting up of the continuum into numerous literate communities of few users; and this outcome in turn depends on whether people will even bother to become users of a written language, particularly of one useful to communicate with only people in their town and the one immediately neighbouring it.

Summing up, contemporary Nahuatl is a continuum of geographically scattered varieties. Historical contact might have caused the complex linguistic relation between contemporary varieties that hinders a satisfactory delimitation of dialectal areas. It is notable that the possibility of contact between contemporary varieties has not been explored, despite the fact that recent migration might offer opportunities for contact between speakers from across the continuum. It is also worth noticing that the official stance is to treat the continuum as a linguistic group of thoroughly classified varieties, despite the difficulty of defining linguistic limits between them. The lack of attention to contact between varieties, and the tendency to discuss varieties as delimited entities combine to encourage a perception of fragmentation in the continuum.

2.1.2 Classical Nahuatl

Classical Nahuatl (CN) is neither a very precise nor easily definable concept. Dialectological and sociolinguistic research identified a problematic oversimplification underlying the concept of CN. Further considerations are necessary to attempt an understanding of CN in relation to the contemporary continuum.

CN can be seen first as a subset of the Nahuatl written corpus of colonial times. The term has served as a cover term for the written Nahuatl material from the sixteenth and seventeenth centuries (i.e. the first two centuries of the colonial presence of Spain in Mexico) excluding manuscripts from highly divergent dialects from the Western and Eastern periphery (Canger, 1988:50). CN is thus largely related to written texts from around central Mexico, but the delimitation of the CN corpus is not only geographical. The spectrum of written Nahuatl during this period has in one extreme mundane texts like wills, and court records, and in the other, religious poetry, historical chronicles, and prescriptive grammars and vocabularies. However, Launey (2011:xvii), for example, defines CN as "the literary language of the century following the conquest". CN is

thus mainly associated with the High written varieties of the colonial period, particularly as described and encouraged by Spanish missionaries in prescriptive grammars and vocabularies (Flores Farfán, 2004a, 2010a).

CN can also be seen as the codified written variety described in colonial prescriptive grammars and vocabularies, and largely reflecting the Nahuatl High variety (understanding High in the diglossic sense discussed in section 1.2.1) used in central Mexico in the sixteenth and seventeenth centuries. The term CN is indeed mainly associated with the language of Mexico-Tenochtitlan, the Aztec capital and the centre of power in the sixteenth century, because this was the variety chosen by grammarians and lexicographers for codification[5]. The bilingual dictionary of Alonso de Molina (1571), for example, warns the reader that it records the vocabulary used in Mexico and Tezcoco, thus leaving out the lexical variation existent in the provinces; the grammars of Olmos (1547) and Carochi (1645) make recommendations about the grammatical forms and words that are rustic, and not very elegant (Flores Farfán, 2010a:197-9). A sociolinguistic perspective, Flores Farfán (2004a, 2010a) argues, reveals an internal diglossia in pre-Hispanic times. The existence of a pre-Hispanic diglossia is exemplified by the use at the time of the opposing terms *macehuallahtolli* 'the language of the common man' (L) and *pillahtolli* 'the language of the nobles' (H). In such relation, the H variety can be largely identified as the Mexica High social dialect, Mexico being the centre of political and economic power (Flores Farfán, 2010a:194-5). The survival of the word *macehualli* as opposed to *pilli* in contemporary varieties evidences, Flores Farfán considers, that the L varieties, in general, are the ones which survived the invasion (Flores Farfán, 2004a:173). The Nahuatl continuum before and immediately after the arrival of the conquest was as diverse as one could expect. Flores Farfán warns us, nevertheless, that it is important to bear in mind that grammatical studies of CN, e.g. Andrews (1975), rely largely on prescriptive sources and High corpora, rather than on Low corpora like the wills and petitions of the common people (Flores Farfán, 2010a:199).

Before the awareness caused by sociolinguistic research and LR enthusiasm, the value assigned to CN was not necessarily shared by the contemporary spoken varieties. Outside academic contexts, contemporary varieties are still commonly regarded as lower-prestige deformations of an idealised original variety. Canger

[5]As mentioned in section 1.1.1, codification is the prescription of a set of largely unvarying orthographical, grammatical and lexical rules.

(2011) notices that the abundance of sources for studying CN, plus a logical assumption that the new develops from the old, makes it too easy to assume that the contemporary varieties evolved from a variety spoken in the centre of Mexico in the sixteenth century. She has proposed, in contrast, that the language of the Mexica capital –whose H sociolect was chosen for codification by Spanish missionaries– cannot be identified with only one regional dialect, and most likely was the result of the meeting of speakers from the provinces who arrived there over the centuries previous to the conquest. Canger has introduced the term *urban Nahuatl* to identify this innovative variety of the Aztec capital, which includes innovations, according to her, not found in other varieties of the language. This implies, Canger emphasises, that the urban Nahuatl was not the origin of the modern varieties; urban Nahuatl surged as an innovation developed from the contact of dialects spoken in the provinces which, in turn, were the basis for the modern varieties (Canger, 2011:255). Akin to the historical misconceptions about the origin of contemporary varieties signalled by Canger, CN is commonly seen as a superior variety. Flores Farfán (2010a:190) not only identifies comparisons with CN as one significant source of negative attitudes towards contemporary varieties; he is also wary of CN being used as linguistic capital and thus exposed to the opportunistic manipulation of 'cultural brokers' (Flores Farfán, 2005:319).

The observations of Canger and Flores Farfán have raised awareness about the sociolinguistic complexity of colonial Nahuatl. However, the propositions about H and L varieties, and urban Nahuatl were made without access to the colonial corpora of many different regions, and sociolinguistic interpretations about colonial Nahuatl and CN might change as more data becomes available. Mundane sources outside the Valley of Mexico reveal many of the characteristics claimed by Flores Farfán to be the H variety, or the innovations supposedly found only in Tenochtitlan at the time[6].

Summing up, CN Nahuatl can be seen as both a subset of the corpus of colonial Nahuatl written texts, and as a codified written variety described in colonial prescriptive grammars and vocabularies. Both concepts are closely related to the Nahuatl High variety used in central Mexico in the sixteenth and seventeenth centuries. This work will use CN in both senses, recognising the difficulty of achieving a satisfactory definition. In relation to the contemporary continuum, CN is neither the "original" nor the "best" Nahuatl, despite stand-

[6]Justyna Olko, personal communication, January 2019.

ing out in historical terms. When one considers the contemporary continuum and the CN corpus as a whole, Nahuatl appears best described as a multidimensional linguistic cluster, with variations along synchronic, diachronic and sociolinguistic dimensions.

2.2 Nahuatl morphology

The term *morphology* denotes a part of the language system, namely the internal structure of words. The term also refers to a sub-discipline of linguistics for which two alternative definitions of morphology are possible: 1) morphology is the study of systematic covariation in the form and meaning of words; 2) morphology is the study of the combination of *morphemes*, the smallest meaningful constituents that form words (Haspelmath & Sims, 2013:1-3). In definition 2, morphemes would be subdivided into stems and affixes. The *stem* would be the main morpheme, bearing a concrete meaning, around which the *affixes*, indicating grammatical function, group. Affixes following the stem are called *suffixes* and affixes that precede it are called *prefixes*. The second definition implies the acceptance of morphemes as fundamental building blocks combined according to rules, an assumption that cognitive approaches to morphology do not take as axiomatic. The first definition, therefore, serves better current theoretical discussions regarding morphology. One such discussion focuses on *lexical access* (Haspelmath & Sims, 2013:72), the process of looking up a word in the *lexicon*, i.e. the language user's mental dictionary[7].

The precise nature of morphological processes is not the focus of this dissertation, so the second definition suffices in this work to describe the morphological complexity of Nahuatl. It also conforms with previous descriptions of Nahuatl, which have followed more traditional approaches to morphology. Nahuatl is typically described as an *agglutinative language*, i.e. a language in which almost all

[7]There is no consensus on the precise form of the lexicon, but three major positions can be identified (Haspelmath & Sims, 2013:61). One possibility is that the lexicon contains to the extent possible only individually meaningful parts called *morphemes*, and that virtually all complex words, i.e. words formed by more than one morpheme, are created by rules rather than being listed. Idiosyncratic complex forms, nevertheless, are also listed as entries. This is a *morpheme-based lexicon*. The opposite view, the *strict word-form lexicon*, proposes that all forms are listed in the lexicon, whether they are predictable or idiosyncratic. The intermediary position, the *moderate word-form lexicon*, posits that word-forms, morphemes and derived stems might all be listed in the lexicon, and that the presence of a particular word form depends on a variety of factors.

words are formed of monofunctional morphemes (Matthews, 1991:20-1) whose form is retrievable owing to clear morpheme boundaries (Pirkola, 2001:336).

CN verbs, for example, have a fairly regular morphology (Launey, 2011:69), and they can be expressed, up to a certain point, as morphemes concatenated according to given formulae (Andrews, 2003:46). Compared to languages like English, Nahuatl words contain a typically higher number of morphemes, and Nahuatl is traditionally described as a *synthetic* language. The terms *analytic* and *synthetic* refer to the degree to which a language makes use of morphology: in analytical languages, the role of morphology is relatively modest, whereas in synthetic languages morphology has a more important role. Such a distinction is really a continuum ranging from *isolating* languages, i.e. languages with almost no morphology, to *polysynthetic* languages, i.e. languages with an extraordinary amount of morphology and perhaps many compound words (Haspelmath & Sims, 2013:4-6). The traditional description of Nahuatl as a synthetic language is largely influenced by descriptions of CN, but some modern varieties show increasingly analytical characteristics. Olko et al. (2018) argue that since the colonial period contact with Spanish have driven typological change in the language, and that similarity with an element of Spanish structure have gradually made minor internal patterns more dominant. In extreme cases, it has been claimed, highly Hispanised varieties would be typologicaly closer to Spanish than to CN or other more conservative varieties (Flores Farfán, 2004b).

2.2.1 Nuclear clauses

The most notable linguistic elements of Nahuatl are the morphosyntactic constructs that Andrews (2003) calls a *nuclear clause* (NC) in his description of CN. According to Andrews, all vocables in CN Nahuatl represent NCs, with the exception of the invariant words called particles. NCs are syntactically complete entities obligatorily containing a subject and a predicate, explicit or not. A predicate is not necessarily marked with a verbal stem, for nominal predication is common (see below). NCs consist of a *stem*, nominal or verbal, and inflectional *affixes* arranged in a rigid structure, which means that, although syntactical constructs, they are formed with morphological means.

Andrews notices that NCs are found in many other Native American languages, and because of their morphosyntactic peculiarity have been sometimes

called "sentence-words", i.e. words used as a complete syntactical unit. Andrews, however, highlights that "wordal sentences" or "wordal clauses" would be a more accurate terminology, and that even these terms do not transmit the essential nature of NCs Andrews (2003:45). One tendency, at least in many colonial sources and among academics, is to render Nahuatl NCs in writing as *orthographical words*, i.e. strings of characters delimited by spaces, which explains why they are easily taken as words in this broad sense of the term.

There are two main types of nuclear clauses, *verbal nuclear clauses* (VNCs) and *nominal nuclear clauses* (NNCs) (Andrews, 2003:45-9). The general arrangement of constituents of VNCs for CN is schematised in **Table 2.3**. The prefixes used in VNCs are listed in **Table 2.4**, and the suffixes used to express tense or modality are listed in **Table 2.5**. Launey (2011:227) recognises for VNCs of CN nine basic tense and mood forms, and two directional conjugations. The basic series of tense and mood forms are the present, preterite, future; imperfect, pluperfect, counterfactual; the NI form, which denotes habitual behaviour; the optative, and the vetitive. The two directional conjugations are the directional of motion toward, and the directional of motion away. These eleven series of forms are built on one of three variations of the verb stem called bases by Launey (2011:69). The most notable base is Base2 or *short base*, used to form the preterite. Its formation for one class of verbs in CN[8] involves the dropping of the final vowel –i or –a, and might include as well a change of the consonant preceding the final vowel. The formation of the preterite was used by Canger to define one of her isoglosses (shown in **Figure 2.2**) for the modern continuum, since there are contemporary varieties that form the preterite without dropping the final vowel. Examples of VNCs are *tinechitta* 'you see me', *(o)tinechittac* 'you (Sg) saw me', *tinechittaz* 'you will see me'; and *niquiza* 'I go out', *(o)niquiz* 'I went out', *niquizaz* 'I will go out'.

The general arrangement of constituents of NNCs for CN is schematised in table **Table 2.6**. The prefixes used in CN NNCs are listed in **Table 2.7**, and the suffixes used to express size/attitude and state –possessed or unpossessed- are listed in **Table 2.8**. It is worth noticing that nominal predication is common, and that the notion 'to be' is implicit in NNCs and thus the nominal stem itself serves as predicate in the present tense, receiving the subject prefixes just like a verbal one does. To grasp what a Nahuatl NNC really signifies, one should consider that *cihuatl*, for example, does not mean simply '(a) woman', but 'to

[8]Canger (1980) tags this class of verbs as *Class IV*. See table **Table 4.3**

be a woman' (Launey, 2011:18). Examples of CN NNCs are *nitlacatl* 'I am a man'; *nimotlacauh* 'I am your (sg.) man'; *ticihuatl* 'you are a woman'; *cihuatl* '(she is a) woman', *cihuatzintli* '(she is a) beloved woman; *tinocihuatzin* 'you are my beloved woman (i.e. wife)'; *ocuilin* '(it is a) worm'.

Relational nominal nuclear clauses (RNNCs) are one notable subgroup of NNCs used to express adverbial modification. RNNCs are built using what Andrews (2003:445) calls *relational noun stems*, a small number of noun stems that have relational meanings, and which, when incorporated in NNCs, denote adverbial relations in terms of place, time, etc. Some examples of relational nominal stems are listed in **Table 2.9**. Launey treats these relational noun stems like suffixes, but he highlights that they are "a somewhat peculiar sort of noun that can either be put in the possessed construction or form a compound noun by being added as a suffix to a regular noun stem" (Launey, 2011:240). RNNCs are not used with a subject other than the third person singular.

RNNCs perform in Nahuatl the function that prepositions perform in languages like English and Spanish. In consequence, they are usually translated into these languages using prepositions, a practice that Andrews regrets, because RNNCs are thus easily taken to be prepositions. For example, *nohuan* 'it is in my company', i.e. 'in my company' is normally translated as 'with me', and thus *–huan* could be erroneously taken as the direct equivalent of 'with'. Other examples of RNNCs are *caltitlan* 'it is in the vicinity of the house', i.e. 'near the house'; *calihtic* 'it is at the belly of the house', i.e. 'inside the house'; and *icalihtic* 'it is in the belly of his/her house', i.e. 'inside his/her house'.

PREFIXES							STEMS	SUFFIXES
(Augment)	*Subject*	*Definite Object*	*(Directional)*	*Reflexive*	*Indefinite Object (People)*	*Indefinite Object (Things or animals)*	***Verbal Stem***	*Tense/Modality and number*
		PREDICATE						
	SUBJECT							*SUBJECT*

Table 2.3: The general outline of VNCs for CN based on Andrews (2003) and Launey (2011).. The subject is expressed with a subject prefix which has to agree in number with a suffix expressing tense or modality. According to the conceptualisation of Andrews (2003) more granular analyses are possible, and thus the suffixes, for example, can be further analysed into constituents expressing the tense/modality and the number. The present work has chosen an intermediate level of analysis where suffixes are treated as one unit.

PREFIXES						
(Augment)	*Subject*	*Definite Object*	*(Directional)*	*Reflexive*	*Indefinite Object (People)*	*Indefinite Object (Things or animals)*
	Subj1Sg+ : **n(i)**	Obj1Sg+ : **nech**		Reflex1Sg+ : **no**		
	Subj2Sg +: **t(i)**	Obj2Sg+ : **mitz**		Reflex2Sg+ : **mo**		
	Subj3Sg+ : **∅**	Obj3Sg+ : **c** or **qu**	DirectionThere+: **on**	Reflex3Sg+ : **mo**		
Augment : **o**					IndefObjTe+: **te**	IndefObjTla+: **tla**
	Subj1Pl +: **t(i)**	Obj1Pl +: **tech**	DirectionHere+: **hual**	Reflex1Pl+ : **to**		
	Subj2Pl+ : **am**	Obj2Pl+ : **amech**		Reflex2Pl+ : **mo**		
	Subj3Pl +: **∅**	Obj3Pl+ : **quim**		Reflex3Pl+ : **mo**		
	Subj2SgOPTATIVE+ : **xi**					
	Subj2PlOPTATIVE+ : **xi**					

Table 2.4: Prefixes that can appear in VNCs for CN based on Launey (2011).. The orthography used in the table follows a Spanish-like orthography that does not mark vowel length. The augment can appear with the preterite and pluperfect tenses. The –i- of the Subject prefix can be dropped before a vowel. The directional prefix is also optional, but "can be found at the front of virtually any verb, not merely verbs of motion" (Launey, 2011:52). The appearance of object prefixes is limited to one, except for bitransitive verbs, which can take two objects (Launey, 2011:178).

SUFFIXES used with **Base1** of verbal stems e.g. **quiza**, from *quiza*, 'to go out'; **xima**, from *xima*, 'to shave''
Present
+PresSg : **Ø**
+PresPl : **h**
Imperfect
+ImperfectSg : **ya**
+ImperfectPl : **yah**
NI form (habitual behaviour)
+NIformSg : **ni**
+NIformPl : **ni(me)h**
SUFFIXES used with **Base2** of verbal stems e.g. **quiz**, from *quiza*, 'to go out'; **xin**, from *xima*, 'to shave'
Preterite
+PreteriteSg : **Ø** (for Drop verbs) **c** (for NoDrop1 verbs)
+PreteritePl : **queh**
Pluperfect
+PluperfectSg : **ca**
+PluperfectPl : **cah**
Vetitive
+VetitiveSg : **h** (for NoDrop1 verbs) **Ø** (for other verbs)
+VetitivePl : **tin** or **tih**

(a) Base1 is used to form the Present, Imperfect and NI form. Base2 is used to form the Preterite, Pluperfect and Vetitive

Table 2.5: The eleven basic series of suffixes that can appear in VNCs for CN (Launey, 2011). Suffixes follow one of three variations of the verb stem called bases. Base1 and Base3 tend to be the same, whereas Base2 might differ from Base1 by dropping the final vowel, and in some cases by changing the preceding consonant.

SUFFIXES used with **Base3** of verbal stems e.g. **quiza**, from *quiza*, 'to go out'; **xima**, from *xima*, 'to shave'
Future
+FutureSg : **z**
+FuturePl : **zqueh**
Optative
+OptativeSg : **Ø**
+OptativePl : **can**
Counterfactual
+CounterfactualSg : **zquia**
+CounterfactualPl : **zquiah**
Directional Conjugations
Motion Toward
+ImperfTowardsSg : **quiuh**
+ImperfTowardsPl : **quihui**
+PerfTowardsSg : **co**
+PerfTowardsPl : **coh**
+OptTowardsSg : **qui**
+OptTowardsPl : **quih**
Motion Away
+ImperfAwaySg : **tiuh**
+ImperfAwayPl : **tihuih**
+PerfAwaySg : **to**
+PerfAwayPl : **toh**
+OptAwaySg : **ti**
+OptAwayPl : **tin** or **tih**

(b) Base3 is used to form the Future, Optative, Counterfactual and two Directional Conjugations

Table 2.5: The eleven basic series of suffixes that can appear in VNCs for CN (Cont.)

PREFIXES		STEMS	SUFFIXES	
Subject	*Possessive*	***Nominal Stems***	*Size/attitude*	*State (Possessed vs Unpossessed) and number*
	PREDICATE			
SUBJECT				*SUBJECT*
RELATIONAL NOMINAL NUCLEAR CLAUSES (RNNCs)				
Type 1	*Possessive*	***Relational Nominal Stem** (position, direction, instrumentality, etc.)*		
Type 2	*Possessive*	***Nominal Stems***	*Size/attitude*	***Relational Nominal Stem** (position, direction, instrumentality, etc.)*
Type 3	*Possessive*	***Nominal Stems***	*-ti-*	***Relational Nominal Stem** (position, direction, instrumentality, etc.)*

Table 2.6: General outline of NNCs for CN based on Andrews (2003) and Launey (2011). The appearance of the size/attitude suffix is optional. The appearance of the possessive depends on the state expressed by the clause, but is mandatory with type 1 of RNNC. RNNCs are a notable type of NNCs as they express adverbial relations of time, place, etc. by means of a relational nominal stem. The relational nominal stem is incorporated in one of three ways: directly to a possessive prefix (Type 1); to another nominal stem, which can bear a size/attitude suffix (Type 2); to another nominal stem by means of the ligature **-ti-** (Type 3).

PREFIXES	
Subject	***Possessive***
Subj1Sg+ : **n(i)**	Poss1Sg+ : **no**
Subj2Sg +: **t(i)**	Poss2Sg+ : **mo**
Subj3Sg+ : **∅**	Poss3Sg+ : **i**
Subj1Pl +: **t(i)**	Poss1Pl +: **to**
Subj2Pl+ : **am**	Poss2Pl+ : **amo**
Subj3Pl +: **∅**	Poss3Pl+ : **im**
	PossIndet+ : **te**

Table 2.7: Prefixes that can appear in NNCs for CN according to Launey (2011). The –i- of the Subject prefix can be dropped before a vowel.

SUFFIXES	
Size/Attitude	*State*
+HonDimSg : **tzin**	+SgPoss : **uh** or **∅**
+HonDimPl : **tzitzin**	+PlPoss : **huan**
+DimSg : **ton**	+SgUnposs : **tl**, **tli**, **li**, **in**, or **∅**
+DimPl: **toton**	+PlUnposs: **h**, **tin**, **meh**
+Depr : **zol**	
+DimPitySg : **pil**	
+DimPityPl : **pipil**	
+AugSg : **pol**	
+AugPl : **popol**	

Table 2.8: Suffixes that can appear in NNCs for CN according to Launey (2011). The form of the singular unpossessed depends on the ending of the nominal stem. The deprecatory –zol appears only with inanimate nouns, which in CN appear only in the singular.

Relational nominal stem	*Approximate interpretation*
+Rel_co :**c(o)**	'on, in', e.g. *tepec*, 'on the mountain'; *calco*, 'in the house'.
+Rel_ca :**ca**	'by means of', e.g. *noca* 'because of me'; *chimaltica*, 'with shields'; *tetica*, 'with rocks'
+Rel_huan : **huan**	'in company of', e.g. *amohuan*, 'in your(Pl) company'; *mohuan*, 'in your(Sg) company'
+Rel_ihtic : **ihtic**	'within, inside', e.g. *iihtic*, 'inside him/her/it'; *calihtic*, 'inside the house'
...	...
+Rel_tech : **tech**	'on, touching, in relation to', e.g. *caltitech*, 'onto the house'; *motech*, 'regarding you, in relation to you'
+Rel_tlan :**tlan**	'alongside of, beside', e.g. *notlan*, 'beside me', *nocaltitlan*, 'beside my house'

Table 2.9: Some examples of relational nominal stems and their approximate interpretation. Based on Launey (2011), Wright-Carr (2007).

We recognise that NCs have been conceptualised by Andrews, whose work has drawn upon the CN High corpus (Flores Farfán, 2010a). This work acknowledges also that, according to Flores Farfán (2004b, 2012) and Olko et al. (2018), the contact with Spanish over centuries seems to have influenced a typological change of many contemporary varieties, manifested in a reduction in the productivity of incorporation and compounding. These changes imply, for example, that a number of RNNCs might have been gradually substituted by Spanish prepositions; and that the subtleties and complex concepts expressed by a compound stem in CN could now be expressed supported by a variety of Spanish linguistic means. Olko et al. argue that the borrowing of Spanish prepositions *para* 'for' and *hasta* 'until', beginning in the first half of the seventeenth century, has led to the development of adpositions, via the grammaticalisation of RNNCs. In adition, the borrowings of *para* and *hasta* has led to the development of new kinds of complement clauses: purpose and temporal (Olko et al., 2018:482-9).

In general, however, the presence of NC is evident in any contemporary text, and the relevance of the concept of NC allegedly extends to contemporary varieties as well. Although the differences between contemporary varieties cannot be underestimated, and of them with CN, NCs still seem a crucial structure in the Nahuatl cluster.

2.3 Nahuatl LR and literacy

2.3.1 Research on the diversity of contemporary Nahuatl

Research on CN has obviously been restricted to written sources, and thus has produced many valuable linguistic, literary and historical studies to support the study of old texts (see, for example, the review of Wright-Carr, 2007:25). In contrast, research on contemporary Nahuatl has focused on the spoken language. It is worth noting that knowledge about contemporary varieties, nevertheless, has proven useful to envisage explanations for features otherwise obscure in old texts. Examples are the formation of the applicative of verbs (Canger, 1980), the vowel length and the glottal stop (Karttunen & Amsler, 1983; Karttunen, 1992).

Linguistic research on contemporary Nahuatl has largely focused on the spoken language, and has been approached mainly with a sociolinguistic, anthropological or dialectological perspective. The focus on spoken forms so far is in principle explained by the sparseness of contemporary texts and the developing literacy, but they also seem influenced by the importance given to the oral component of Nahuatl culture. In fact, for a long time there was a strong tendency to associate anything coming from Nahuatl communities or authors, written texts included, with anthropologists' work (Galarza, 1992:227). Discussions around contemporary literacy have expressed concerns about the potential gap between written attempts and the spoken language, or about the impact of writing on the spoken varieties. There is, for example, a certain scepticism about the usability and intelligibility of texts produced by intellectuals or the textbooks issued by the government, and it is stressed that such essays are far from 'lay' speakers' language or interests (Flores Farfán, 2007, 2009). In addition, the prescriptive impact written texts could have in Nahuatl communities and their language has been pointed out by Canger (1994). The works of Francis are notable for their investigation of the development of literacy among Nahuatl children, albeit paying overall attention to the potential influence of Spanish on such development (Francis, 1999a,b, 2000a,b; Francis & Navarrete Gomez, 2003; Francis & Nieto Andrade, 2006).

Contemporary Nahuatl has been an interesting topic for sociolinguistics due to its internal variation and its contact with Spanish. The studies of Hill & Hill (1980, 1986) on the Nahuatl spoken around the Malinche volcano, and studies by Flores Farfán in the Balsas River region (Flores Farfán, 1999, 2000, 2009, 2010b)(Flores Farfán, 1999, 2000, 2009, 2010b) identified the linguistic adaptation strategies of Nahuatl speakers caused by their contact with Spanish. Hill & Hill (1980, 1986:57-8) highlighted how difficult it is to trace a division between Nahuatl and Spanish in certain language events, and coined the term "syncretic language" for Nahuatl varieties spoken in one region of Tlaxcala. An important theme in sociolinguistic literature, raised by Flores Farfán, is the divergence between contemporary spoken varieties and the assumed 'canonical' forms described in CN sources, and between conservative and innovative communities of users. In the light of the dynamic contact with Spanish, Flores Farfán (2005), criticises excessive purism in revitalisation efforts, and the negative attitudes towards modern variants resulting from judging their value against CN as the idealised, 'correct' form (Flores Farfán, 2009). Worth mentioning, although not specifically sociolinguistic, are the studies of Lastra de Suárez & Horcasitas

(1976, 1979, 1978) which include a report on the spectrum of attitudes towards Nahuatl in different regions where it is or was spoken.

Sociolinguistic research has created awareness of Nahuatl diversity, and clearly is guided by a conviction that one must counter the discrimination against spoken contemporary varieties, either in relation to Spanish or in relation to CN. Despite this relevant contribution to Nahuatl LR, sociolinguistic research alone does not seem enough to support a transition towards the strong side of Nahuatl LR which implies a trans-regional literate usage of Nahuatl. As argued in chapter 1, sociolinguistic research might even have a slight bias against the written language. Sociolinguistic research, like the one carried out by Flores Farfán, focuses on the distinctiveness of the varieties spoken in specific communities, and on the importance of preserving this spoken diversity; such concerns are undeniable triggers for LR enterprises, but can in time present challenges when the question of trans-local literacy is discussed, as exemplified by INALI's work.

The aforementioned dialectological work of Lastra de Suárez (1986) and Canger (1988, 2011) showed the complex diversity of the Nahuatl continuum. The idea of convergence, in the form of isoglosses, is implicit in these dialectological studies, but its main, explicit concern is the understanding of internal differentiation. On their own, not being specifically intended to inform LR or LP, they support better an idea of how the Nahuatl continuum can be divided than an idea of how it could be approached as a whole to promote a trans-regional literate usage.

It is sociolinguistic and dialectological research which mostly informs our knowledge on contemporary Nahuatl. If this is the case with other indigenous languages in Mexico (and it seems to be), it is not surprising that the official strategy to encourage literacy for LR focuses on the diversity of the spoken usage as departing point to 'design' the written usage.

2.3.2 The official stance: INALI

In 2003 the Mexican federal government published the Ley general de Derechos Lingüísticos de los Pueblos Indígenas (LGDLPI) (General Law of Indigenous Peoples' Linguistic Rights). Its purpose was to regulate the recognition and protection of the linguistic rights of indigenous peoples, and to promote the usage and development of the Mexican indigenous languages. In symbolic terms,

the LGDLPI emphasises that indigenous languages have the same official value that Spanish, and declares them 'national' languages. In more practical terms, the LGDLPI orders all levels of government to warrant the usage of indigenous languages in legal, educational, health and social security institutions, and the mass media. Finally, the LGDLPI orders the creation of the Instituto Nacional de Lenguas Indígenas (INALI).

INALI is the LP agency for indigenous languages in Mexico, and has as one main objective to support the LR of indigenous languages. INALI holds in general the premise that written standardisation is important to change the social status and expand the functions of indigenous languages (INALI, 2012:46-7). The process of norm-creation for writing systems and the resulting orthographical conventions, therefore, are core objectives of its work (INALI, 2012:26). In this enterprise, INALI seeks to combine contemporary linguistic findings with the agreements reached by social actors (INALI, 2009). *Norma*, 'norm', and *normalización*, 'norm-creation' are terms employed by INALI regarding writing, perhaps because they appear less controversial than *standard* and its related terminology, but they seem to be used in the same sense, as Flores Farfán (2017:81) has also noted[9].

Since its creation in 2001, INALI has put considerable effort into outlining a thorough classification of the Mexican indigenous languages, and currently distinguishes thirty Nahuatl varieties. Its provisional classification is based on general criteria of linguistic genealogy, dialectology and sociolinguistics (INALI, 2010:34). With this background informing its work, it is not surprising that INALI is very conscious and respectful of spoken diversity. Thus, it advises against the error of considering that the speech of every indigenous group is so homogeneous as to suppose they speak a single language; furthermore, INALI considers that although the varieties identified around the same indigenous peo-

[9]Although *norm* and *standard* are related concepts, Bermel (2007) makes a distinction between them. He states that a norm is based on evidence about the preferred standing of a feature within a language code or variety; standard, on the other hand, entails approval by opinions considered authoritative by a language community, and there is no need to make reference to its frequency or use. There can be norms without the existence of a standard based on those norms, and it is possible "to speak of norms within, for instance, a dialect, because every community recognizes speech that belongs to it and speech that does not" (Bermel, 2007:5). INALI considers speakers and academics the agents that discuss and make decisions "*acerca de la norma que se establecerá para una lengua*", 'about the norm to be established for a language' (INALI, 2012:25-6). The 'norms' INALI is promoting are 'standards' in as much as they require the approval of speakers and academics. *Norm-creation* and not *normalisation* will be used here to translate the Spanish term *normalización*. Much more than normalise writing, i.e. make writing usual or ordinary for a community, INALI is actually aiming to create norms for writing.

ple show a close structural relation, "*en términos de la genealogía lingüística, la dialectología y la sociolingüística, se trata de lenguas diferentes*", 'in terms of linguistic genealogy, dialectology and sociolinguistics they are different languages' (INALI, 2010:36). INALI, therefore, avoids using the term 'language', adopting instead the category of "*agrupación lingüística*", 'linguistic grouping'. In such a taxonomy, the Nahuatl continuum is considered a linguistic grouping, and sub-divided into the aforementioned 30 varieties.

Although procuring political correctness in the terminology employed, INALI's classification still seems heavily influenced by political, non-linguistic criteria. The role of political criteria in the distinction of 'different languages' cannot be overemphasised. Besides multiple examples like Danish-Norwegian, Serbian-Croatian, Catalan-Valencian and a long etc., the Nahuatl varieties in the region where the states of Puebla, Oaxaca and Veracruz meet, show just how the classification of linguistic varieties can often follow political borders (section **2.1.1**, around **Figure 2.3**).

As fundamental to its work, INALI states that a high degree of internal homogeneity is required to declare a linguistic group to be a 'language'. This statement, however, notably misses the fact that in no 'language' is speech, for example, homogeneous. Even in highly standardised systems like English (Wardhaugh, 1999:97-100) or German (Johnson & Braber, 2008:50-6) there is no fixed standard pronunciation, and the everyday language is often a non-standard regional dialect, leaving aside the consideration of sociolects. Besides, it has been mentioned the role a standardisation process plays in the definition of a language. The common assumption of, for example, English being a single, homogeneous language can largely be related to a relatively homogeneous written version of it (a written standard) superposed onto a spoken continuum. In such a continuum 'ungrammatical' usage and 'non-standard' pronunciations are not uncommon, and at times are so evident that English has been described as "one tongue, many languages" (Svartvik & Leech, 2006).

Other problematic criteria employed by INALI in its classification depend on the perception of speakers. One is the name given by the speakers to their linguistic variety (INALI, 2010:35). This is complex in the case of Nahuatl, which in different regions is called *mexicano tlajtol*, *maseual tla'tol*, or *mexcatl*, just to cite some examples found in the catalogue of INALI. Based on inter-dialectal intelligibility, another problematic criterion, INALI considers that a variety can

be assumed for those localities where the speakers declare that they have mutual intelligibility. It has been proposed for other linguistic continua, however, that dialectal distances perceived by speakers could change depending on factors such as attitude (Watts, 1999), or even on a single very deviant word, sound or morphosyntactic structure (Beijering et al., 2008). In the Nahuatl case, Olko & Sullivan affirm that in two inter-dialectal encounters, a high enough degree of intelligibility became immediately apparent as to allow animated monolingual discussions on a diverse array of topics. This, they believe, counters the "false propaganda regarding the mutual incomprehensibility of modern 'dialects' of Nahuatl" (Olko & Sullivan, 2014:389-91). Olko & Sullivan have organised, until October 2018, seven inter-dialectal encounters, three of which focused on reading colonial texts. Olko sustains that all of these meetings demonstrated not only high mutual intelligibility, but also a very close understanding of colonial texts, and the survival of terms unattested or ambiguously explained in colonial lexicographic sources (Personal communication with Justyna Olko, January 2019).

INALI intends to proceed scientifically and works with the conviction that speakers must be involved in LP. In the normalisation of writing, INALI takes "*fundamentación científica*", 'scientific basis', as one non-negotiable principle, and it seeks to promote decisions based on "*estudios fonológicos serios*", 'serious phonological studies'. According to INALI, "*el principio fundamental para la normalización de las convenciones ortográficas es la alineación de la escritura con la oralidad*", 'the fundamental principle for the norm-creation of orthographical conventions is the alignment of writing with orality' (INALI, 2012:26). The intention is "*evitar que el sistema de escritura del español sea la guía para trabajar con lenguas tipológicas e históricamente diferentes*", 'to avoid having the Spanish writing system as the guide for working with languages which are typologically and historically different' (INALI, 2012:27). INALI also considers it essential, nevertheless, to reach agreements between all speakers and involve them in the making of decisions. Thus, another non-negotiable principle of action states that, once there is adequate basis to prove that two varieties of the same linguistic group cannot share the same normalising decisions, there is no reason to force their respective speakers to proceed in such a way (INALI, 2012:23-8). This approach has been criticised as populist by Olko & Sullivan (2014:386), who have highlighted that the process of consolidating a written norm is more complicated than just voting on an alphabet.

The three versus five vowels debate in Quechua demonstrates the difficulty of reaching LP agreements by voting. Peruvian linguists argued that Quechua should have only three vowels <i, a, u>to represent the phonemes /i/, /a/, and /u/, and that [e] and [o] occur as allophones of /i/ and /u/ when in the proximity of the uvular consonant /q/. In contrast, the members of the Quechua Academy in Cusco insisted on having an alphabet with five vowels <i, e, a, u, o>, among other reasons, because Quechua has been written with five vowels since the colonial period. Unable to reach agreement, voting was used as way to make a decision in two meetings – in 1975 five vowels were approved, whereas three vowels were approved in 1983 - in which the difference to approve a proposal was only one vote. As in most political debates, it seemed that there were always procedural grounds on which the validity of decisions taken might be challenged by one or another group of interest, and thus the cycle of meetings and votes was potentially endless (Hornberger & King, 1998:396). Thus, conciliation between actors seems difficult to achieve by voting. Besides, settling the question of an orthography by appealing to scientific 'objective' studies could also be complicated. As the Czech orthographic reforms exemplify, collateral informal debates can be a messy picture; within them, "rational bean-counting is subordinated to individual aesthetic reactions and beliefs" (Bermel, 2007:149).

INALI's commendable awareness of diversity and its stance on the importance of the spoken language as a point of departure for important decisions like written normalisation is not surprising, taking into account that the research on, for example, contemporary Nahuatl has focused overwhelmingly on the spoken language. However, achieving a careful alignment of writing to the oral usage, respecting at the same time the perceptions of a spectrum of users, implies that INALI might need to design as many written models as there are spoken variations and opinions.

2.3.3 Bottom-up initiatives

There have been various groups of linguistic and political activism in favour of Nahuatl LR, as is shown in one review by Brambila Rojo (2004:137-43). He noticed, however, that one common denominator of this activism was a rather low involvement of the Nahuatl communities themselves.

Currently, Flores Farfán and Olko & Sullivan lead the two most visible initiatives, which are explicitly presented as Nahuatl LR endeavours. Both projects agree on the importance of bottom-up revitalisation, i.e. one originating in the Nahuatl communities and having the speakers as main actors, although the two initiatives have contrasting views on written standardisation.

Flores Farfán (2002, 2007, 2011b, 2013)Flores Farfán (2002, 2007, 2011b, 2013) has promoted audio-visual strategies to go beyond approaches primarily based in schools and written media, and seeks to revitalise traditional practices like painting on amate paper. He presents his approach as sensible because it is based on the revitalisation of traditional culture, and is proposed to overcome the top-down, one-sided approaches of school instruction (Flores Farfán, 2011b:199-201)(Flores Farfán, 2011b:199-201). Such practices seem indeed sensible to trigger a revalorisation of Nahuatl and an interest for LR among the communities. The emphasis on orality and a 'traditional' Nahuatl culture does not seem to counter, however, and might even reinforce the folklorisation of Nahuatl and its association with only anthropologists' work. In addition, the project is generally sceptical towards a shared standardised written usage across the continuum. Flores Farfán maintains that imposed standardisation, and imposed hegemony in general, is a threat to linguistic diversity and to the survival of indigenous languages. For him, the "distorted institutional ideologies" looking to standardise indigenous languages threaten the possibility of "celebrating linguistic diversity" (Flores Farfán, 2017:81).

Flores Farfán has overtly aligned his work with the interests of the 'common man'. He rejects efforts to draw upon CN for contemporary practice, and to assert that Nahuatl is a single language. Such pretensions, he contends, benefit only 'cultural brokers' and the Nahua intellectuals "who study Nahuatl and make a living out of it" (Flores Farfán, 2017:80). His project believes in celebrating internal linguistic diversity, and intends to produce books in as many diverse varieties as possible (Flores Farfán, 2011b:201). Although Flores Farfán has envisaged the relevance of contact between users of different varieties for LR in virtual communities like social media (Flores Farfán, 2017), his approach towards LR echoes the concerns for the preservation of distinctiveness and alignment of written practice to the spoken local varieties.

In contrast, the approach of Olko & Sullivan (2014, 2016a,b) highlights the need for a broad, unified LR effort, and aims to encourage the empowerment

of the speakers through literacy and in close connection with CN. Their idea of effective research and revitalisation has as a fundamental axis the active involvement of speakers as students and researchers in educational, social and political tasks. Olko & Sullivan (2014:391-92) emphasise the importance of contact between speakers from different varieties in broad networks. The most notable feature of their approach is their assertion that it is fundamental to link modern speakers with their past in order to encourage them to actively promote the survival and development of their culture. That is why their most ambitious aim is to revive and extend the literacy developed during the colonial period. They aim to encourage the implementation of one standard orthography across varieties, in close relationship with the older tradition of writing in Nahuatl (Olko & Sullivan, 2014:381-2). They support the publishing of contemporary works from different variants, and of colonial texts in the book series *Totlahtol*. Variations other than orthographic are preserved in these books, following the premise that the standardisation of orthography should not be at the expense of local features. For example, two versions (La Huasteca and Tlaxcala varieties) of the pre-hispanic myth of the Four Suns have been published as *Nahui Tonatiuh* (2015), using the same orthographic conventions.

Olko & Sullivan's experience working with users from different Nahuatl communities, they affirm, counters the idea of CN being a dead language that cannot be understood by modern speakers. Based on their study of archival and field data, as well as on their experience with Nahuatl students, they argue that a close correspondence between CN and many modern variants is easily recognizable, and that speakers of modern dialects do not have trouble reading and understanding colonial manuscripts (Olko & Sullivan, 2014:373, 382, 393).

Regarding a progression towards the strong side of LR, the initiative of Flores Farfán does not seem promising on its own as a way of uniting users of different localities. The accommodation of writing to spoken diversity might be a valuable point of departure for local LR, but as the interest in LR increases within different localities, it is necessary to add potential users along the Nahuatl continuum. Olko & Sullivan's proposal seems more promising for that purpose, as written contact could be a way to overcome geographical dispersion. The acceptance of a standard orthography for overcoming spoken variation, however, seems unlikely in the short term without a central authority to enforce it. Unlike Basque, the Nahuatl continuum does not have at its disposal a widely recognised academy or a political unity over a 'home' region.

Nahuatl users, however, could nowadays contribute to the authentication practice of writing. An incipient literacy in Nahuatl communities, albeit heterogeneous, seems promising for gradually moving Nahuatl LR towards the strong side.

2.4 Nahuatl contemporary writing

Attending only to the panorama presented by most of the dialectological and sociolinguistic literature discussed above, the Nahuatl cluster appears to occur in a pre-literate, oral context, and thus to necessitate designing one or a number of writing systems. Notably, the lack of a recognised standard usage has never stopped Nahuatl writers, and writing is a practice well incorporated in the Nahuatl culture. Olko & Sullivan (2013) notice how the adaptation of alphabetic writing to Nahuatl supported the quick development of written usage in colonial times. The research of McDonough (2010, 2014) on Nahuatl intelligentsia shows that Nahuas have uninterruptedly used alphabetic reading and writing in their first (and even second, and third) language(s) since the sixteenth century. The role of the intellectual along with the genres, forums, tools, and discursive codes he/she used has changed as the needs of their specific communities evolved, but it is certain that writing has long been indigenous (McDonough, 2014:11-21).

Orality might be a key feature of Nahuatl culture, but many young Nahuatl speakers are not necessarily illiterate, at least in Spanish, according to the Mexican General Census of 2000. Literacy rates reported then seemed relatively high among young speakers (83.2% in 15-29 years-olds), and the percentage of 8-14 years-olds able to read and write (85.2%) were 10 points below the national percentages of Spanish monolinguals of the same age (INEGI, 2005:47-9). The figures can be attributed to the bilingual education programmes established in the 1960s by the *Secretaría de Educación Pública*, Ministry of Public Education (SEP). Bilingual education programmes have in practice faced many obstacles to promoting both the use of indigenous languages and Spanish. Among others, parents placed little value on Nahuatl literacy (Rolstad, 2001:14-6), and not surprisingly schools and written media easily turned into instruments to acquire Spanish (Flores Farfán, 2002:231). However flawed or controversial the bilingual education programmes might be, the figures from INEGI suggest that literacy rates might be increasing among younger generations.

The report on the census of 2000 clarifies that the reported literacy levels can be considered to refer in practice to Spanish (INEGI, 2005:43). Francis & Nieto Andrade (2006), however, conducted a study on literacy acquisition in a school where no formal teaching programme in Nahuatl literacy existed. They estimated that in an atmosphere of 'active tolerance' the cognitive/academic abilities gained for Spanish do transfer to Nahuatl: "even under assuredly unfavourable conditions of language conflict, marginalization, and displacement of the vernacular, underlying competencies are still available to the young bilinguals in both languages" (Francis & Nieto Andrade, 2006:150). Preliminary comparisons from student's writing samples suggested that they exploited rather successfully the typical advantages of the written modality, e.g. opportunity to plan, reflect and revise, and more processing time (Francis & Nieto Andrade, 2006:153). Francis & Nieto Andrade noticed, nevertheless, that the transfer of literacy competences from Spanish to Nahuatl is neither automatic nor assured, and that a long term trend pointed to an eventual erosion of skills in Nahuatl (Francis & Nieto Andrade, 2006:149). It seems then that although instruction in Spanish does not necessarily run counter to the development of Nahuatl literacy, the latter does require successive practice and support to consolidate.

There are enough examples of contemporary Nahuatl texts to safely assume that literacy, however incipient or first gained through Spanish, is not strange in contemporary Nahuatl communities, especially since more positive attitudes towards the language have emerged. Literary, official and academic texts are only some examples of the slow but steady increase in the production of Nahuatl texts derived from a renewed interest, official or not, on using Nahuatl in writing. Nahuatl writers from different regions have been active since the mid-twentieth century (Hernández, 2015; León-Portilla, 1986). Besides these literary texts, the federal government has promoted the translation of official texts by Nahuatl translators as part of the linguistic policies begun in 2003 with the LGDLPI. The products of these endeavours, especially the translations, are criticised by Flores Farfán (2017, 2009) as top-down approaches which are "very limited in terms of intelligibility and use" (Flores Farfán, 2017:74). Therefore, in line with his sociolinguistic approach, he has led a series of publications compiling traditional stories and riddles from the Balsas River Region, in an attempt to make them closer to the "layman" and not only to an intellectual elite which might cultivate an excessive purism. Another core objective of Flores Farfán's bottom-up approach is to create written materials which actually have a social function because, he emphasises, this is the only way in which writing can acquire

any relevance in the communities (Flores Farfán, 2009:148). The approaches to Nahuatl literacy might differ, but writing is arguably a significant component of Nahuatl culture, and of Nahuatl LR endeavours in spite of their respective ideological assumptions.

As old prejudices against indigenous languages recede, the spoken and written academic usage of Nahuatl is also starting to be promoted. Academic spaces like the IDIEZ and the Universidad Veracruzana Intercultural (UVI), a branch of the Universidad Veracruzana[10], are noteworthy for encouraging students to conduct academic discussions in Nahuatl, thus making it a "vehicle" language and not only an object of study. Notably, in 2015 Eusebia Martinez Silva wrote and defended her master's dissertation in Nahuatl (Universidad Veracruzana, 2015:14); and so did Eduardo de la Cruz Cruz and Abelardo de la Cruz de la Cruz in 2016 at the IDIEZ.

Academic spaces like the IDIEZ and the UVI are important because they allow for the face-to-face contact between speakers of different Nahuatl varieties, and also because the academic usage of Nahuatl encourages writing in Nahuatl as well. Pharao Hansen (2016:358-62) reports how speakers from different Nahuatl communities (including students, faculty and non-academic staff) concurred at the Tequila campus of the UVI. Nahuatl was initially merely part of the curriculum: a language to be studied in a single semester. However, members of the faculty, one of them a PhD Nahuatl speaker, began encouraging the students to use Nahuatl in the classroom, in the halls, in presentations at a monolingual colloquium, in contributions to a bilingual magazine and other activities. This engagement in Nahuatl in an academic space and for academic purposes boosted the enthusiasm of the students to use Nahuatl in and outside the classrooms. It also led to discussions among themselves about how to best express a specific context in Nahuatl, or how to best spell it (Pharao Hansen, 2016:362). The idea of turning Nahuatl students into researchers and creators of knowledge, rather than mere consumers, encourage written works like the monolingual Nahuatl dictionary compiled by students of the IDIEZ (Sullivan et al., 2016).

Nahuatl, thus, is being written, albeit for varied purposes, by writers with varied ideological backgrounds and different levels of literacy, and maybe even

[10]In the Universidad Veracruzana there is a project to launch an MA course completely taught in Nahuatl, although the lack of teaching materials in many subjects is seen as a first issue to tackle. It is to be hoped that the newly regained confidence of Nahuatl professionals in using their language will help to alleviate this shortage in the middle term.

different levels of knowledge about the structure of the language (Flores Farfán, 2017:78). Two important implications are to be noted. First, it seems that the need and desire to write in Nahuatl (supported by the literacy priming provided for better or worse by Spanish, and encouraged in academic spaces) are getting stronger despite the lack of a standard orthography, or other norms.

Besides, like in the Romani case (Matras, 2015), basic literacy and technology might actually be contributing to overcoming the geographical dispersion of Nahuatl speakers. The written contact between users from different regions, a main objective of (Olko & Sullivan, 2014:390-4) is now possible thanks to the internet. In line with their concerns, they have created a closed Facebook community where participants are encouraged to communicate exclusively in Nahuatl. computer-mediated written interactions have also caught the attention of Flores Farfán as bottom-up initiatives in which divergence in orthography and dialectal variations do not hinder communication. Notably, he also envisages that such written interactions might eventually produce mixed varieties, and that inter-dialectal convergence and poly-dialectism could emerge (Flores Farfán, 2017:76, 81). He highlights that, although the overwhelming referent for 'real-life' literacy is mostly Spanish, these Facebook users are more immediately concerned with communicating, gaining visibility and affirming an identity, than with an orthographic choice (Flores Farfán, 2017:77)[11].

Writing is thus gradually acquiring relevance to fulfill social functions from the bottom, and not only as a top-down goal. This work agrees with Flores Farfán's observation that such contemporary interactions, and a review of the history of Nahuatl literacy, show that for Nahuatl people the choice to write in Nahuatl has been more important than the question of how to write in Nahuatl.

[11]Flores Farfán actively participated in the last inter-dialectal encounter organised by Olko & Sullivan in October 2018 (Personal communication with Justyna Olko, January 2019)

2.5 Bringing CN into LR and LP: NCs as points of convergence

2.5.1 (Re)imagining 'a Nahuatl' community

Literacy, however incipient or heterogeneous, is important for putting users from different varieties of the Nahuatl continuum in contact with each other. Significantly, as Olko and Sullivan's approach to LR emphasises, literacy is also an opportunity for Nahuatl speakers to have contact with the written heritage contained in the CN corpus and with old documents in general. This is a very relevant perspective towards Nahuatl LR: from a linguistic point of view, it approaches Nahuatl as a whole cluster, and not only as a dialect continuum or a collection of separated varieties. From an extra-linguistic point of view, the contact between speakers of different communities could strengthen a trans-local Nahuatl identity to form wide political networks.

Contact between users of different Nahuatl varieties depends of course on their willingness to engage in the first place. This raises the question of whether there is a sense of Nahuatl ethnolinguistic belonging. Pharao Hansen (2016) has identified four language ideologies in Nahuatl LR. Among them, the ideologies Pharao Hansen calls *localist purism* and *pan-Nahuan purism* reveal contrasting views on the idea of 'a Nahuatl community' among Nahuatl speakers. Localist purism sees any identifiable local variety as a language in itself, and as the unique symbolic vehicle for social cohesion and continuity at the community level (Pharao Hansen, 2016:170-3). According to Pharao Hansen, the type of LR projects showing localist purism are community-based, and involve speakers who often have limited educational background, and have little acquaintance with or interest in indigenous affairs beyond their local community. Pan-Nahuan purism, on the other hand, seems more commonly expressed by Nahuatl intellectuals, and seeks to promote standardisation and dialect levelling to create an imagined community encompassing all Nahuatl speakers. The LR project of Olko & Sullivan at the IDIEZ is associated by Pharao Hansen with pan-Nahuan purism. It seems thus that a sense of supra-local belonging among Nahuatl speakers is still developing, and that its consolidation will largely depend on a balance between the localist and pan-Nahuan approaches.

There is, according to Pharao Hansen (2016:385), a positive side of localist purism. He considers that linguistic vitality is better maintained in communities where localist purism creates a local public sphere that resists the mainstream public sphere. Localist purism, he argues, transform the local linguistic variants into symbols of a political community. By 'mainstream public' Pharao Hansen mainly means the Mexican Spanish-based public sphere. However, he also considers that some pan-Nahuan supporters might have lost a connection to their communities of origin, and are now part of the mainstream Mexican public. Pan-Nahuan approaches are, in his view, less conducive to LR than localist purism. It is the ties to local communities more than the ties to a wider historical pan-ethnic community, he believes, that makes the language valuable to most of their speakers and has maintained it alive (Pharao Hansen, 2016:73). In this way, Pharao Hansen highlights the value of the experiential communities built via face-to-face engagements for political and LR endeavours, in a similar way in which Fishman highlights the importance of 'real' communities in the intergenerational transmission of the language (see **section 1.3.2.1**).

As it has been argued, however, 'virtual' communities seem eventually necessary in LR to authenticate new forms of language, and to transcend to the strong side of LR. Likewise, and in close connection with the linguistic side of the matter, the participation of Nahuatl users in the imagined community promoted by pan-Nahuan approaches is arguably promising to authenticate new ways of belonging that may strengthen politically Nahuatl communities. Pharao Hansen argues, for example, that Mexican nationalism has expropriated the Nahuatl culture to build a 'Mexican Nation'. He also recognises, on the other hand, that different degrees of chauvinism caused by localist purism has the risk of hindering cross-community collaboration and joint political projects of speakers of different varieties. It is true that pan-Nahuan purism has the risk of promoting ideologies of homogeneity and supremacy, as Pharao Hansen argues; but even if it is not free of risks, a pan-Nahuan approach seems a chance to involve many Nahuatl users in the authentication of new forms of language and community. The imagined community promoted by a pan-Nahuan approach seems useful to establish trans-local political communities that can re-appropriate 'a Nahuatl' identity, and that are strong enough to work for the benefit of the Nahuatl people. The present work accepts the core importance of the experiential communities for Nahuatl LR. It is only argued that an imagined Nahuatl community could complement the experiential communities that have kept the language alive so far, as well as strengthen Nahuatl speakers in political terms.

An imagined pan-Nahuan community may be not only a space to negotiate and re-create forms of language, but also forms of belonging. Like the standardisation of Norwegian, Nahuatl LR is an enterprise which involves reimagining the composition of a community around language. Like Norwegians, who were looking to differentiate themselves from Danish using language, Nahuah might be looking to differentiate themselves from the Mexican mainstream society by reclaiming and revitalising Nahuatl language. But just like in the Norwegian case, this aiming for differentiation in Nahuatl LR can backfire if only experiential, 'real', 'authentic' forms of language and community are promoted, and seen as irremediably confronted with the imagined, 'unauthentic' forms of language and community of an intelligentsia influenced by the mainstream 'exterior'. As it has been argued in **sections 1.1.5** and **1.2.1**, drawing limits with respect to an 'exterior' based on an emphasis on 'authenticity', can nurture also divisions within a potential community of language users.

How is it possible to find a balance between a local component and a mainstream component for LR and for the configuration of a wide Nahuatl community? Maybe the answers can be best envisaged by Nahuatl users participating in what Pharao Hansen calls *spaces of possibility*, where approaches to the relations and hierarchies of the everyday can be challenged and experimented with. The intercultural ideology in spaces like the UVI, he argues, allows for the negotiation and not only enforcement of cultural boundaries, and for the choice and combination of two cultural spheres (Pharao Hansen, 2016:380). It is especially relevant to highlight that such negotiation should happen not only in relation to the Spanish-speaking mainstream Mexican sphere, but also in relation to other Nahuatl-speaking local spheres.

The idea of spaces of possibility agrees with the idea of a community of practice, whose importance has been noted for the potential development of written conventions (**section 1.3.2.2**). Academic spaces-of-possibility/communities-of-practice like the UVI and the IDIEZ are important because they foster face-to-face interactions in a context that boosts the need, confidence and desire to engage in a literate 'virtual' sphere. The consolidation of projects like the UVI and IDIEZ are thus of the utmost importance to authenticate new forms of language, spoken and written. Moreover, they are key opportunities for speakers of different varieties to imagine and shape together 'a Nahuatl language' alongside 'a Nahuatl communtiy'. Just as the communities of fellow-readers were the embryo of the nationally imagined communities (Anderson, 2006:42-4), the literate

Nahuatl community currently emerging in spaces like the UI and the IDIEZ could be the seed of an imagined Nahuatl community.

2.5.2 CN to bridge across the Nahuatl cluster

One cannot deny the difficulties related to the linguist and social –e.g. educational– differences in the Nahuatl continuum, and recognise that these differences will ultimately condition the number of users that can participate in an imagined Nahuatl community founded on Nahuatl literacy. The overlap between the old and emerging traditions of writing, however, could offer a common ground to build upon to benefit as many Nahuatl users as possible, and to encourage uniform writing to the degree that it is possible at this point of the history of Nahuatl literacy. That is where the importance of CN resides, and this work concurs with Olko & Sullivan on the importance of bringing CN into LR.

Besides the emotional connection to an experiential community, built upon a local variety, the emotional connection to an esteemed past, built upon CN, can strengthen the appealing of LR among Nahuatl communities. The CN corpus is valuable because it can be used to return to contemporary speakers a sense of value and pride about themselves and their culture. McDonough's experience with Reading Circles[12], where old Nahuatl texts are discussed, suggests that access to cultural heritage sources has the potential to increase the self-esteem of Nahua community members, and to reinforce a positive image of their language and culture. Equally important, this repatriation of old texts can counter the effects of a "cultural bomb" intended to annihilate a people's belief in their language, in their capacities, and ultimately in themselves (McDonough, 2014). It is notable that many Nahuatl speakers have no idea of the existence of a prestigious written heritage, and the pleasure when they realise their connection with an ancient, esteemed culture is noteworthy (see e.g. Olko & Sullivan, 2016a:173; Rolstad, 2001:11). Thus, making Nahuatl people aware of this connection is one step towards the re-appropriation of symbols that have been exploited only by the Mexican Nationalism.

[12]In these Reading Circles, organised by McDonough, old Nahuatl texts were read and discussed by Nahuatl native speakers who were undergraduates or had degrees in law, accounting, linguistics and social sciences. Although such experience could be (unfairly) criticised as elitist because of the profile of the participants, arguably these kind of exercises can be also attempted among the general public in Nahua communities, expecting to achieve similar positive results.

Equally important, the CN corpus is a rich source of examples to support contemporary literacy, and it makes sense to take advantage of its points in common with the contemporary continuum. Debates around writing give the impression that no previous usable literary tradition exists for Nahuatl. As Olko & Sullivan point out, existing scholarship does not sufficiently highlight the continuity between contemporary local varieties and old Nahuatl; and yet this continuity is essential to connect the entire cluster. Among other reasons, "usually [...] researchers specialize exclusively either in colonial or in modern data, making no attempt to connect these phenomena by seeing the language and cultural development over the long-term" (Olko & Sullivan, 2016a:163).

The possibility of bringing historical Nahuatl into LR has not been ignored by Flores Farfán (2002:233) either. His sociolinguistic wariness, however, exemplifies how the gap between past and present, can be further widened by authenticity-related tensions, or by a distrust of intellectuality. It is undisputable that contemporary varieties should not be devalued in comparison with CN. Nevertheless, attempts to recover CN forms should not be entirely disqualified either, however artificial they initially might look in comparison with the 'real' language of the layman. The gap between past and present should not be broadened either way, nor feed the impression that present-day indigenous culture has no past, and that past culture has nothing to do with the lives of indigenous people today (Olko & Sullivan, 2016a:173).

Nahuatl LR can benefit from attending to the overlap not only between contemporary varieties but also of them with the CN corpus. Exercising sociolinguistic awareness, the CN corpus can play a positive role in LR, not as an authentic norm to judge the value of contemporary usage, but as a common resource to boost pride and provide examples to the developing literacy along the continuum. Besides exercising sociolinguistic awareness to counter discrimination, Nahuatl LR should focus on the overlap between written contemporary varieties and CN to attenuate the underlying isolationism and/or mundaneness of authenticity-based approaches to LR.

2.5.3 Study convergence in writing using NCs

If a focus on the differences in the spoken language implies, as INALI's work suggests, a subsequent subdivision of a continuum to accommodate literacy to

spoken diversity, maybe the discussion about literacy for Nahuatl LR should not rely only on insights gained from the study of the spoken language. Nowadays, Nahuatl LR initiatives could also learn from the overlaps in the increasing written practice as another source of knowledge to *support* rather than just *regulate* writing in the short term across the continuum.

Olko & Sullivan argue that close correspondences between contemporary and old Nahuatl are easily recognisable at the lexical, morphological, syntactical and phonological levels (Olko et al., 2018). An emphasis on the morphological features common to all varieties rather than on their divergent sounds, they argue, could facilitate inter-dialectal communication, make colonial texts accessible to contemporary readers, and be a window to the structure of the language (Olko & Sullivan, 2014:386).

The potential importance of morphological overlaps could thus help to deal with the authenticity-based divergence manifested in divergent orthographic practices. Arguably, in spite of the set of characters used by a writer, encouraging a more abstract understanding of the Nahuatl linguistic structures among its users could prove useful. Flores Farfán (2009:139-148, 2017:76-9) has noticed the trend to treat written Nahuatl as written Spanish by writing bound morphemes as separated pronouns (Flores Farfán, 2009:139-48; 2017:76-9). Although he refrains from criticising such usage, he hints at the desirability of facilitating a reflection on the structure of the language and its history.

NCs are a central morphological concept for studying a plausible overlap of the contemporary written practice in Nahuatl with the CN corpus. NCs seem to represent a concept applicable to any Nahuatl variety, notwithstanding that it was developed based on a study of CN. In addition, taking into account the morphological complexity of Nahuatl varieties, NCs need to be systematically recognised and analysed in contemporary writing before extending the exploration of convergence to other levels, e.g. lexical and semantic. NCs thus are not only a plausible link between all the varieties of the language, but also a first point of concern towards the extensive analysis of contemporary writing at different linguistic levels.

To persuade users of the convenience of a given orthography might not be easy in the short term. It must be acknowledged that an orthography could bear emotional attachment to a local Nahuatl variety (Pharao Hansen, 2016:192), and that the influential actors involved in the practice of writing, namely published

authors and publishers, tend to be, for good or bad, the ultimate selectors and shapers of conventions. In the meantime, the investigation of convergence could rely on the analysis of NCs as a linking structure, below the orthographical divergence, of contemporary written varieties and the CN corpus.

2.6 Conclusions

Awareness of spoken diversity is commendable to avoid discrimination; INALI's work, however, exemplifies the complications caused by an exclusive focus on the spoken language as basis for the design of literacy for LR. A written usage based mainly on descriptions studying the spoken diversity might subdivide the Nahuatl continuum into small communities of users, thus hindering a trans-local literate usage that is necessary to take Nahuatl LR to the strong side of the GIDS scale. Despite the lack of recognised standards, Nahuatl users are writing, and they might need not a written norm as much as they need to keep writing. It cannot be overemphasised that a community of practice of Nahuatl writers is necessary to develop conventions through practice rather than by explicit design or agreement. In the meantime, the exploration of Nahuatl contemporary written practice is an opportunity to complement the body of knowledge available to support literacy for Nahuatl LR. Such exploration could, for example, be the basis for creating pivotal resources that could be more or less adaptable to assist, rather than just to regulate, writing along the continuum.

Grenoble & Whaley (2006:136) mention technical support as one of four kinds of support needed to begin creating literacy in a pre-literate community. Literacy in the Nahuatl continuum already exists, however incipient, influenced by Spanish, or heterogeneous it might be. Therefore, the technical support needed for Nahuatl LR is already beyond the design of a suitable orthography; it must focus on how to encourage and study convergence in the written practice.

In the study and encouragement of convergence, CN seems a valuable asset, providing examples of previous written usage and boosting the pride and interest of Nahuatl users when they access their heritage. NCs are the morphological structures that can support the investigation of convergence between contemporary writing and CN despite orthographical divergence. NCs need also to be the first focus of an extensive exploration of convergence in the written practice at different linguistic levels.

The next chapter gives an overview of the resources currently available to attempt a systematic, large-scale exploration of the overlaps between contemporary texts using a CN model as pivot. It will be argued that the current orthographic divergence hinders the compilation of large data sets that could be used in a probabilistic approach to analyse NCs. It will also be argued that a description of CN, for which extensive grammatical studies and dictionaries written with a fairly standardised orthography are available, can be formalised in a Finite State computational model.

Chapter 3

Resources and challenges to explore contemporary written Nahuatl: tackling morphological complexity and orthographical variation

The recognition and analysis of NCs is a fundamental task for any attempt to explore Nahuatl written texts. In their argument for a common standardised orthography, Olko & Sullivan (2014:389) have emphasised that all Nahuatl varieties share the same morphological behaviour, and this observation seems justified by the evident presence of NC in contemporary texts. Morphological analysis, therefore, is central to the exploration of the possible overlap of contemporary written samples and CN.

From a computational point of view, *morphological parsing*, the process of finding the constituent morphemes in a word (Jurafsky & Martin, 2009:113) seems an unavoidable first task in working with an agglutinative language. It is hardly relevant for languages like English, but of primary importance for Nahuatl, taking into account the importance of NCs. In addition, morphological parsing is the basis for other applications including tokenisation, spelling checking, non-trivial dictionary lookup, language teaching and comprehension assistance, and, in general, for expanding existing resources for minority languages (Beesley, 2004). Segmenting NCs into their morphemes is useful, for example, for determining lexical correspondences between Nahuatl and Spanish, since one

Nahuatl orthographical word can correspond to a group of orthographical words in Spanish (Gutierrez Vasques, 2015).

Variation is one evident challenge for performing morphological analysis of texts in the Nahuatl continuum. Texts not only reflect the lexical and structural variation between varieties, but are also written using divergent spellings. This practice hinders our ability to identify, parse and lemmatise what may be an otherwise identical form in two or more different texts. In addition to morphological parsing, therefore, the exploration of Nahuatl texts also requires us to tackle the heterogeneity caused by diverging orthographies fragmenting a growing and potentially valuable set of written data.

3.1 Available resources and morphological analysers for Nahuatl

3.1.1 Resources for contemporary varieties

The texts produced by diverse revitalisation efforts constitute a heterogeneous set, and heterogeneity is largely evident in orthographic terms. A concise review of contemporary orthographies is provided by De la Cruz Cruz (2014) and Pharao Hansen (2016). Divergence occurs, of course, also in lexical and structural terms, but in order to systematically and extensively identify converging features across such a continuum, orthographic variation is a first very significant hindrance.

Data sets from contemporary Nahuatl texts are available in the form of wordlists thanks to the web crawling carried out by the project *Crúbadán* (Scannell, 2014). The data for Nahuatl is divided into 17 groups, according to the variety name declared in the constituent documents. The lists of words and bigrams with corresponding frequencies given for each group might aid statistical approaches for developing applications, but they do not allow for the exploration of each word in context. More importantly, the usefulness of the data for statistical techniques is significantly reduced by the fact that the same word would be counted separately if it is spelled differently. The wordlist nhx (Ishtmus-Mecayapan variety), for example, includes the VNC *quicaqui* 'he hears it'; the same VNC appears as *kikaki* in the wordlist labelled nsu (Sierra Negra variety).

This means that the approximately 4 million word tokens (see **section 4.1** for a definition of *word tokens* and *word types*) collected by Crúbadán for the Nahuatl continuum are more likely to be used in subsets ranging from 15,284 (Morelos variety) to 502,418 word tokens (Guerrero); the respective count of word types would range from 2,526 (Morelos) to 15,050 (Guerrero). The project *Axolotl* (Gutierrez-Vasques et al., 2016) compiled a parallel Spanish-Nahuatl corpus mostly from non-digitised documents. The texts in the corpus can be consulted and downloaded in a web interface. Like the data sets compiled by crawling the web, Axolotl is synchronically, diachronically and orthographically heterogeneous due to its aim to be as comprehensive as possible. Consequently, the data present the same sparseness as the sets compiled by Crúbadán.

The work of Lastra de Suárez (1986) provides lexical comparisons between 26 modern vocabularies collected in an extensive dialect survey with a questionnaire of around 400 Spanish words. The data, however, is difficult to incorporate in modern systems, as it is only available in printed format. Vocabularies compiled by SIL are available for diverse varieties. Following the traditional practice of the SIL, the entries in these works tend to be rather alphabetical transcriptions, close to the phonological distinctiveness of the variety covered. Although there have been studies that compare the vocabulary across two varieties (e.g. Aburto & Mason, 2005), the aim to emphasise the distinctiveness of each variety in SIL vocabularies translates, again, in a dispersion of lexical resources potentially valuable for all the continuum. The analytical dictionary of Amith (2002:256) focuses on the variety of Ameyaltepc, Guerrero. This work envisages the potential of electronic formats to aid inter-dialectal lexicography. However, a key characteristic of Amith's work is his criticism of the focus of Western lexicography on written texts, and Amith's objective is giving indigenous languages a particular treatment as mostly spoken languages. His dictionary, therefore, largely consists of phonemic transcriptions of spoken data collected in field. Amith's strategy is sociolinguistically defendable, and methodologically necessary in the absence of written texts. Undeniably, the dictionaries of Amith and SIL are valuable additions to the knowledge of the Nahuatl continuum. They are noteworthy efforts to document the spoken varieties as they are now, but their direct contribution to the revitalisation of the continuum may prove to be limited. One could wonder, for example, how these resources could be used by Nahuatl readers outside the communities in which they are compiled, as they are closely focused on a thorough description of the local spoken language, and provide explanations for each entry in either English or Spanish. The present

work uses the term *reader* instead of *speaker* because, after all, a dictionary, electronic or not, is fundamentally a written resource, most likely to be read, without the aid of spoken interaction, when used outside the specific place of its creation. The dictionary of Chicontepec Veracruz (Sullivan et al., 2016), covers only one variety, but intends also to target users from other Nahuatl varieties, and not only locals and non-speakers. This work states as its aim to connect with the tradition established by previous lexicographers and with other important resources, as attested by the format and orthography chosen for the entries. Notably, the dictionary is, to the best of my knowledge, the first monolingual Nahuatl dictionary, aiming to be consulted mainly by Nahuatl speakers from across the continuum.

3.1.2 Resources for CN

The orthographical variation and, consequently, the dispersion of data for CN is somewhat alleviated thanks to scholarly work. CN was never written uniformly, but a review of orthographies for CN (Wright-Carr, 2007:45-56) reflects a converging tendency over the last years largely influenced by the work of Andrews. The conventions called ACK orthography by Olko & Sullivan (2014:386) are the culmination of decades of academic research and publication. Thanks to this academic work, there are nowadays some more or less established conventions, mainly among US scholars, regarding, e.g., the representation of NCs as one word, the treatment and representation of the glottal stop as a consonant, and vowel length. Most of these tend to be used in the publication of texts or extracts, as well as dictionaries. The enormous CN corpus is, nevertheless, far from being entirely available in a uniform orthography, as shown by the challenges faced by the project of normalisation of the *Codex of Florence* (Thouvenot, 2011). The normalisation is carried out with the aid of *Tecpana*, a procedural orthographical normaliser designed for CN that encodes about 270 rules to transform a given string.

There are numerous lexical sources for CN, apart from the essential dictionaries of Alonso de Molina (1571) and Remi Simeon (1885), as shown by the review of Wright-Carr (2007:25-41), and some incorporate comparisons with modern varieties. In the dictionary of Karttunen (1992), available in hard copy, colonial sources are complemented with vocabularies from modern varieties, namely Tetelcingo (Morelos), Xalitla (Guerrero) and Zacapoaxtla (Puebla). The dic-

tionary is also notable for denoting the vowel quantity and glottal stop. The Spanish-Nahuatl dictionary of De Wolf (2003) is an impressive work collected from core classical sources, and providing related contemporary variants as part of an entry. Unfortunately, to locate a Nahuatl word the Spanish translation must be known, and only a printed version is available.

Online dictionaries spare us the digitisation process necessary for integration with computer applications. Pharao Hansen (2014) provides a useful review of some of them. The *Grand Dictionnaire du Nahuatl* (GDN) is an ambitious, extensive source integrating in one interface Nahuatl lexical sources from different times and places. This project began as a desktop application developed by Thouvenot (2008b), and was published online by the National University of Mexico (UNAM) in 2012 as *Gran Diccionario del Nahuatl*. The GDN is part of a collection of resources called *CEN*[1], which includes a morphological analyser of CN, and two word-in-context explorers for the corpora included. CEN was also first available as a desktop application (Thouvenot, 2008a) and is also being implemented online by UNAM as the *Compendio Enciclopédico Nahuatl* (which conveniently has the acronym CEN). The format of the dictionaries in GDN is not easily portable for use outside of the dedicated interface. More importantly, in order to be integrated in other applications, the dictionaries in GDN have to be treated separately: the data in GDN, albeit extensive, becomes dispersed due to the decision to maintain the original spelling of the component sources. The *Copenhagen Nahuatl Dictionary* (CONDIP) (Canger, 2002) is another large-scale effort to integrate a dictionary with in-context explorations of a word in a corpus. It provides morphological analyses of the words in three different orthographies: one palaeographic, one normalised, and one phonemic. Although it is online, CONDIP was created in the late 1980s, and despite the potentially useful information provided, the obsolete text encoding and the database format hinders its integration in other applications. The *Wired Humanities Dictionary of the University of Oregon* (WHDO) (Wood, 2016) integrates core sources with Nahuatl definitions composed by Nahuatl students from the IDIEZ. The entries, however, do not present explicitly key information to classify nominal and verbal stems, e.g. transitivity (see **section 4.2.1** and **Table 4.3**). This information has to be located or inferred from the fields recovered from the component sources, which are provided as part of an entry.

[1]the name is an euphonic pronunciation of *ce* 'one', which is translated by the creators as 'together'

The dictionary by Wimmer (2006) is another comprehensive work of mainly colonial sources, notably the dictionaries of Molina and Simeon. Its entries are complemented and cross-referenced with other dictionaries like Karttunen (1992), and with other numerous sources, including early editions of the grammars by Launey (1979) and Andrews (1975). Entries contain examples of usage in colonial texts, references to the sources, and employ the normalised orthography ACK. Most importantly, it is more accessible and acquirable than other online dictionaries. It is part of the GDN, making up 17% of the GDN entries, but is also available on its own website. Being written in plain HTML, the dictionary is in a format that does not easily become obsolete, nor does it require special interfaces to be read, and the HTML tags can be readily removed.

3.1.3 Morphological analysers

There is only one morphological analyser for CN called *Chachalaca* (Thouvenot & De Pury, 2008). It is a procedural application (see below) whose desktop version is available as part of the CEN. At the time writing, an online version is reported to be in development[2]. Chachalaca performs its analysis using the Wimmer dictionary as one of its two lexical databases, and based on the grammatical summary available in the dictionary, which is based, in turn, on the grammars by Launey (1979) and Andrews (1975) and Sullivan (1992). Although the programme is certainly useful in the morphological analysis of a word, it presents certain pitfalls that make it difficult to integrate in larger batches of work. Apart from the considerable time taken to analyse a word (see **section 5.1** and **Table 5.1**), the programme is completely dependent on its dictionaries: if it is unable to find the stem of a word, the analysis fails with no further suggestions. In addition, the strings to be analysed have to be written in the normalised spelling used by Chachalaca, and deviations from this spelling causes a failure, as the documentation explains. The programme is able to propose some substitutions of certain characters to reattempt the analysis, but this adds to the already considerable analysis time.

To the best of my knowledge, there is no dedicated morphological analyser for contemporary written Nahuatl. Maxwell & Amith (2005) have documented

[2]The four applications of the project CEN are available online at http://cen.iib.unam.mx [last accessed on 16/08/2018]. Chachalaca is still the only application not accessible online yet.

the grammar of two contemporary varieties as a finite state grammar. Although one functionality of this research product is to perform morphological parsing, it has been largely built with pedagogical and documentation goals in mind. Moreover, as it is the case with most research on modern varieties, it overtly aims to register the spoken particularities of the documented varieties as faithfully as possible. As a result, the work needed to adapt this application to other purposes is equivalent to creating a new one.

Gutierrez Vasques (2015) approached morphological segmentation of NCs using *Morfessor* (Creutz & Lagus, 2005; Smit et al., 2014) a family of probabilistic machine learning methods for inducing morphological segmentations from unannotated data. Morfessor has proved useful for work with agglutinative languages like Finnish and Turkish. A relevant limitation for Morfessor in approaching Nahuatl is the aforementioned sparseness of available data. For version 1.0, for example, the best segmentations were achieved when learning from word types in contrast with the learning from word tokens; the size of the data set, likewise, increased the precision of the results (Creutz & Lagus, 2005:20). The data set used in demonstrations of the version 2.0 for agglutinative languages (Finnish and Turkish) is considerably larger than the sets provided by Crúbadán. It comprises, among others, 36 million tokens and 2,206,719 types for Finnish, and 12 million tokens and 617,298 types for Turkish (Smit et al., 2014). The amount of data available for Nahuatl, therefore, might affect the performance of Morfessor. Gutierrez Vasques (2015:156) recognised the dialectal, diachronic and orthographical variation in her corpus as a source of noise for statistical methods in general. She applied the set of orthographical normalisation rules from Thouvenot (2011), managing to reduce the variation in many but not all texts. In consequence, only a subset of documents (around 30%)[3] was usable for the rest of the project. In short, variation, particularly orthographical, adds noise to statistically induced morphological analysis, unless the set of training data is fairly homogeneous, either by normalisation or by using a homogeneous subset.

Summing up, this work aims to explore convergence in a morphologically complex language, using a pivot model (CN) and a set of contemporary data largely dispersed as result of heterogeneous spellings. It is necessary, therefore to tackle the questions of morphological analysis and orthographical divergence at the same time.

[3]Personal communication (April 2017)

3.2 Probabilistic versus non-probabilistic methods in NLP

Processing texts to explore convergence in written Nahuatl is a Natural Language Processing (NLP) task. Unlike other data processing applications, NLP entails not just manipulating bytes, lines and characters; NLP systems performing such processing also incorporate linguistic knowledge like phonology, morphology, syntax, etc., to achieve their purpose (Jurafsky & Martin, 2009:36-8).

There are philosophical considerations underlying computational approaches to language and NLP and, in general, to the question of how to capture knowledge of language in computational models. According to these considerations, approaches to the study of language could be broadly classified as rationalist and empiricist, and approaches to NLP, in a more less parallel, albeit non-strict correspondence, as probabilistic and non-probabilistic. This broad classification does not imply whatsoever that approaches to NLP do not complement to achieve practical goals, or that a non-probabilistic technique necessarily accepts rationalist assumptions or vice versa; quite often, the adoption of a certain technique in NLP is a practical necessity rather than a theoretical linguistic stance.

A concise comparison between empiricism and rationalism, as well as the implications for NLP is given by Manning & Schütze (1999:4-7). A rationalist approach to language assumes that the mind starts with a detailed set of principles and procedures specific to the various components of languages. Rationalism was largely predominant in linguistics, psychology, artificial intelligence, and NLP between 1960 and 1985. According to it, a significant part of the knowledge in the human mind is fixed in advance, not acquired through the senses. The *generative linguistics* of Chomsky, for example, assumes that there is an innate language faculty, and seeks to describe the corresponding module in the human mind. For this aim, data such as texts provide only indirect evidence, and the intuition of native speakers is favoured as source of explanations. Chomskyan linguistics depends on categorical principles which are or are not satisfied: a sentence, for example, is grammatical or ungrammatical (Manning & Schütze, 1999:5). A rationalist approach to NLP might consist of an attempt to supply systems with a hand-coded starting knowledge about a language; linguistic knowledge would be thus 'given' to a system, e.g. by coding a set of rules. Such

non-probabilistic techniques do not necessarily depend on statistics (and, for extension, probabilities), such as the frequency of patterns found in data sets. One advantage of non-probabilistic techniques is that they allow for the encoding of facts (or assumed facts) already known/assumed about the language, thus alleviating the lack of large training data sets required by probabilistic methods. A drawback is that the encoding of comprehensive rules can be complicated and time-consuming, and might prove incorrect if our assumed, observed 'rules' are not borne out in usage.

Empiricism assumes that some cognitive ability and initial structures are present in the brain. They are not, however, detailed principles and procedures specific to language, but rather only those that allows for general operations for association, pattern recognition and generalisation. These general operations make possible the learning of the detailed structure of natural languages when they are applied to the rich sensory linguistic input available to a child. Language principles are learned from experiencing the linguistic surroundings. Empiricist approaches are interested in describing language as it actually occurs, not as coded in a linguistic mental module. In consequence, data provide by, e.g. textual corpora, are the base of empiricist attempts to understand language. In an empiricist approach to NLP, linguistic knowledge is induced from training data by probabilistic means. In consequence, *probabilistic techniques* need data to be trained on, based on the statistics and probabilities determined for the training dataset. The need of training sets (Jurafsky & Martin, 2009:125) is a potential pitfall for probabilistic techniques if not enough data is available to attest as many repetitions of as many features of the language as possible. One advantage is that the linguistic knowledge encoded in a probabilistic model will be induced from real examples of usage, and not for assumed facts about a language.

Statistical/Probabilistic models are nowadays favoured in computational linguistics and NLP for two main reasons. First, statistical NLP is relevant to theoretical linguistics, because it may be able to account for things such as non-categorical phenomena and language change better than the criteria of grammaticality and ungrammaticality of generativist approaches. Cognitive processes, it seems, are best formalised as probabilistic processes, or at least by means of some quantitative framework that can handle uncertainty and incomplete information (Manning & Schütze, 1999:11, 15). Second, more technical reasons, like the availability of better computational resources and large data sets, also

favoured the development of empirical and statistical methods as the preferred approach to language processing (Jurafsky & Martin, 2009:46-7). The availability of textual data required by probabilistic approaches, however, is a luxury that not many minoritised and indigenous languages can afford: as shown by the example of Gutierrez Vasques (2015:156), a parallel corpus of 1 million tokens is still very small for Statitstical Machine Translation (SMT). Moreover, although availability of data is slowly increasing thanks to revitalisation efforts, statistical techniques might need to be restricted to small subsets in order to avoid potential noise caused by heterogeneous spellings and other sources of variation.

The aim of our research is one first reason to adopt a non-probabilistic technique to perform morphological analysis on our data. Our purpose is to use a given description (CN), available in multiple grammars and vocabularies, to identify points of convergence with contemporary written samples. A non-probabilistic model can encode such description of CN, which will be the theoretical pivot to be confronted against the empirical data provided by the contemporary texts.

The available resources are another reason to use a non-probabilistic technique. Let us take into account that, although certainly less dispersed than the contemporary corpus, CN texts are still undergoing a process of digitisation and normalisation. However, there is available one on-line comprehensive dictionary of CN (Wimmer, 2006), compiled in a uniform orthography, and thorough descriptions of its morphology (Andrews, 2003; Launey, 2011), and this knowledge can be implemented with non-statistical techniques. The encoding of CN vocabulary and grammatical descriptions in a non-probabilistic model would therefore avoid the need to collect the vastly larger amounts of data needed to generate a probabilistic CN model.

Our choice of a non-probabilistic NLP technique, which might seem closer to a generativist approach, is therefore not a theoretical stance, but is instead guided by practical motivations. This decision does not imply a rationalist view on language, but rather a strategy to best use the resources found in our current circumstances. An empiricist point of view and a probabilistic model could better approach the variability and competing forms in the Nahuatl continuum, not deeming them as wrong/unacceptable but rather as infrequent. Considerations of frequency of use are essential to understand language when categorical assumptions fail to explain non-categorical phenomena like language change and

sociolinguistic variation (Manning & Schütze, 1999:11). However, it is crucial to bear in mind that in order to employ probabilistic models using textual data collected from all the Nahuatl continuum, one still has to tackle the pressing question of the orthographical heterogeneity that fragments the available data into small subsets.

3.3 Finite States Morphology

The Finite State (FS) formalism is the base of non-probabilistic techniques that have proved useful for morphological parsing (Jurafsky & Martin, 2009:60-75), a central task for this research. The FS approach to morphology departs from a rather rationalist point of view, in which morphology can be largely explained in terms of rules. According to this view, the two central problems in morphology would be *word formation*, also called *morphotactics* or *morphosyntax*, and *phonological* and *orthographical alternation*. Morphotactics dictates the legal combinations and orders of the morphemes that make up a word; e.g. *un-guard-ed-ly* is a valid word in English, but **un-elephant-ly-ed* is not. The alternations determine the shape (either as a sound or as a spelling) of a morpheme depending on its environment; e.g. *fly* is realised as *flie* when followed by an *s*, and *die* is realised as *dy* when followed by *ing*.

Koskenniemi (1986, 1983) was the first to realise that morphological descriptions and the phonological alternation that determines the shape of a morpheme in a given context can be implemented in the form of FS automata. The work of Kokenniemi is the basis of FS morphology. FS morphology main claim is that morphotactics and alternations can be modelled and computed using FS networks (Beesley & Karttunen, 2003:xvi).

3.3.1 Finite State Networks

From a computational point of view a language can be first approached as a set of strings (words) composed of symbols from an alphabet (Beesley & Karttunen, 2003:259). In the FS paradigm each word of a language is represented as a path in a network consisting of states and arcs. Each arc represents the transition from one state to the other and is labelled with the symbol that causes the transition.

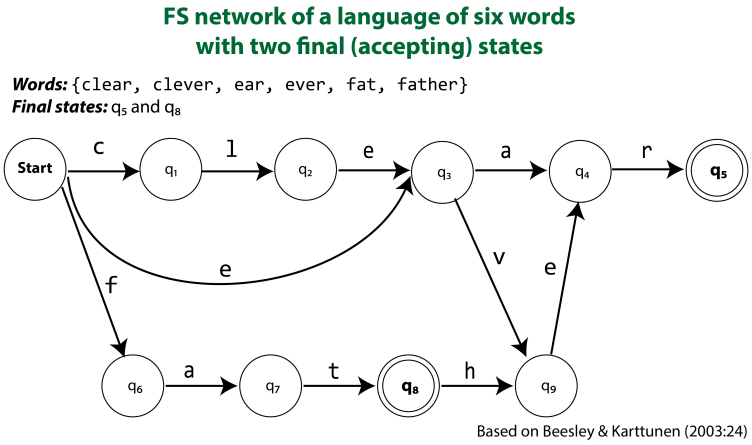


Figure 3.1 FS network of a language of six words.

The network has a START STATE and one or more FINAL or ACCEPTING STATES. The number of states in a path is finite, hence the name *Finite State*. In order to be accepted as a word pertaining to the language (network), each of the symbols of a candidate string must match a corresponding symbol in a valid path to cause a transition to the next state; if the candidate string activates all transitions through a final state, it is accepted as belonging to the language (Beesley & Karttunen, 2003:2, 17, 44). **Figure 3.1** shows an example of a simple FS network of a language of six words.

A FS network is a very efficient representation of a language because the language in question is denoted by a regular expression. A *regular expression* (regex) is a formula that specifies simple classes of strings. Formally, it is an algebraic notation for characterizing a set of strings (Jurafsky & Martin, 2009:52). Any regex has a FS implementation, just like any FS implementation can be described with a regex (Jurafsky & Martin, 2009:60). A regex denotes a *language* (i.e. a set of strings like {a, b, c}) or a *relation* (i.e. a set of ordered string-pairs like {<a, A>, <b, B>, <c, C>}) that can be compiled into a FS network that compactly encodes the corresponding language, which may well be infinite (Beesley & Karttunen, 2003:44). An example of a regex is $a+$, 'the language of words consisting of one or more concatenations of the symbol a'. This regex denotes, therefore, the infinite set {a, aa, aaa, ...}. Such a regex can be compiled, i.e. translated to machine language, into a FS network that compactly encodes the infinite language (**Figure 3.2**).

All the paths in a network share the same structure of arcs and states as can be seen in **Figure 3.1**. Consequently, the list of all possible words can be

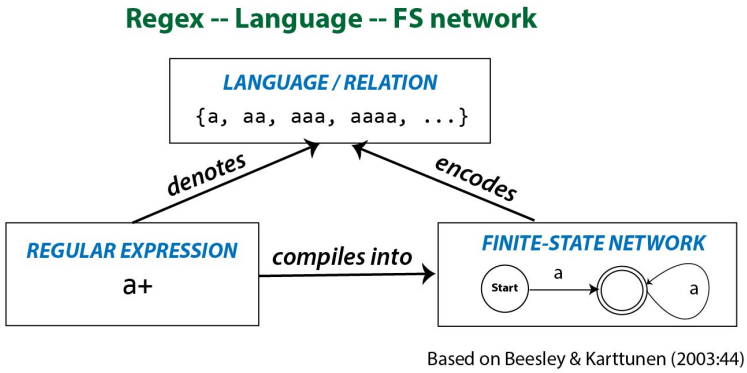


Figure 3.2 Regex, language and FS network.

'packed' efficiently in a FS network which is *minimal* in the sense that it is impossible to encode the same paths using fewer states and arcs. *Minimality* is an important property of FS networks for practical applications; long lists of words in text format that can take several megabytes in disk space can be encoded in a FS network using a few hundred kilobytes (Beesley & Karttunen, 2003:17). The infinite language denoted by $a+$ exemplifies the potential efficiency of FS networks in terms of data storage.

3.3.2 FS transducers and its advantages for this research

Previous work with Basque (Aduriz et al., 1997; Alegria et al., 2002, 1996), Finnish (Linden et al., 2009; Pirinen & Linden, 2010) and Arabic (Shaalan & Attia, 2012) have shown the utility of FS approaches to develop NLP applications for morphologically complex languages. Examples include also low-resourced languages like Quechua (Rios, 2011; Rios & Castro Mamani, 2014), one Nahuatl variety (Maxwell & Amith, 2005), Igbo (Iheanetu & Adeyeye, 2013), and the Odawa variety of Ojibwe (Bowers et al., 2017). These works have used FS techniques aiming to create a basis for applications that can support the revitalisation of these languages. FS techniques have been applied to document linguistic varieties, normalise texts, create spell checkers, and, in general, to facilitate the production of resources like grammatical descriptions and written texts in contexts of null or emerging literacy and divergent usage. The pedagogical application of FS applications has also been noticed (Bowers et al., 2017; Maxwell & Amith, 2005), as well as the potential of these tools to speed the creation and proof-reading of high quality texts and, by extension, to help enforcing a particular written form as standard if desired (Bowers et al., 2017).

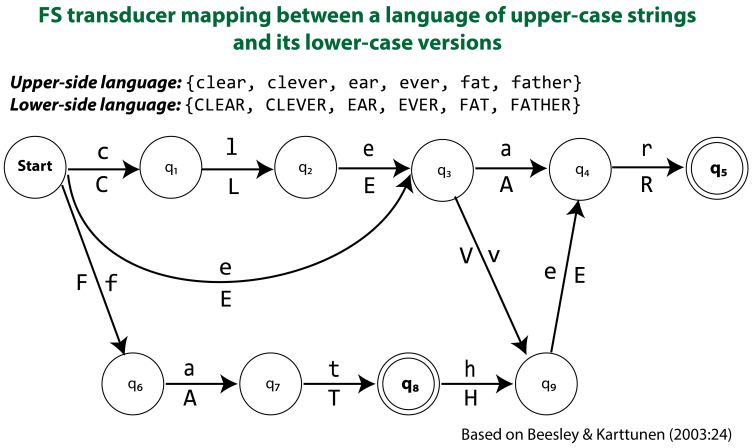


Figure 3.3 A transducer mapping a language of upper-case strings to a language of lower-case strings.

The language resulting from legal combinations of morphemes can be seen as a set of strings, and thus encoded as a FS network. The strength of the FS approach is that networks can indeed do more than just accept or reject strings: they can encode compactly more information and map between two corresponding strings, which is a key advantage for our research. A FS network can encode a morphological analysis of a word, and at the same time map this same analysis to alternative spellings.

As explained above, an FS network can encode also a relation, i.e. a set of ordered pairs of strings. The arcs in the paths encoding a relation have two labels, thus causing the network to have an *upper language* and a *lower language* (see section 3.3.2.1 below). These two-level networks are called *transducers* in reference to devices that convert energy from one form to another, e.g. a microphone which converts vibrations of air into electrical signals (Beesley & Karttunen, 2003:13). Likewise, a *FS transducer* (FST) relates one string to another by mapping between their comprising symbols. A transducer can, for example, map a lower-case string into its correspondent upper-case version, like the one shown in **Figure 3.3**. The upper language of this transducer is {clear, clever, ear, ever, fat, father} whereas the lower language is {CLEAR, CLEVER, EAR, EVER, FAT, FATHER}. A similar transducer could map between orthographical alternations to allow relaxed spelling, for example, the omission of accents in Spanish, or the use of <e>, <oe>, <ae>for <ü>, <ö>, <ä>in German.

3.3.2.1 FST compactly encodes diverse information of a language

Morphological information can be compactly encoded in a FST as a relation, and be used in morphological analysis. FS morphological parsing represents a word as a correspondence of a lexical [upper] level, which represents a concatenation of morphological tags making up a word, and a surface [lower] level, which represents the concatenation of letters making up the spelling of the word (Jurafsky & Martin, 2009:94). Both morphological analysis and generation can be achieved with one FST encoding the relation between the upper and lower languages. According to this perspective, mapping from the surface level to the lexical level is called *look-up* or analysis, and the inverse mapping is called *look-down* or generation (Beesley & Karttunen, 2003:13-14). In conventional FS notation the colon (**:**) is used to express a relation between the upper and lower language. The symbol to the left belongs to the upper language and the symbol to the right to the lower language. A *lexicon transducer* is a transducer encoding a morphological analysis in its upper language and an abstract, not-yet-orthographical string in its lower side. This initial network can be composed with a transducer of alternation rules into a single network to create a *lexical transducer*, a transducer incorporating all the morphological information about the language and about the phonological/orthographical realisation of the morphemes making up a valid word. The lexical transducer is thus a morphological analyser/generator and can be used to relate, either by analysis or generation, a morphological parsing and an orthographical word (Beesley & Karttunen, 2003:xv-i).

The composition of two FSTs helps to map a morphological analysis to its spelling realisation. In terms of relations, *composition* is a finite state operation that brings together the 'outside' components of two relations into a new relation, eliminating the common one in the middle (Beesley & Karttunen, 2003:28). For example, if A is the relation {<cat, gato>} and B the relation {<gato, Katze>}, the composition of A and B, expressed in FS notation as $A.o.B$, is the new relation {<cat, Katze>}. It is important to notice that composition is not a commutative operation, and that the composition of two relations with no intermediate linking element is the *empty language*, the language that contains no strings at all (Beesley & Karttunen, 2003:18). Following our example, the composition of {<gato>, <Katze>} and {<cat>, <gato>}, i.e. $B.o.A$, is the empty set {}.

In terms of transducers, composition links the upper language of one transducer with the surface language of another, eliminating the intermediate languages, i.e. the lower language of the transducer 'above' and the upper language of the transducer 'below'. For morphological analysis, composition maps an orthographical word to its corresponding analysis and vice versa. On one hand, the lower language of a lexicon transducer can include abstract symbols representing affixes whose orthographical realisation is conditioned by its context of occurrence. A given lexicon transducer *MORPH*, for example, could map the symbol ˆC (representing the third person singular direct-object prefix) in its lower language to the tag **DirObj3Sg**+ in its upper language. On the other hand, the alternation rules to determine the realisation of each morpheme, and the context in which the rule applies can be expressed as *replace rules* (Beesley & Karttunen, 2003:132-169). Being an extended notation for regular expressions, replace rules can be implemented as FS transducers. One such transducer, *ORT*, can encode the relation mapping ˆC, in its upper language, to its orthographical realisation, either <c>or <qu>depending on the context, in its lower language. The transducer resulting from the composition *MORPH .o. ORT* will map the tag **DirObj3Sg**+ to the appropriate spelling <c>or <qu>eliminating the intermediate symbol ˆC.

Composition is a very useful operation as successive FSTs can be composed to map between alternative spellings. In this way, the morphological analysis in the upper language of the top transducer can be efficiently linked to whichever spelling is specified by the lower language of the bottom transducer. The resulting FST will encode not only morphological analysis, but also information about the expected alternative spellings of the same word. The advantage of approaching morphological analysis in a context of variation is that one initial core network can be modified with FS operations (e.g. union, to extend the network, and composition, to map between different levels of successive orthographical alternations) to create different systems that support, for example, multiple dialects of the same language, multiple orthographies and multiple registers with different levels of strictness (Beesley & Karttunen, 2003:287-310).

3.3.2.2 FS versus procedural applications

FS applications for morphological analysis are faster and more extensible than procedural matching algorithms, i.e. analysers that work based on a strategy

of breaking down input words into constituent parts by consulting lexicons and rules designed for a particular language (Hulden & Samih, 2012). Whereas procedural analysers have to perform diverse tasks over a string, including searching potentially long lexicons in text format, a FS application only attempts to map a candidate string to a path in the network which is already compactly compiled. This being a single task, it translates into a speedy analysis.

The extensibility made possible by FS operations like composition is a key advantage in tackling the problem of orthographic variation. A procedural analyser might deal with orthographic variations as an extra replacement task, using the regex module available for the particular programming language in which the application is coded. In contrast, in a FS analyser, the alternative spellings can be compiled as part of the language network, without requiring any extra lookup time (Hulden & Samih, 2012). Some FS tools even allow for the alternative spellings to be composed on runtime to map variations without modifying the original network (Beesley & Karttunen, 2003:432-3). This capability has proved useful, for example, in the normalisation of Quechua texts (Rios & Castro Mamani, 2014). Notably, the mapping of alternative spellings does not consist only of a replacement of characters with no incorporated knowledge about the morphology or phonology as happens in, for example, the CN normaliser *Tecpana* (Thouvenot, 2011:163). In contrast, a FS application can incorporate in the same network knowledge about morphology, phonology and the orthographical variation one might expect, thus integrating in one the capabilities of two procedural applications, e.g. the CN morphological analyser *Chachalaca* (Thouvenot & De Pury, 2008) and the normaliser *Tecpana*.

Another important advantage of a FS approach is the possibility of creating a *guesser* (Beesley & Karttunen, 2003:444-451). Assuming the morphotactics and morphophonology have been correctly described, a normal FS analyser will not recognise a word unless its stem is part of the network. A guesser is a FST designed to analyse a word based on any phonologically possible stem, which can be described in terms of a regex. This is useful to identify newly borrowed or coined words (Shaalan & Attia, 2012), and in general to split a word into morphemes around a hypothetical stem when the main transducer fails (Rios & Castro Mamani, 2014). Thus, a guesser allows for the exploration of plausible morphological convergence with a sample text even if there is a lexical gap in the FS network.

Summing up, morphological analysis is necessary to explore convergence between written Nahuatl texts which are heterogeneous in dialectal and orthographical terms. For such task, the non-statistical approach of FS is a good chance to circumvent the scarcity of uniform training data, while addressing the dialectal and orthographical variation in the analysed texts. FSTs can encode our knowledge about CN and can be adapted to the divergent orthographies found in our text sample. Moreover, a core FST can be used to create a guesser that points to plausible morphological convergence despite divergent/unknown vocabulary in a text.

3.3.3 Applications to create, manipulate and search FS networks

FSTs are computational representations relating two languages. FS networks for a complex language are built by describing such a language in terms of set operations and regular expressions, i.e. a notation that denotes a set of strings (language). These regular expressions can then be compiled into a FS network using dedicated applications. Particularly notable are the Xerox Finite State Tools (XFST) (Beesley & Karttunen, 2003), a collection of interfaces, compilers, run time applications and libraries of FS algorithms to build and test linguistic FS applications including morphological analysers/generators. Much work on FS applications and particularly on FS morphology has been carried in the Xerox Research Centre Europe and in the Palo Alto Research Centre; XFST has been the culmination of much pioneering work on FS linguistic theory first formulated in the early 1970s and recovered and developed over the next two decades (Beesley & Karttunen, 2003:xi-ii).

XFST include two development tools, *xfst* and *lexc*. The xfst utility provides an interface to the basic algorithms of FS calculus, and a compiler for a metalanguage of regular expressions. It permits the creation of the corresponding FS networks, and their manipulation with FS operations like concatenation, union, intersection, composition, etc. FS networks can be completely defined with regular expressions compiled by *xfst*. However, the syntax of regular expressions (e.g., required spaces between characters, or curly braces around whole words) means that adding words to the network is rather tedious, and that the resulting expression can be extremely complex to read. *lexc*, for Lexicon Compiler, is a high-level declarative language and associated compiler for defining FS networks

and transducers. A *lexc* source file describes the morphotactics of a language in terms of *lexicons* and *continuation classes*. Each entry in a lexicon consists of two parts, a stem or affix and a continuation class, i.e. the lexicon containing the forms to be concatenated with a given entry in order to create a path (word) in the FS network. Every *lexc* description has a lexicon called Root which corresponds to the Start state of the network to be compiled. In contrast, the (possibly various) final or accepting states can be distributed in different lexicons: when an entry in a lexicon is a valid end of word, it has # as continuation class. The hash # is a special symbol indicating the end of a word, i.e. an accepting state of the FS network. The source files of *lexc* descriptions can be produced with a plain text editor, e.g. Emacs, Vim or Notepad, and compiled into a FS network with *xfst*. FS networks compiled from a *lexc* description can be further manipulated using *xfst*, for example, to compose a core transducer with others encoding alternative spellings, or to create a guesser.

The FS networks created with *xfst* and *lexc* are used with a run time application called *lookup* which applies a FS network to input tokens, e.g. to perform morphological analysis/generation. It can be used in command-line, and its output piped to other applications as a component of a larger system. A key feature of the XFST is the use of *flag diacritics*. These are multicharacter symbols that can be strategically inserted in the strings of the network, allowing for overcoming the limitations of simple unrestricted concatenations. Flag diacritics manage separated dependencies, i.e. constraints on the co-occurrence of non-contiguous morphemes within words. A suitable combination of flag diacritics can mark a path (word) as illegal. Such paths are blocked during the analysis and generation routines performed with *lookup*.

The notation and syntax used by the tools designed in Xerox have been extended to other sets of tools, notably the Helsinki Finite State Tools (*hfst*) (Linden et al., 2009) and FOMA (Hulden, 2009). The *hfst* tools are particularly noteworthy because it is possible to incorporate *weights* into the compiled FS networks. Weights tell how probable a word or its analysis is, and they can be thought as penalties to mark words/analyses with bigger weights as less probable. Therefore, they can facilitate the disambiguation among several possible analyses of a given word, although not all operations or rules support weights (Axelson, 2013). FOMA is a compiler and C library containing implementations of all classical automata/transducer algorithms: determinisation, minimisation, epsilon removal, composition and Boolean operations. It is compatible with the

XFST regex and scripting syntax (*xfst* and *lexc*), supports flag diacritics and its own lookup utility for applying FS transducers, called *flookup*.

FOMA was chosen as the development tool for this research for practical reasons. Unlike *hfst* and XFST, all the tools available in FOMA are available under a GNU license. The potential of using weights with *hsft* makes it a great option to discard unlikely analyses; however, successive work on annotated corpora is still necessary to determine weights. In addition, FOMA proved more easily installable than *hsft* in the operative system used by this research (Ubuntu 16.04). The last version of XFST was unable to handle flag diacritics operations that FOMA executed with no further problem. Finally, the same *lexc* and *xfst* source files used with FOMA can eventually be compiled in either *hfst* and XFST.

The next chapter will describe the methodology followed in this research to explore the overlaps of eight contemporary Nahuatl texts. The description of the test corpus is followed by an explanation of how the comprising modules of a CN FS model were created; and of how different alternation rules are expected to map alternative graphemes used in the test text to the canonical orthography defined in the CN model. Finally, it will be explained what the different outputs of the FS model for a given word type indicate regarding the points of convergence between a test text and CN.

Chapter 4

Methodology

The goal of this methodology is to locate points of convergence between samples of contemporary texts and CN. For this purpose, a Finite State model of Classical Nahuatl (CN) was tested against eight modern texts from different varieties to find an overlap between them, indicated by a plausible morphological similarity between the NCs used in them.

4.1 The test corpus

The test texts (**Table 4.1**) are eight versions of the New Testament (NT) published by *La Liga Bíblica Internacional*, a publishing partner of the Summer Institute of Linguistics (SIL). The versions of the NT are presented as pertaining to varieties of Guerrero, Michoacan, Central Huasteca, West Huasteca, East Huasteca, Puebla Highlands, North Oaxaca and North Puebla. The name used in this research is the English approximate translation of the Spanish geographical name given in the text to distinguish it. The present work did not aim to obtain a proportional sample according to any classification of major dialectal areas or varieties.

This work does not seek to achieve the ideal balance and representativeness of a snapshot corpus (McEnery & Hardie, 2012:9) for two main reasons. First, as has been mentioned above, different sources recognise different numbers of varieties, and there is no clear delimitation between them in many cases. That is why the idea underlying this methodology is to approach Nahuatl as

a continuum, rather than as a set of discrete entities with well-defined, closed boundaries. From this point of view, therefore, variety labels can be taken into account as metadata, but not as absolute categories to be sampled proportionally. Second, many more published texts are available in paper, but digitizing them is beyond the possibilities of this work; likewise, it might not be possible to obtain published texts from varieties with very few speakers. In brief, the lack of clear categories and the scarcity of immediately accessible data make it difficult to achieve balance and representativeness in terms of the regional varieties covered. Our corpus could be described as an opportunistic corpus (McEnery & Hardie, 2012:11-3), and the conclusions derived from its exploration are intended to illustrate how our methodology could be applied to explore convergence and written language in further work.

A small *gold standard* (GS) of CN (**Table 4.1**), i.e. texts for which morphological analysis were human-generated, was used to evaluate the performance of the CN model against CN texts alone. It consists of translations of two of Aesop's fables found alongside other texts in the manuscript MS 1628 owned by the National Library of Mexico. The GS is based on one palaeography available on-line of the texts(Ayac, 2014), which includes a morphological analysis of each word. According to the author of the palaeography, it was done after a facsimile of the original manuscript. The granularity of the morphological segmentation, the tags used and some spellings are not the same used by the model, but the correspondence was determined carefully to keep the morphological analysis faithful to the original.

The texts in the test corpus are approximately between 168,000 and 275,000 *word tokens*[1] in length, excluding footnotes, legal pages and appendixes. In terms of *word types* the texts contain between 12,930 and 21,819 word types. Word types are distinct word tokens (Manning & Schütze, 1999:21); one type can occur many times in a text or corpus, but all the tokens consisting exactly of the same characters are considered occurrences of the same type (McEnery & Hardie, 2012:252). A list of word types was generated for each text using dedicated scripts written in Python. Each wordlist sorts alphabetically the word types found in a text with its corresponding frequency. The word types

[1]Tokens are the individual units, including numbers and punctuation marks, in which the text is divided. Although tokenization entails in practice more complex considerations, *word token* means here "the strings of alphabetic characters delimited by whitespace" (Manning & Schütze, 1999:124-125).

Versions of the NT (Contemporary Varieties)			*Gold Standard (Classical Nahuatl)*		
Text	*Word Tokens*	*Word Types*	*Aesop's Text (CN)*	*Word Tokens*	*Word Types*
Guerrero (G)	185,499	15,608	The Old Man and Death	135	85
Central Huasteca (HuC)	199,793	13,182	The Fishermen	55	41
West Huasteca (HuW)	198,542	13,452			
East Huasteca (HuE)	198,541	13,656			
North Oaxaca (OaxN)	190,441	12,930			
North Puebla (PuebN)	168,364	15,265			
Michoacan (Mich)	254,756	14,028			
Puebla Highlands (PuebH)	275,368	21,819			

Table 4.1: The test corpus and the CN gold standard. The names of the contemporary varieties used here to distinguish each text are those under which the text is presented. The number of dialectal areas vary from source to source and thus the names in this table do not necessarily imply a correspondence to a dialectal area.

collected in the wordlists are only alphabetic strings obtained after filtering out, for example, chapter and verse numbers.

The texts were chosen because their electronic versions were available online. This means that they are easier to obtain and to convert into plain text (and with fewer errors to correct) than texts scanned from paper sources. In addition, they are of a reasonable extent, so one can expect to find more word types in them than in shorter texts. In general, it seems that the number of word types grows in proportion to the number of tokens (Jurafsky & Martin, 2009:120). Finally, they have undergone a revision to be published online, which implies that they already reflect an effort to achieve a relative uniformity in the orthography and language used, unlike, e.g., manuscripts used for personal communication.

The texts have also been chosen following the idea that linguistic revitalisation entails mainly the transformation of a language to fulfil new roles, rather than a restoration of what is taken to be its authentic version. In general, the texts in the corpus are relevant as good examples of endeavours to take the local language into new roles. They are not transcriptions of spoken interchanges or local stories, but introspective, more mediated attempts to use the everyday local language to communicate messages which are far from everyday local life. In particular, the Biblical texts are interesting because the publishers, partners

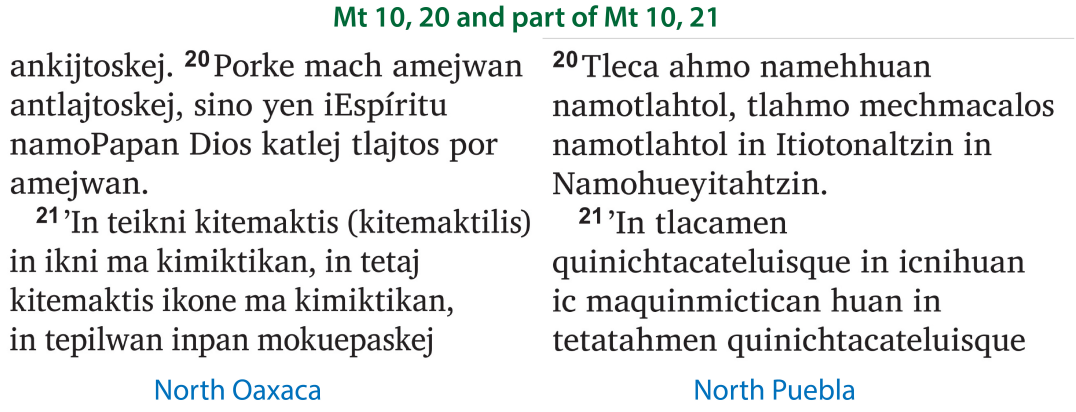


Figure 4.1 Mt 10, 20-21 as rendered in the North Oaxaca and North Puebla texts.

of SIL, seem to share with SIL the tendency to make the language used as close as possible to local usage.

The texts present divergent lexical and grammatical features, but the most immediately evident differences tend to be orthographical. The verses in **Figure 4.1** correspond, for example, to MT 10, 20-21 in the version for North Oaxaca (OaxN) and North Puebla (PuebN). The texts show an orthographical alternation to represent the phonemes /k/ and /w/, namely <qu> and <hu> in PuebN, and <k> and <w> in OaxN. Besides, the texts show a different second person plural pronoun 'you (Pl)' (*amejwan* in OaxN, *namehuan*, in PuebN), and differ also in the lexical item used to introduce an explanation (the Spanish conjunction *porque* 'because' in OaxN, and tleca in PuebN). Stylistic variations, e.g. the use of different constructions to convey equivalent ideas, also contribute to increase the sensation of difference between both texts. For example, the clause 'for it will not be you speaking' at the beginning of verse 20[2] has as equivalents:

(1) *Porke mach amejwan antlajtoske*

'because it is doubtful you (Pl) will speak' (OaxN)

(2) *Tleca ahmo namehhuan namotlahtol*

'because it is not you [nor] your words' (PuebN)

[2]The verse 20 reads in English "for it will not be you speaking, but the Spirit of your Father speaking through you". Verse 21 reads "Brother will betray brother to death, and a father his child; children will rebel against their parents and have them put to death" (New International Version Bible available on http://www.biblestudytools.com/matthew/10.html [Accessed 8/8/2017])

Whereas the OaxN text relies on *antlajtoskej*, a future form of the verb *tlahtoa* 'to speak', the PuebN text uses *namotlahtol*, a possessive form of the noun *tlahtolli* 'words'. It is to be noticed also that the possessive form *namotlajtol* is also found in the OaxN text, as well as the negative *ahmo*. The point to highlight is that below their evident differences, there seem to be overlaps between both texts, and they should be found despite the orthographical and stylistic variation.

4.2 The implementation of the transducers of the model

The FS model of CN is implemented as a group of FS transducers, i.e. FS networks with two levels. The goal, in brief, is to create with these transducers a 'black box' that accomplishes a string-to-string mapping of word-types found in the texts to CN morphological analyses.

The implementation of each of the component transducers involves, broadly speaking, two tasks: one, the creation of a transducer whose upper 'words' consist of tags of morphemes concatenated according to CN morphotactics, and whose lower 'words' are an intermediate, abstract representation of a CN word; and two, the composition of this transducer with others encoding rules to map the intermediate words to an orthographical realisation. The resulting transducer, thus, relates the surface orthographic forms in its lower language to a corresponding morphological analysis represented by the tags in its lexical level. A first composition of orthographical alternations maps to orthographical forms following the conventions of our 'canonical' orthography. Successive compositions can be used to map between these spellings and the variations found in contemporary texts. The strength of a FS model is that such mappings are not trivial alternations of symbols, but mappings that consider at the same time a plausible common underlying morphology of the related strings.

The model is mostly concerned with the morphosyntactic constructs that Andrews (2003:45-9) calls *nuclear clauses* (NC), particularly with *verbal nuclear clauses* (VNC) and *nominal nuclear clauses* (NNC). The internal morphological structure of these constructs is modelled as a FST using *lexc* scripts compiled using FOMA (Hulden, 2009). In general, FOMA follows the syntax used by the Xerox FS Tools (XFSTs) (Beesley & Karttunen, 2003), so the *lexc* scripts

compiled in FOMA can also be compiled in XFSTs. Other Nahuatl lexical items called *particles* (Andrews, 2003:39-44) are largely invariant, i.e. have no internal morphological structure, and are compiled in a separate transducer.

4.2.1 Grammatical and lexical sources

To model the core CN network this research relies mainly on the grammatical descriptions of Launey (2011) and the dictionary compiled by Wimmer (2006).

The plausible morphological structures of NCs were defined according to Launey's grammar. As explained by Andrews (2003:46), a Nahuatl nuclear clause consists of affixes arranged around a stem in a rigid sequence, so its structure can be expressed in a formula. To deal with such formulae, two things must be known: what the positions (slots) represent in terms of informational categories (e.g., person, number and tense), and what fillers can occupy the positions. This description translates well into the FS-morphology strategy (Beesley & Karttunen, 2003:222) of defining plausible morphological structures in order to reduce further work to lexicography; in other words, one can define first a suitable combination of affixes and stems, and then add new stems into the appropriate category/slot. Launey's descriptions of VNC and NNC structures were modelled by means of *lexicons* (lists of suitable entries to fill a slot in the structure) and *continuation classes* (which indicate the next lexicon with morphemes to concatenate) declared in a *lexc* file. In such arrangements, appropriate generic slots are reserved to be filled with sets of stems that share the same morphological behaviour. The ordering of constituents for NNCs, VNCs and the inventory of affixes used in the model are those shown in **Tables 2.4** to **2.9**. As the table **Table 4.2** sums up, the model generates/analyses four sets of tense forms, four sets of modality forms, and two sets of directional conjugations, which according to Launey (2011:227) are the basic sets of forms of VNCs in CN. As for NNCs, the model generates/analyses sets of possessed and unpossessed forms of noun stems, which can also express size or attitude. NNCs include as well RNNC generated/analysed using 48 relational stems.

The Wimmer dictionary provided the nominal and verbal stems to fill suitable morphological structures. The stems to populate the morphosyntactical structures modelled following Launey's grammar were obtained from the dictionary. Following the simile of NCs being formulae, categorising stems is most

Forms of VNCs modelled	Forms of NNCs modelled		
Tense	Size/Attitude	State	Relation
Preterite			
Future	Honorific/Diminutive(-tzin)		
Imperfect			
Pluperfect	Diminutive (-ton)	Possessed	
Modality			
Counterfactual	Deprecatory (-zol)	Unpossessed ('Absolutive')	RNNCs formed on 48 relational stems
NI form (habitual behaviour)			
Optative	Diminutive with pity (-pil)		
Vetitive			
Directional Conjugation	Augmentative (-pol)		
Motion Toward			
Motion Away			

Table 4.2: Forms generated/analysed by the FS model.

relevant for the FS approach, since the type of stem determines the affixes it can take, and therefore the composition of an acceptable NC.

Four main categories were defined for nouns based on their *absolutive suffix*, namely <tl>, <(t)li>, <in> or 0. *Absolutive* is the simplest state of a NNC, when "no possessor pronoun [prefix] occurs in its predicate and the subject pronoun [prefix] shows singular/common number [i.e. the third person singular 0]" (Andrews, 2003:101). In simpler terms, the absolutive is the non-possessive form of a NNC with no visible subject prefix, e.g. *tah-tli*, '(he is) father' as opposed to *ti-no-tah-tzin*, 'you are my beloved father'. Being the simplest, the absolutive is the dictionary form of a noun, from which the nominal stem, e.g. *tah-*, can be found. The main four categories of verbal stems are based on whether or not the preterite tense is formed by dropping the final vowel of the dictionary form (the third person singular of the present). **Table 4.3** shows the eight main categories of stems used in the model.

Other relevant subcategories are based on other considerations. Nominal stems, for example, were further classified as animate and inanimate, as in CN this distinction determines whether a noun can be pluralised. Besides, if a plural can be formed, the ending of the stem (vowel or consonant) seems to determine the plural suffix to be taken (Launey, 2011:19-21). Verbal stems, among other criteria, are further classified depending on whether they take an object prefix (transitive) or not (intransitive), and on the last vowel of the dictionary form.

	Verbal stems			*Nominal stems*	
	e.g.	*Launey (2011) section*	*Others*		*e.g.*
DROP1	quiz-a (intr.) tlacamat-i (tr.)	8.3, 8.4	Class IV, Canger (1980) Class B, Andrews (2003) Class 2, Lockhart (2001)	**TL**	tlaca-tl (an.) te-tl (inan.)
DROP2	tlahpalo-a (tr.)	8.5	Class II, Canger (1980) Class C, Andrews (2003) Class 3, Lockhart (2001)	**TLI (includes LI)**	nan-tli (an.) cal-li (inan.)
NODROP1	itt-a (tr.) pan-o (intr.)	8.7	Class III, Canger (1980) Class A, Andrews (2003) Class 1, Lockhart (2001)	**IN**	oculi-in (an.)
NODROP2	cu-a (tr.) zom-a (tr.)	8.6	Class I, Canger (1980) Class D, Andrews (2003) Class 4, Lockhart (2001)	**CHICHI**	chichi-θ (an.)

Table 4.3: The eight main categories of stems used by the model. The verbal categories defined in the model according to Launey (2011) are presented along the equivalent class given in three other sources. The nominal categories are further classified as animate (an.) and inanimate (inan.).

Launey's grammar was chosen as the principal guide mainly because it is the most up-to-date CN grammar available. Its recent publication in English entailed a revision of the original French edition of 1979. Like Andrews, Launey explains grammatical points with rigour and provides plenty of examples, taking advantage of phonological studies about issues generally missed in old grammars. Launey's explanations, however, are more approachable than those of Andrews, as they do not hang too much on linguistic arguments and use less obscure terminology. Likewise, the model intends to be approachable beyond the sphere of specialists, and its performance might be more easily understood if it goes along the most pedagogical explanations of Launey. Andrews's grammar, however, was invaluable as a reference source. Grammars by Carochi (1645) and Wright-Carr (2007) were other valuable sources, particularly regarding irregular verbs and invariant words.

Wimmer's dictionary was chosen because it departs from the two principal dictionaries produced for CN: the vocabulary of Alonso de Molina and the dictionary of Remi Simeon. In addition, Wimmer confronts and enriches these important sources with other recent dictionaries, e.g. Karttunen (1992), and with other works, most significantly the Launey and Andrews (2003) grammars. It also contains abundant examples, in a normalised orthography, extracted from colonial texts, most notably the *Florentine Codex*. In addition, the dictionary is available online, which saved us a potentially complicated conversion from paper to electronic form. All the entries under an initial letter, nevertheless, are presented as running text in a single web page[3]. It was necessary, therefore, to segment the running text to generate a database. The dictionary uses certain tags to mark different information for each entry, but these anchors are not uniformly used in many cases. Consequently, much potentially useful information was lost in the compilation of the database, but the most important data for our purpose, e.g. the type of transitivity of a verbal stem, was extracted and used for the necessary classification of stems.

[3] The information in the dictionary is not available in any other format that distributes the data necessary for the classification of a stem in a manageable way (Personal communication with Alexis Wimmer, November 2016). The dictionary is also found as one of various vocabularies compiled in the project *Gran Diccionario Nahuatl*. The web interface of this project, however, does not allow for the easy exporting of a single dictionary. Moreover, although the information in this website is presented in a tabular format, the data needed to classify a stem are merged into a single column with no significant anchor available to assist segmentation, just like the text available in the web page presenting Wimmer's dictionary alone.

PREFIXES		*STEMS*	*SUFFIXES*
Lexicon Subj	*Lexicon **DefObject***	*Lexicon **VerbsTr***	*Lexicon **Present***
Subj1Sg : ni	Obj1Sg : nech		
Subj2Sg : ti	Obj2Sg : mitz		PresSg : **0**
Subj3Sg : 0	Obj3Sg : ^C	e.g.	
		^tlahpaloa (to greet)	
Subj1Pl : ti	Obj1Pl : tech		
Subj2Pl : an	Obj2Pl : amech		PresPl : h
Subj3Pl : 0	Obj3Pl : quin		

Table 4.4: Four lexicons to exemplify the modelling paths in a FST.

4.2.2 Modelling the paths of the NNC and VNC transducers

The use of lexicons, of morphemes and stems, and of continuation classes in lexc script results in paths representing concatenations of morphemes. To provide examples in the following explanations, **Table 4.4** illustrates four lexicons used to model the paths for the present tense of transitive stems in the VNC transducer.

The lexicons and continuation classes used to compile the VNC and NNC transducers are indeed much more complex. For starters the lexicons used for VNC prefixes, for example, are more numerous (7 in total), as they reflect the following arrangement:

Augment+Subj+DefObj+Directional+Reflexive+IndefObjTe+IndefObjTla +^**Stem**+ Tense/Mood Suffix

More important, valid paths are defined not only as simple concatenations of all the morphemes in every lexicon, but as concatenations which are valid according to CN morphotactics. The VNC transducer required the consideration of certain restrictions and requirements, like the *valence* of a stem, i.e. the absence or presence of an object (Andrews, 2003:46), and the agreement in number between a subject prefix and the tense/mood suffix.

Flag-diacritics are added strategically in the lexicons to manage these separated dependencies. When the model analyses or generates, flag diacritics enforce, e.g., that a singular subject prefix must relate only with a singular tense suffix, that an intransitive stem must not take an object, and that the *Augment*,

not properly a prefix, but traditionally written as such (Launey, 2011:73), must appear only with the preterite and pluperfect tenses.

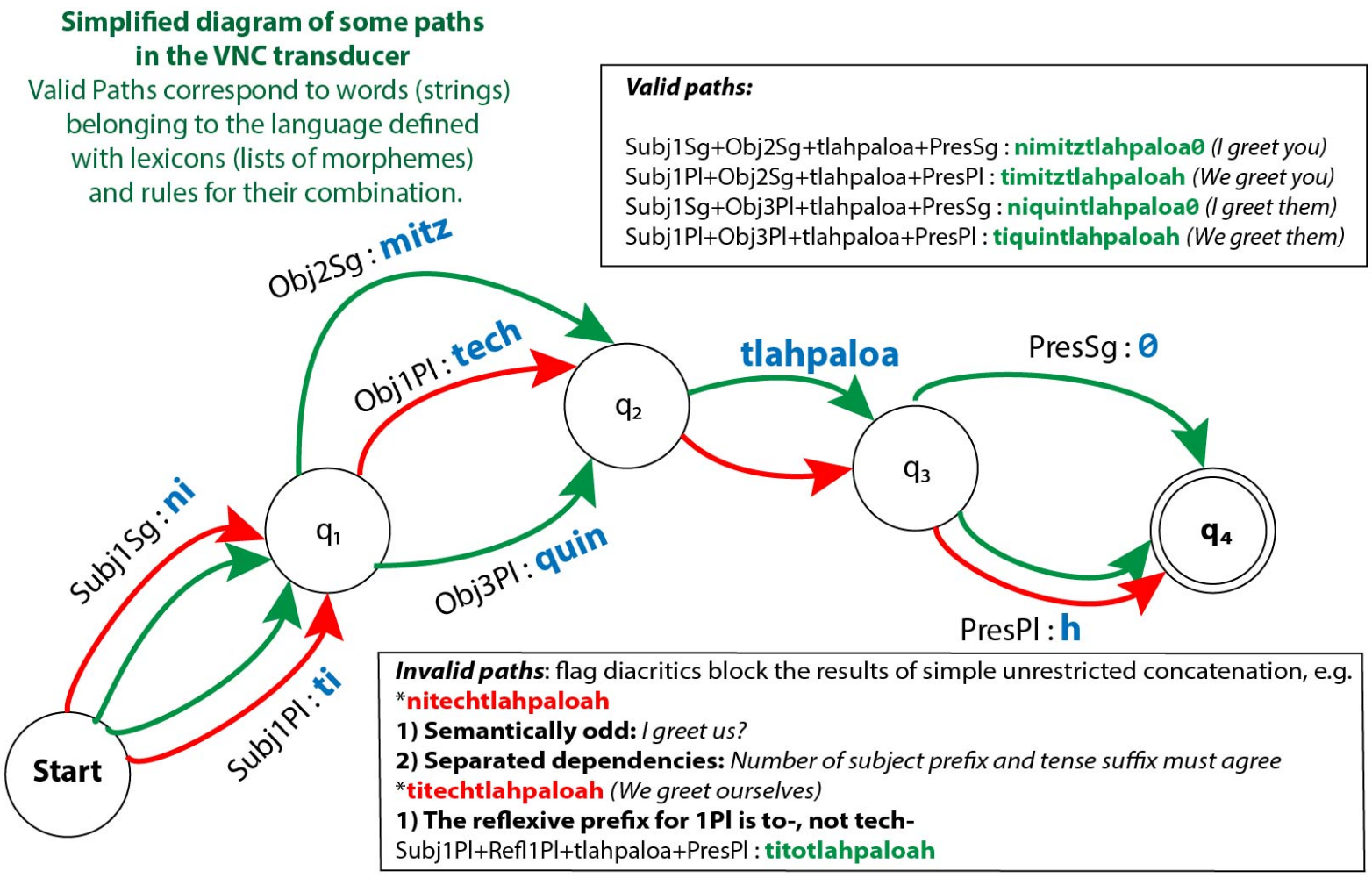


Figure 4.2 Some paths in the transducer of VNC.

Figure 4.2 shows a simplified diagram of some paths in the transducer of VNC. The strings in the upper and lower sides are separated by a colon (:), and, for simplicity's sake, the transitions (arcs) are represented between strings –like 'ni'- and not between single-character symbols –like 'n', 'i'. The upper language is formed with concatenations of a stem (like *tlahpaloa*, appearing in both sides of the transducer) with the tags for the grammatical function of concatenated affixes. The lower language contains abstract symbols like `0` which represents the present-singular suffix. Flag diacritics block the illegal path containing a singular subject prefix *ni-* combined with a plural suffix for the present *–h*. Although such invalid paths, the result of simple unrestricted concatenation, are still part of the network, they are not considered for analysis and generation thanks to the flag-diacritics.

The granularity of the analyses performed by the model is somewhat restricted, but enough for our exploratory purposes. The NNC and VNC transducers use nominal and verbal stems as collected from the dictionary; *stem* means here bases which could be further analysed into component morphemes. No attempt has been made to define or use exclusively *roots*, i.e., monomorphemic bases. Likewise, there was no attempt to model, for example, other morphological relations like derivation or compounding. If a derivative or compound is found as a dictionary entry, its corresponding base is used in the appropriate lexicon. Take for example, the *causative*, a verbal derivative meaning 'to cause the (object) to VERB', and which changes the valence of the verb, i.e. intransitive verbs become transitive, and transitive ones become bitransitive (Launey, 2011:189). Instead of modelling the formation of the causative for all verbal roots, the base of causative forms found as entries in the dictionary are used in the appropriate category (transitive verbs) in the lexicon of verbal stems. This means that for a word like *quimictia*, 'he kills it', the model will produce only the analysis:

Subj3Sg+Obj3Sg+mictia+PresSg[4]

The stem *mictia*, 'to kill/he kills', is certainly the causative of *miqu-i*, 'to die/he dies', and could be traced back to it and interpreted therefore as 'to cause somebody/something to die'. In such case, one could expect the analysis:

[4]The analyses of a string by our model is presented according to the tags defined in our model for each affix listed in tables from **Table 2.3** to **Table 2.9**

Subj3Sg+Obj3Sg+miqui+CAUSATIVE+PresSg

However, no such deeper analysis is achieved because: one, *mictia* was found in the dictionary and added to the model as a transitive verbal stem; and two, the model has not defined the morphotactics to form a causative from roots like *miqu-*.

Reduplication, a non-concatenative process consisting in prefixing a stem with a variant of its first syllable, is not modelled either. Reduplication is used, for example, to form an alternative plural of some nouns (Launey, 2011:20-1), and to form "intensive" versions of verbs or significantly change their meaning (Launey, 2011:286-7). Stems resulting from reduplication, if found in Wimmer, are added to a suitable lexicon, just like compounds and derivatives. Thus, both *nemi*, 'to live/he lives' and its reduplication *nehnemi*, 'to walk/he walks' are included in the model in the appropriate verbal category. In consequence, the stem *nehnemi* is not analysed by the model in relation to the root *nemi* just as *mictia* is not analysed in relation to *miqu-*.

Although not modelling reduplication in the core module can have an impact on the recognition of certain plural nominal forms, this omission is somewhat alleviated by the use of the guesser (see **section 4.3.3**). *Mich-in*, 'fish', for example, has two possible plurals[5], *mich-meh* and *mimich-tin*. The latter will not be recognised by the core module, because only *mich-* is a defined nominal stem in it; however, thanks to the plural ending –tin, the guesser will propose it as a plural form built around a hypothetical stem *mimich-*. Regarding the verbal forms, it might even be better to treat reduplications, just like derivatives and compounds, independently from their originating stems, especially if a semantic distinction, potentially important for other applications[6], exists.

The construction of a more efficient or complete FS model, in which only roots (monomorphemic stems) are used in the lexicons and all possible morphological relations are modelled, is beyond the aims and possibilities of this research. Our FS model is expected only to achieve plausible mappings between orthographical strings in the analysed texts and CN morphological analysis de-

[5] In fact, the plural was never very fixed and many doublets can be found (Launey, 2011:21). This is the reason for allowing for alternative plural forms in the model.

[6] This seems to be the case for Gutierrez Vasques (2015) who focuses on the extraction of a bilingual lexicon from a parallel Spanish-Nahuatl corpus; for this purpose, it might be more important to distinguish mictia, 'to kill' and miqui, 'to die' as two different verb stems than relating them based on a common root (Personal communication, April 2017).

fined in the transducer, and not, for example, to support the study of derivation or compounding.

4.2.3 Adding orthographical levels to the transducers

Replace rules (Beesley & Karttunen, 2003:132), an extended notation for regular expressions in XFST, are used to map the morphological analysis in the transducers of NNC and VNC to its orthographical realisations. Replace rules are shorthand notations for complicated regular expressions built with basic operators. As they represent a regular relation of an upper and lower language, they are compiled as transducers that can be *composed* (see below) with the ones built for VNC and NNC.

Transducers of replace rules are used, for example, to link symbols representing an affix in the CN transducers with its orthographical realisation, which is conditioned by the context of occurrence of the symbol. Thus, the symbol ˆC, representing the object prefix for the third person singular, can be realised orthographically as <c>, <qu> or <qui> depending on the context of appearance, which can in turn be specified in the replace rules notation. Phonological processes like the devoicing of certain consonants in word/syllable-final position (Launey, 2011:8) or the assimilation of nasals (Launey, 2011:14), among others, are also given orthographical realisation by the replace rules.

The 'canonical' orthography used in the lower language of the model comprises the following sets:

Consonants C = {c, ch, cu, uc, h, hu, uh, m, n, p, qu, t, tl, tz, x, y, z}

Vowels V = {a, e, i, o}

The character <u> never represents a vowel, and appears only as part of digraphs representing consonants. Launey (2011:6) uses a grave accent <ˋ> above a non-final vowel and a circumflex <ˆ> above a final vowel to indicate that that vowel is followed by the glottal stop /ʔ/; the orthography of the model uses the character <h> for this consonant in both cases. Vowel length is not currently represented in the orthography because it rarely appears marked in either colonial or contemporary texts. Although vowel length can be important for semantic purposes (see Carochi's early awareness of this fact in Carochi,

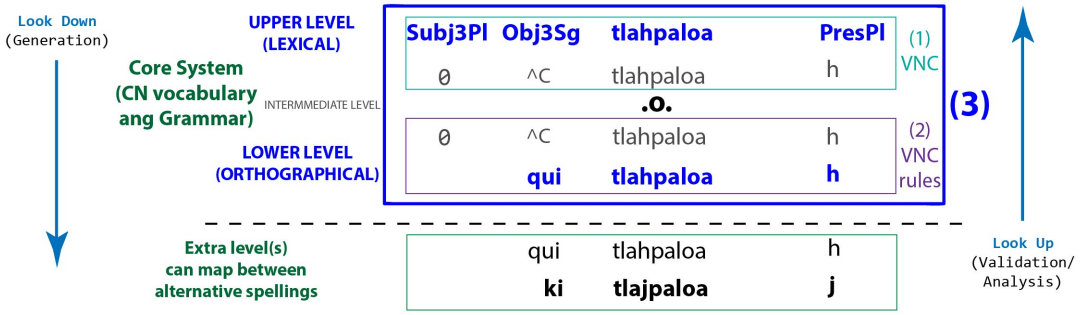


Figure 4.3 Composition of transducers.

2001:469), it has not seemed essential for the morphological processes modelled so far. In any case, vowel length can be added to the core transducers in a future version without many complications.

Composition is a finite state operation between two transducers that eliminates the intermediate linking strings and conserves only the two outermost languages, namely the upper level of the transducer above and the lower level of the transducer below. Therefore, the result of a composition between a transducer of CN nuclear clauses and its corresponding replace-rules transducer is one with morphological analyses in the upper level and orthographical words in the lower level.

Figure 4.3 illustrates in an impressionistic way the result of adding orthographical alternations to the model by means of composition. If (1) is a path in the VNC transducer and (2) one in the transducer of replace rules for VNC, (3) will be the corresponding path in the transducer resulting from the composition:

(1) **Subj3Pl**+**Obj3Sg**+**tlahpaloa**+**PresPl** : 0^Ctlahpaloah

(2) 0^Ctlahpaloah : **quitlahpaloa**

(3) **Subj3Pl**+**Obj3Sg**+**tlahpaloa**+**PresPl : quitlahpaloah**

In a similar way, successive rules can be compiled in other transducers and used to map alternative spellings (e.g. kitlajpaloaj) to the 'canonical' orthography used by the model. The use of replace rules and composition are one valuable advantage of FST for the exploration of convergence in our contemporary texts: if a NC in a text resembles a CN one, it can be related to a CN morphological analysis in spite of a different spelling. This can be achieved without modifying the core CN transducer or the text.

4.3 The construction of the model

The model consists of three main modules that are used in cascade. The first is a stop-list transducer including proper names and Spanish words found across the NT texts. The second, the core module, is a FS transducer of NNC, VNC and particles. The third is a guesser that proposes plausible stems of strings resembling nominal and verbal clauses.

4.3.1 The stop-list transducer

A list of names and Spanish words was first compiled through an exploration of the texts using scripts written in Python. These tasks are largely trivial when working with pre-processed, annotated corpora or with texts in languages with established written norms. They required careful preparation, however, when approaching our texts for the first time, without the aid of a Nahuatl-dedicated tool like our FS model.

In our test corpus, some biblical names appear in a Spanish version (*Babilonia*) and in an English version (*Babylon*), so available lists of biblical names are only partially useful. In addition, names change from text to text, and variations can also appear in the same one: *Arfaxad* (G) appears as *Arfacsad* in Mich; *Abraham* and *Abrajam* are found in G, whereas *Abrahám* is found in Mich. The initial list of proper names was compiled from the text G; using this first list and the edit-distance method (available in the NLTK library for Python) variations of the names were spotted based on the similarity between strings, and the list of proper names successively updated with each new text.

Spanish-looking strings were first identified using the PyEnchant spell-checking library in a Python script. These strings are called here Spanish-looking, because they could not be immediately taken as Spanish words; there can be Nahuatl NCs that resemble Spanish words. Using a spell-checker as only aid, word-types like *chocas* ('you crash into' in Spanish, '(s)he/it will cry' in Nahuatl, according to the spelling used in G), could easily be erroneously discounted as Spanish.

The first tests of the FS model complemented this weak initial approach to the compilation of a stop-list. It became clear that one first useful application of the model was indeed as a recogniser of Nahuatl words. Applying the model,

the identification of more Spanish strings, names, text-conversion errors and typos became more straightforward. The rejection of strings is caused not only by the presence of a typical Spanish character, neither defined nor mapped with alternation rules to one in the alphabet of the model, like <r> or <ñ>; rejections also have to do with the lack of recognisable structure, phonological or morphological, of a Nahuatl-looking string, e.g.:

**qujtohuaj*	typo for *quijtohua*
**otocalistliy*	foot-note marker 'y' attached to the NNC *otocalistli*
**yunta*	Spanish word for 'yoke' in the model <u> is not a vowel and therefore <yun> is not a valid syllable
**ltijtinemi*	two initial consonants

Unlike *chocas*, 'you will cry' (Subj3Sg+choca+FutSg), such strings will not be given an analysis, and so they can be spotted as candidates to update the stop-list or clues to correct typos in a text. The compilation of our stop-list exemplifies how and why FS transducers can be the base of useful applications, e.g. spell-checkers, for morphologically complex languages like Nahuatl.

4.3.2 The core FST

The core module of the model is largely formed of the union of two transducers: Nominal Nuclear Clauses (NNC) and Verbal Nuclear Clauses (VNC). Each transducer is compiled separately and composed with its corresponding orthographical-alternation rules before performing the union for the core module. The core module was complemented with a transducer of irregular verbs and one of invariant words. **Figure 4.4** illustrates how the core module is created.

The irregular verbs were modelled following examples from Launey (2011), Carochi (1645), and Wright-Carr (2007). They are not numerous (Launey considers only three, Carochi and Wright-Carr, seven), but the choice made was to put them in a separate transducer to avoid an increment in size and complexity of the VNC transducer. This decision entails maintaining three separate lexicons for the Subject Prefix: one in the VNC, one in the NNC and one in the irregV transducer. However, the maintenance of a single separated lexicon in three transducers is a low cost compared to the modularity given to the system: keeping a simple VNC makes it easier to test and update it with new forms for

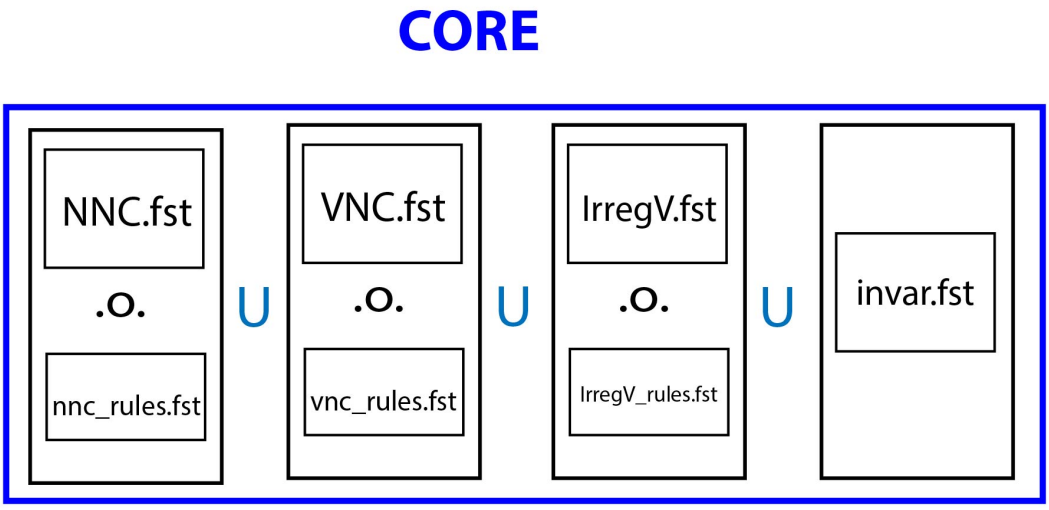


Figure 4.4 The core transducer and its comprising modules.

regular verbs without having to add restrictions applying to irregular verbs only. The irregular verbs transducer has its own alternation rules, just like the NNC and VNC ones.

A transducer of invariant forms includes the items called *particles* (Wright-Carr, 2007:171), and others like emphatic pronouns, interrogatives and 'conjunctions'. All of these invariant words are listed in a single lexicon of 255 entries without continuation classes that compiles in a simple transducer.

Particles, along with some more or less adverbial expressions, are called 'adverbs' by Carochi, which is still an invaluable guide on the subject (Lockhart in Carochi, 2001). Launey (2011) covers them, although in different places along his grammar. This research used, therefore, the lists in Wright-Carr (2007:177-87) and the whole book 5 of Carochi's grammar as useful guides for its compilation, although comparing to compare this sources with Launey's explanations. Although they are not inflected, Nahuatl particles are forever clustering and scattering (Lockhart, 2001:103). Wright-Carr lists various apparent compounds (e.g. *quinihcuac*, 'then, after something referred', perhaps resulting from quin, 'after', and *ihcuac*, 'when, then') and attempts the analysis of many. However, it is not clear to me how systematic such a compounding process can be, nor whether listed particles as *quinihcuac* or *inihcuac* are really compounds or rather orthographical variations representing as a single word what could be otherwise written as a 'particle phrase', e.g. quin *ihcuac* or in *ihcuac*. It seems, after all, that although the clusters of particles become very significant, "they never become 'words' as we know them in English" (Lockhart, 2001:103). All the par-

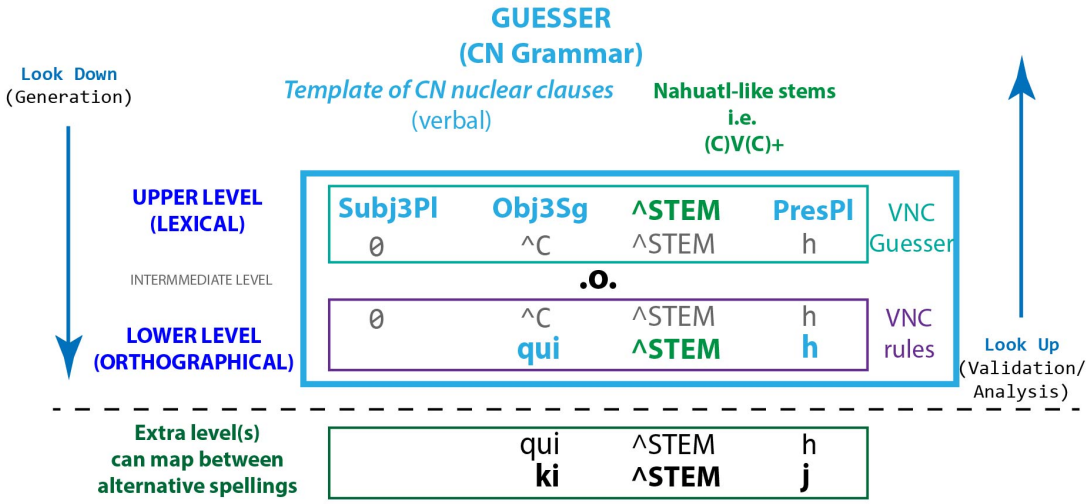


Figure 4.5 A VNC generated by the guesser. The VNC is a form corresponding to the third person plural of a transitive hypothetical stem formed by one or more syllables with the form (C)V(C). The guesser can be further composed with transducers mapping between alternative spellings, just like the core transducer.

ticles found in Wright-Carr and Carochi, isolated or in clusters, are just listed in lexicons without continuation classes. In consequence, the upper words of this transducer do not form a series of tags of comprising morphemes, but rather single tags describing the function of the particle, and each analysis produced is just one of such tags. The analysis 'Temporal_achtopa', for example, reads as 'the particle *achtopa*, "first", which is a particle with temporal meaning'. The transducer of invariant forms is not composed with any rules.

4.3.3 The guesser

The definition of morphological structures with a dedicated slot for stems sharing the same morphological behaviours allows for the creation of a guesser (Beesley & Karttunen, 2003:444-51).

The guesser is a transducer that proposes plausible stems for strings whose structure resembles a VNC or a NNC. Like the core transducer, the guesser is formed of legal paths defined by lexicons, continuation classes and flag diacritics. Unlike the core transducer, that uses stems compiled from Wimmer's dictionary, the guesser generates VNCs and NNCs by concatenating valid combinations of affixes around a hypothetical stem which is phonologically plausible. **Figure 4.5** illustrates a VNC generated by the guesser.

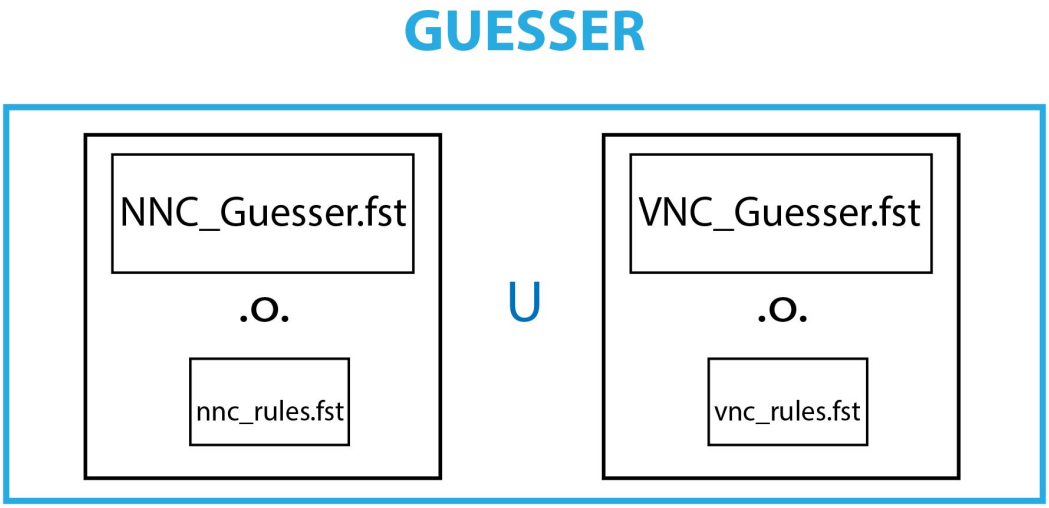


Figure 4.6 The guesser and its comprising modules.

In CN a syllable has the form (C)V(C), i.e. a vowel optionally preceded and/or followed by a consonant (Launey, 2011:13). In general, the guesser uses stems defined with the regular expression ***(C)V(C)+***, which stands for 'a combination of one or more syllables with the form (C)V(C)'. Such general definition was refined with two restrictions. First, strings beginning with the character <h>, which represents the glottal stop /ʔ/, were excluded as hypothetical stems; no Nahuatl word begins with the consonant /ʔ/, at least from what can be seen in the dictionaries available. Second, only strings ending in <i> or <o> are considered as hypothetical stems for DROP2 verbs; this secures the characteristic endings <–i-a, -o-a> of this category of verbs.

The guesser is used to suggest new stems that might need to be added to the appropriate lexicon. Using the guesser, a string in the text can be recognised as having the morphological structure of a CN NC, even if its stem is not included in the core FST. This functionality is useful to recognise nuclear clauses that have the morphology of a CN but might be built around a contemporary lexical variant not registered in Wimmer's dictionary. No guesser was built for invariant words, as they do not have a recognisable structure, nor for irregular verbs. As shown in **Figure 4.6**, the guesser is created by the union of the VNC guesser and the NNC guesser only, each of which has been previously composed with its respective set of alternation rules.

4.3.4 The orthographical alternation rules for individual texts

Before analysing a text, the CN core and guesser are adapted by composing them with the transducers of alternation rules for the text in question. Each text has its own series of alternation rules. These sets of rules are different to the ones used to map morphological analysis of CN to their corresponding 'canonical' orthographical realisation, and which supress intermediate abstract strings like `0^C^tlahpaloah` in the core or `0^C^STEMh` in the guesser.

The orthographical alternation rules for each text, in contrast, map alternative graphemes used in the text to the ones defined in the core CN transducer, e.g. <k> to <qu-, c> representing the phoneme /k/ in *kitlahpaloah*. The purpose of these sets of rules is to tackle the orthographic heterogeneity of the corpus. When exploring the morphological convergence between CN and the texts in the corpus, these rules can map nuclear clauses in the texts to a plausible morphological analysis modelled in the core CN transducer.

Although individual rules can be used to map orthographic alternations that are recurrent in most texts of the test corpus[7], their combination and order vary from text to text. Therefore, the choice made was to keep a separated set of rules for each text.

4.4 The analysis of a text

The word-types alphabetically sorted in the word-lists of a text are used as input for the model in each test. Each word-type of the word-list is passed as input to the model and a corresponding output obtained. The input strings are either accepted by one of the transducers inside the 'black box', or rejected. The transducers are used in cascade, which means that the rejected strings of one serve as input for the next. Only the strings rejected by the last transducer, i.e. the guesser, are taken as rejections of the model.

The analysis in cascade helps to spot plausible points of convergence and divergence between the CN model and the text in question (**Figure 4.7**). The core FST produces a morphological analysis of an accepted string, which points

[7] e.g. `z -> s // _ [a|o]`, which reads as '<s> maps to <z> in front of <a> or <o>'

to a plausible morphological and lexical convergence between CN and the text. The accepted string plausibly is a CN NC or particle. If not recognised by the core FST, a string is passed to a 'guesser', which proposes plausible analysis of strings whose morphology resembles that of a CN NC although is not fully recognised by the core due, perhaps, to lexical divergence. Finally, the strings not recognised by the guesser are likely to point to divergence in both morphological and lexical terms.

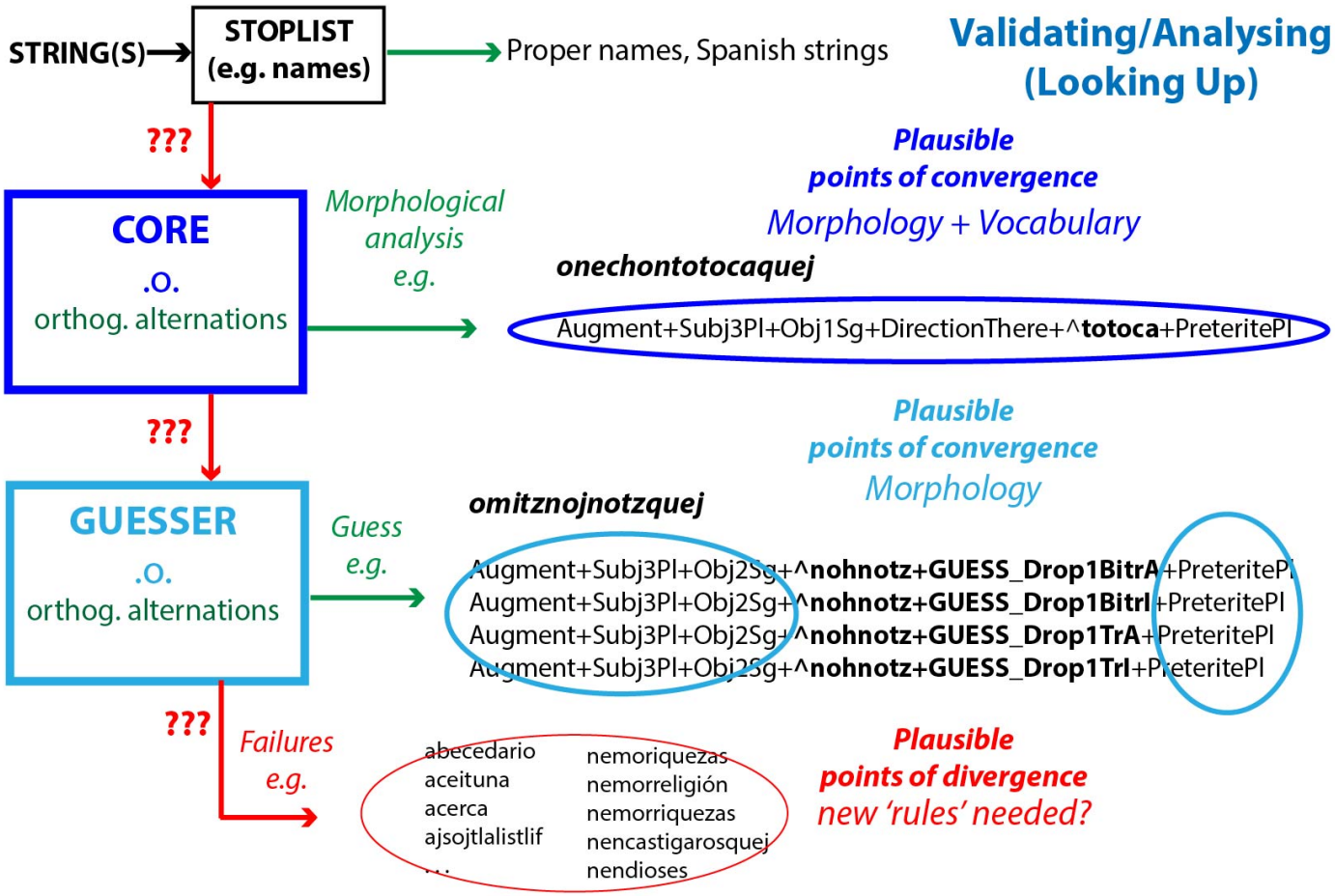


Figure 4.7 Cascade analysis of a string. A series of transducers, each composed with suitable orthographical alternations for a text, are used in cascade to point to plausible points of convergence between CN and the text analysed

4.4.1 What does the core show in terms of convergence?

The core FST produces a morphological analysis of a recognised string, which points to plausible morphological and lexical convergence between the string and a CN nuclear clause. If a string from the text can be mapped to a morphological analysis in the CN model, the string is equivalent to one generated by the model through a valid combination of CN affixes with a defined stem. In **Figure 4.7**, for example, the string *onechontotocaquej* 'they chased me', is analysed as a plural form of the preterite tense, as explained by Launey (2011), of the verbal stem *totoca* 'to chase, to drive away', listed in the dictionary of Wimmer (2006). This string, therefore, is a plausible point of convergence between the text and CN in terms of morphology and vocabulary.

The recognised strings followed by their morphological analysis are compiled in a file. Besides a list of stems and recurrent combinations of affixes, an index of ambiguity can be obtained from this file by counting the average number of analyses a recognised string receives. The strings that cannot be given an analysis by the core transducer are compiled in a file that serves as input for the guesser.

4.4.2 What does the guesser show in terms of convergence?

The guesser helps to identify plausible morphological convergence despite lexical divergence. The guesser makes proposals of morphological analysis for the strings that were not recognised by the core FST. Each analysis is based on the CN affixes that could be recognised around a hypothetical CN-like stem. In this way, one can identify strings whose morphology resembles the CN morphology and obtain proposals to fill in possible gaps in the list of defined stems in the model. In theory, the string could be recognised by the core model as soon as the correct stem is added.

The guesser will typically make more than one proposal for each accepted string. This happens because the hypothetical stem is tested in each of the categories defined for nominal and verbal stems. A string like *omitznojnotsquej* obtains, among others, four guesses as a plural form of the preterite of a hypothetical stem *nohnotz-*. As a result of the process to form the preterite, namely

by dropping the final vowel of stems ending in <i> or <a> in the present (the dictionary form), the string in question could have been formed from a stem *nohnotz*-i or *nohnotz*-a. Moreover, in the model there are two categories of verbal stems that can take an object, namely bitransitive and transitive. In consequence, four guesses are contemplated as equally plausible as shown in **Figure 4.7**. Less elaborated forms like the present will receive fewer guesses. However, it will be relevant to observe how dramatically ambiguity could increase after, for example, the addition of orthographical alternation rules.

The recognised strings with their guesses are saved in a file, which serves to determine the average number of guesses a recognised string receives. The strings that cannot be given an analysis by the core transducer are compiled in a file that serves as input for the guesser.

4.4.3 What do the rejected strings show in terms of divergence?

The rejections file contains strings with neither lexical nor structural resemblance to CN. In principle these files show points of divergence but the list requires further assessment to identify different causes of rejection.

There might be strings which are correctly rejected, not being Nahuatl words. This includes names, which can be easily spotted as capitalised initial, 'stand-alone' Spanish words (*aceituna*, *abecedario*) which can be recognised by a spell checking utility, and non-words resulting from errors in the conversion of the text. These strings could have been missed in the compilation of the stop-list, but the cascade analysis allows us to spot them to adjust the model and text.

After discarding the correctly rejected strings, two groups of actual failures of the model, pointing to divergence between the model and the text, can be identified.

One type of failure shows plausible lexical divergences between the text analysed and the CN model. These strings are not given a guess in the cascade analysis because, although they might contain affixes typical of a Nahuatl nuclear clause, they are built around a Spanish stem. Bear in mind that the guesser is designed to make proposals only for stems that have the form of a CN stem, i.e., one or more syllables in the form (C)V(C). After identifying 'stand-alone'

Spanish words with a spellchecker, the presence of characters not contained in the set of valid consonants C of the CN FST (e.g. r, g and d) allows for the recognition of Spanish stems integrated into Nahuatl constructions. Take for example, *nemorreligión* 'your (Pl) religion' and *nendioses* 'you are gods' in **Figure 4.7**.

The restrictive alphabet of the model will result in the rejection of all NCs using Spanish loans bearing "non-traditional" Nahuatl phonemes, which could be a significant number in contemporary texts. Nahuatl writers have long made use of Spanish loanwords, and this and other contact-induced phenomema extend in significant ways to modern varieties (Olko & Sullivan, 2013:199-201). In fact, loanwords present special orthographic problems, which are especially marked in what Sebba (2007:95-100) calls post-colonial texts and post-colonial orthographies, and could cause a significant percentage of rejection by a FS model based on a restrictive alphabet.

The second group is the result of probable morphological divergence. These strings might contain affixes different to the ones modelled for CN, which are characteristic of the variety in which the text is written. Diverging prefixes, for example, can be spotted following the alphabetically sorted list of rejections. Recurring patterns like the aforementioned prefixes *nen-* and *nemo-* are likely variants of the second person plural subject (in CN *am-*, *an-*) and object (in CN *amo-*) respectively. A similar method can be used to identify divergent suffixes after sorting the failures according to its ending.

The next chapter presents the results obtained when analysing our test texts. First it will present the results obtained against the CN gold standard to give an idea of the performance of our model against CN texts. Afterwards, it reports on the points of convergence found between each text and CN, and between the test texts. Finally, the points of convergence will be used to represent graphically the test texts as part of a network where the distance between texts depends on the number of connections they share.

Chapter 5

Results and discussion

5.1 Results for the analysis of the CN gold standard as point of reference

The FS model of CN used in this work is not only an approximate computational representation of the set of 'valid' CN word forms; at the same time, our FS model is a morphological analyser, as it maps between an orthographical word and a concatenation of tags representing its morphological analysis. The tests against the gold standard (GS) gives a point of reference about what percentage of a CN text can be expected to be recognised and analysed by either the core or the guesser, and gives also an idea of which word forms are recognised and which are missed by the model.

Table 5.1 shows the percentage of word-types recognised by the core FST and guesser for each of the texts in the GS. The model has a good rate of recognition as only a small percentage of the texts was not given a morphological analysis or given a guess (1.16% for Text 1, 2.56% for Text 2). Besides, most word-types were recognised by the core in both cases (80.23% for Text 1, 74.35% for Text 2) and less than one quarter of the word-types in the texts required a guess.

Confronting of the model with the GS also illustrates that a FST morphological analyser performs better than a procedural, truncation-based morphological analyser, in this case, Chachalaca version 12.01, which is the only morphological analyser available for CN (see **section 3.1.3**). The percentage of each text

	Text 1. Old man and death (86 word-types)		**Text 2**. The Fishermen (39 word-types)	
	FST	**Chachalaca 12.01**	**FST**	**Chachalaca 12.01**
Time of analysis (m:s)	0:1	23:54	0:1	09:45
Num. of word-types analysed by core	69 (68 correct)	43 (35 correct)	29 (29 correct)	15 (11 correct)
Analysed by core (%)	**80.23**	**50**	**74.35**	**38.46**
Num. of word-types given a guess	16	NA	9	NA
Analysed by guesser (%)	18.6	NA	23.07	NA
Num. of word-types not recognised	1	43	1	24
Not recognised (%)	**1.16**	**50**	**2.56**	**61.54**

Table 5.1: Performance of the FS model and Chachalaca when analysing two CN texts.

recognised by Chachalaca is smaller than the percentage recognised by our FS model (50% against 80.23% for Text 1 and 38.46% against 74.35% for Text 2). Besides, Chachalaca does not seem suitable for the analysis of larger corpora, judging by processing times. Notably, the time required to analyse the same number of word-types (86) is considerably higher for Chachalaca (almost 24 minutes) in comparison with a FST (1 second).

The metrics of precision, recall and F-measure, borrowed from the field of information retrieval are typically used to evaluate the performance of parsers and morphological analyser (Jurafsky & Martin 2009:489-91; Virpioja et al. 2011). In this work case, *precision* measures the percentage of analyses proposed by each application that were correct. *Recall* indicates the percentage of word-types present in the GS that were correctly analysed by each application. Finally, the *F-measure* combines precision and recall in a single metric, allowing for a global evaluation of the morphological analysers. The aforementioned metrics, shown in **Table 5.2**, also indicate that the core transducer of our FS analyser performs better than Chachalaca against our small GS. The FS analyser has a better recall than Chachalaca (0.776 against 0.368), meaning that it provides correct analyses for a higher number of word-types. It is also more precise than Chachalaca (0.989 against 0.793), meaning that a higher percentage of the proposed analyses are correct.

	Text 1. (86 word-types) **Text 2.** (39 word-types) **(T) Total of word-types: 125 word-types**	
	FST	**Chachalaca 12.01**
(A) Num. of word-types analysed by core	98	58
(CA) Num. of word-types correctly analysed by core	97	46
(P) Precision = CA / A	0.989	0.793
(R) Recall = CA / T	0.776	0.368
F-measure = 2PR / (P + R)	**0.869**	**0.502**

Table 5.2: Precision, recall and F-measures of the core transducer of the FS model and Chachalaca.

5.1.1 Discussion

5.1.1.1 Ambiguity

Ambiguity was recognised by Canger (2002:206-7) as one crucial issue in correlating an orthographic variant with a morphological analysis. She gives as an example the string <quioaliaoaloaia>, found in a CN corpus with rich orthographic variation. This string could have 406 readings, among other reasons because in the corpus she analysed the characters <i, j, y> could represent the phonemes /i/ and /y/, whereas <o, u, v> could represent /w/ or /o/.

The highest number of plausible analyses proposed for a recognised word-type in our model was considerable lower than the example provided by Canger, as can be seen in **Table 5.3**. For the Text 1, *nochan* 'my home', is the most ambiguous word-type recognised for the core (3 analyses). The most ambiguous word-type analysed by Chachalaca is *notlamamal* 'my load' with 23 analyses. For Text 2, *topan* 'on us', is the most ambiguous word-type recognised by the core (4 analyses); whereas *toyollo* 'our heart' is given 14 analyses by Chachalaca. The fairly normalised orthography used in our GS thus might be helping reduce the number of possible readings of a string.

	Text 1. Oldman and death (86 word-types)		Text 2. The Fishermen (39 word-types)	
	FST (0910-032202)	**Chachalaca (3010-1402)**	**FST (0910-034249)**	**Chachalaca (3010-1728)**
Ambiguity of analyses (core)				
(C) Num. of word-types analysed by core	69	43	29	15
(SC) Sum of analyses for each word-type proposed by core	82	153	36	45
Highest number of analyses for a word-type	3	23	4	14
Average Ambiguity of core analyses (SC /C)	**1.18**	**3.55**	**1.24**	**3.00**
Ambiguity of guesses (guesser)				
(G) Num. of word-types given a guess	16	NA	9	NA
(SG) Sum of guesses for each word-type	309	NA	192	NA
Highest number of guesses for word-type	55	NA	112	NA
Average Ambiguity of guesser analyses (SG / G)	**19.31**	NA	**21.33**	NA

Table 5.3: Ambiguity of analyses generated by the FS model and Chachalaca. The numeric codes in the header of each column identify the date and time of the test

Chachalaca seems to produce more ambiguous analyses than our core FS analyser. The average ambiguity (the sum of given analyses for each word-type divided by the total number of word-types) was used as rough metric of comparison. Whereas our FS analyser produced an average of 1.18 and 1.24 analyses per input string for Text 1 and Text 2 respectively, Chachalaca proposed an average of 3.55 and 3 analyses per input string. The highest number of analyses proposed for a word-type by the core FST is also notably lower than the figures for Chachalaca: 3 versus 23 for Text 1, and 4 versus 14 for Text 2.

The guesser understandably generates more ambiguity because it considers plausible combinations of affixes around different hypothetical stems. However, the more normalised spellings of Texts 1 and 2 again seems to reduce the number of possible readings for a word-type. The string from Text 1 that received more guesses (55) was <mocehuihticah>, whereas the most ambiguous string in Text 2 was <omociauhcauhqueh> with 112 guesses. The proposed guesses for these strings try different combinations of affixes, which increases the number of candidate analyses as compared with the number of candidate analyses proposed by the core. Text 1 and Text 2, however, do not present the alternations between <u> and <o> or <i> and <j> encountered by Canger in other texts, and the number of guesses are not further increased as they would by these optional alternations. In consequence, the highest number of guesses proposed for a word in the GS is still lower than the 406 readings of <quioaliaoaloaia> reported by Canger.

5.1.1.2 Analyses by the core

One word-type is considered as correctly analysed if one of the analyses proposed by the core is equivalent to the analysis given in the GS. The notation and granularity of the analyses proposed by our model do not correspond exactly to the one used in the GS. For example, the notation in the GS omits the tag for the third person singular and plural subject prefixes, because these affixes have no orthographic value (Subj3Sg+ : `0`, Subj3Pl+ : `0`). Besides, the long vowels marked in the GS with a macron were mapped to its equivalent short vowel, since the lexicons in our model do not mark vowel length. Therefore, the equivalence of the analyses for each string were verified manually to count them as correct.

For NCs, an analysis is considered correct if it matches the analysis in the GS based on three criteria: the same type of NC is identified, i.e. verbal or nominal; the same affixes are recognised; and the boundaries of the identified stem coincide. These last two criteria were disregarded when derivatives like the applicative and causative were marked in the GS. For example, *ittitia* 'to show', lit. 'to make somebody see something', is further analysed in the GS as a causative of *itt-a*, 'see', and the causative is marked instead of the tense:

tech-itt-itia ⇒ *1plO-see-CAUS* ***(GS)***
'he made us see something'

This analysis, available in the GS, had to be found an equivalent according to the granularity and tags used by the FST. The FST does not model the causative and cannot analyse *ittitia* as causative of *itta*. However, the core transducer does include *ittitia* as one of the transitive verbal stems found in Wimmer's dictionary, and is therefore able to analyse *techittitia*, 'he showed us something', in relation to the transitive stem *ittitia*:

techittitia ⇒ *Subj3Sg+Obj1Pl+ˆittitia+PresSg* ***(FST core analysis)***
'he showed us something'

In such cases the analysis is counted as correct[1].

In general, the core analysed correctly short invariant word-types (particles), e.g. *auh* and *zan*, and long NCs like *oquihuicaqueh* and *quinhualtocaya*, all of which were missed by Chachalaca.

5.1.1.2.1 Ambiguous analyses of the core One reason for ambiguous analyses is the omission of vowel length in our model. There is the case, for example of the nominal stem *cōcoh-(tli)*, 'dove' (and also 'throat'), and the transitive verbal stem *coco-(a)*, 'to hurt someone'. Thus, the input string *tecocoh*, is given two plausible analyses, because: in our model both stems are undifferentiated by the vowel length; and *coco-a* forms its preterit by dropping the final vowel and adding an <h>. Under these conditions, the two perfectly plausible analyses are given:

[1] The analyses of a string by our model are presented according to the tags defined in our model for each affix listed in tables from **Table 2.3** to **Table 2.9**. The caret character (ˆ) indicates the beginning of a stem. The right arrow (⇒) separates the candidate string (input) and the analysis proposed by the FST; it can be read as 'string x is given the analysis y'

tecocoh ⇒ *Subj3Sg+IndefObjTe+ˆcocoh+PreteriteSg* ***(Correct)***
'he hurt (preterit) someone'

tecocoh ⇒ *PossIndet+ˆcocoh+SgPoss* ***(Incorrect but plausible)***
'somebody's dove/throat'

This suggests that differentiating the vowel length might reduce the ambiguity of some analyses, although this also depends on whether the input string marks the vowel length or not.

Another reason for ambiguity is the inclusion of alternative forms of one same stem that are listed in Wimmer's dictionary, and recognised in other sources (e.g. Karttunen, 1992). For example, *pāni-tl* and *pāmi-tl* are reported as variants of *pān-tli*, 'flag, banner' and also 'row', and they all are included in the lexicon of nominal stems of the model. Due to the dropping of short /i/, and the articulation of /m/ as /n/ to form the possessive (Launey, 2011:91), the possessive forms of all the stems above would be based on the form *pān-*. Therefore, for inputs like *topan* there are four analyses:

topan ⇒ *Poss1Pl+ˆˆ+Rel_pan* ***(Correct)***
'on us' from the relational stem *–pan*, 'on'

topan ⇒ *Poss1Pl+ˆpan+SgPoss* ***(Incorrect but plausible)***
'our flag/row' from *pan-tli*

topan ⇒ *Poss1Pl+ˆpani+SgPoss* ***(Incorrect but plausible)***
'our flag/row' from *pani-tl*

topan ⇒ *Poss1Pl+ˆpami+SgPoss* ***(Incorrect but plausible)***
'our flag/row' from *pami-tl*

This suggests again that distinguishing vowel length in further versions of the transducer could help to reduce the ambiguity of some analyses. Marking the vowel length, however, would still allow for three analyses of *topān*, because of the alternative stems included in the model.

5.1.1.2.2 Failures of the core The core missed words that seem to be compounds of particles, e.g. *caoc*, a compound of the assertion marker ca and the particle with temporal meaning *oc*, 'again, still'. These compositions are

very dynamic (Lockhart 2001:103; Wright-Carr 2007:171) and thus seem difficult to predict, and in consequence have not been modelled. In any case, attested combinations of particles can be added with a corresponding analysis as irregularities.

NCs around compound stems are also missed, as composition of stems was not modelled. The core is unable to analyse, for example compounds with auxiliary verbs indicating a state or sort of movement (Launey, 2011:274-284):

oncholohtihuetz ⇒ *??*
'it fell into' (*cholo-a*, 'to escape' + auxiliary *huetz-i*, 'to fall')

mocehuihticah ⇒ *??*
'he is resting' (*cehui-a*, 'to rest' + auxiliary *cah*, 'to be')

Although all the verbal stems used in the compounds shown above are listed in the lexicons of the core transducer, it fails to analyse their compositions.

5.1.1.3 Analyses by the guesser

In general, the guesser provided good guesses for the word-types missed by the core. The output of the guesser for each input string is a range of possibilities expected to help identify gaps in the lexicon of nominal or verbal stems, among other limitations of the core transducer.

Naturally, the guesses are only as good as the coverage of forms modelled in the core. *Mocehuihticah*, for example, receives as one of its best guesses:

mocehuihticah ⇒
Subj3Pl+Reflex3Pl+^cehuihti+GUESS_Drop1TrA+PluerfectPl **(Incorrect)**

Although this hypothetical stem (*cehuihti-a*) and combination of affixes seem totally plausible, the model fails to recognise the compound with the auxiliary *cah*, and identifies <cah> as a marker of the pluperfect plural, also making the ligature -ti- part of the hypothetical stem.

One of the strings missed by the core, *omociauhcauhqueh* (translated in the GS as 'they had gotten tired') received 112 guesses. As a rule of thumb, the correct guess tends to be among the ones containing the shortest stem. The GS includes gives the following analysis:

omociauhcauhqueh ⇒ *PST=REFL-tired-PST-PL* ***(GS)***
'they had gotten tired'

for which the guesser proposes the guesses:

Subj3Sg+ˆomociauhcauhqueh+GUESS_Drop1IntrA+PreteriteSg
(Incorrect)

Subj3Pl+ˆomociahucahu+GUESS_Drop1IntrI+PreteritePl
(Incorrect)

Augment+Subj3Pl+ˆmociahucauh+GUESS_Drop1IntrA+PreteritePl
(Incorrect)

Augment+Subj3Pl+Reflex3Pl+ˆciauhcahu+GUESS_Drop1TrA+PreteritePl
(Correct)

The example above illustrates both the utility of the guesser and one limitation of the model in general. The stem recognised in the GS is indeed *ciauhcauh-*, and based on the translation, it could be a compound of two stems listed in the model *ciyahui*, 'to be tired' and *cahua*, 'to leave behind, abandon'. The guesser thus does a good job helping identify compound verbal stems to possibly enrich the model. However, this last example also illustrates another complication in analysing many strings: the lack of distinction between /ia/ and /iya/, often rendered in writing as <ia> (Canger, 2002:206). It is then possible that the core transducer will not analyse many word forms around stems like <ciyahui>, if the input strings are written as <ciahui>.

5.1.1.3.1 Ambiguous guesses

In general, the guesser provides many analyses because it exhausts all the plausible combinations of affixes around a hypothetical stem of the form (C)V(C)+. As the example of *omociauhcauhqueh* shows, long strings can receive many guesses, especially if their extremes resemble a suitable combination of affixes. In these cases, the rule of thumb is to choose the guesses containing the shortest stems.

In contrast, short strings like *caoc*, which is a composition of particles, do not contain a recognisable structure of affixes to help the guesser. In this cases, the whole string is often guessed as a noun stem of the very permissive category CHICHI, i.e. nouns without the suffixes *–tli, -li, -tl*, or *–in* in the non-possessive

form (Launey, 2011:232-4). This is a permissive category because if no other affix can be identified, the whole input will be considered at least a CHICHI[2] noun stem, provided it has the form (C)V(C)+. In fact, such a guess is often among the guesses for a string, thus increasing the ambiguity of a given input:

ˆomociauhcauhqueh+GUESS_An_C_chichi+SgUnposs
animate stem ending in consonant

ˆomociauhcauhqueh+GUESS_In_C_chichi+SgUnposs
inanimate stem ending in consonant

The possibility of being analysed as a CHICHI stem seems higher for short strings whose extremes do not resemble a combination of affixes, as *caoc* exemplifies:

caoc ⇒ *ˆcao+GUESS_In_V_chichiˆ+Rel_co*
Locative built on hypothetical inanimate stem 'cao'

caoc ⇒ *ˆcaoc+GUESS_An_C_chichi+SgUnposs*
hypothetical animate stem 'caoc'

caoc ⇒ *ˆcaoc+GUESS_In_C_chichi+SgUnposs*
hypothetical inanimate stem 'caoc'

Moreover, *caoc* reveals a faulty alternation rule that allows the guesser to propose *caoqu* as a plausible stem. Notice that the digraph <qu> and <c> represent the consonant /k/, but only <c> should occur at the end of a word:

caoc ⇒ *ˆcaoqu+GUESS_An_C_chichi+SgUnposs*
hypothetical animate stem 'caoqu'

caoc ⇒ *ˆcaoqu+GUESS_In_C_chichi+SgUnposs*
hypothetical inanimate stem 'caoqu'

The guesser of the model is a potentially valuable tool, but its utility depends on the coverage of forms by the model, and on the accuracy of alternation rules. Thus, although the guesser analysed a considerable percentage of strings missed by the core, one has to bear in mind some shortcomings, namely: that many forms, especially those not modelled e.g. compounds, might be given CHICHI

[2] *Chichi*, 'dog' is the stereotypical noun that bears no suffix in the absolutive (non-possessive) form. That is why *chichi* was chosen to name this category of suffixless nouns.

guesses, which are not very informative and increase the ambiguity of an analysis; that some missing or inaccurate alternation rules could be producing incorrect guesses. Fortunately, an evaluation of the guesser output reveals its limitations, and offers an opportunity to improve the core transducer, from which the guesser is derived.

5.1.1.3.2 Failures of the guesser

The only strings that did not get at least one guess were the vocative *miquiztlé*, 'Oh, Death', and *tunno*, a borrowing for 'tuna'. These failures are partly explained by another form not modelled in the core, namely the vocative; but they also show that the model will not recognise strings that are not 'valid' in terms of their component symbols and syllable structure, as defined for CN in the model. Thus, another reason for the failure is that accentuated characters like <é> are not part of the alphabet of the model. Likewise, the string *tunno* does not fit even in the permissive category CHICHI, because its syllable structure does not have the form (C)V(C). Bear in mind, that <u> never represents a vowel, but is only part of digraphs representing consonants, <qu, hu, uh, cu, uc> , and thus <tun> is not a valid syllable.

5.2 Results for the analysis of contemporary texts

We now turn to the analysis of a larger corpus of contemporary texts representing a broader range of dialects from the Nahuatl continuum. **Table 5.4** shows an approximate number of speakers related to the varieties covered by the texts in our analysis. Our texts relate to 7 of the 15 most spoken varieties recognised by INALI, and to one of the 15 with less than 8000 speakers (Mich). The NT versions appear under the name of one of the Nahuatl varieties recognised by the SIL, listed in Ethnologue. A rough equivalence between the varieties listed by INALI, Ethnologue and Canger (1980)) was determined using the names of localities (municipalities and villages) associated by each source with a variety.

In comparison to the analysis of the CN GS, the percentage of word-types recognised by the core transducer was lower, ranging from 11.98% (Mich) to

	Variety	Speakers (INALI)	Text (SIL)	Canger (1980)
1	Mexicano de la Huasteca Hidalguense	212,300	**HuE**	I.a. La Huasteca. Eastern Section
2	Mexicano de Guerrero	153,773	**G**	I.c
3	Náhuatl de la Sierra, noreste de Puebla	141,737	**PuebH**	II.b.i. Sierra de Puebla
4	Náhuatl de la Huasteca potosina	133,343	**HuW**	I.a.
5	Náhuatl de la Sierra negra, norte	131,015		II.b.ii. East of Puebla
6	Náhuatl central de Veracruz	130,979		
7	Náhuatl de la Huasteca veracruzana	130,364	**HuC**	I.a. La Huasteca. Eastern Section
8	Náhuatl del noroeste central	76,837	**PuebN**	I.b. North Puebla
9	Náhuatl del centro de Puebla	57,382		
10	Náhuatl de la Sierra negra, sur	32,321		
11	Náhuatl del Istmo	27,210		
12	Náhuatl de la Sierra oeste de Puebla	20,461		
13	Mexicano del oriente central	19,252		
14	Náhuatl de Oaxaca	8,556	**OaxN**	-Not covered Nearest covered locality is Zoquitlán in II.b.ii East Puebla
15	Mexicano del centro alto	8,100		
16	Náhuatl del Istmo bajo	7,707		
17	Mexicano central de occidente	5,999		
18	Mexicano de Temixco	4,199		
19	Mexicano del oriente de Puebla	3,817		
20	Mexicano de Puente de Ixtla	3,069		
21	Mexicano de Tetela del Volcán	2,972		
22	Mexicano del centro	2,623		
23	Mexicano central bajo	1,223		
24	Mexicano bajo de occidente	1,108		
25	Náhuatl alto del norte de Puebla	1,088		
26	Mexicano del centro bajo	841		
27	Mexicano alto de occidente	671		
28	Mexicano del noroeste	591		
29	Mexicano del oriente	286		
30	Mexicano de occidente	84		
	TOTAL	1,319,547		

Table 5.4: Approximate relation of the test texts to the Nahuatl varieties recognised by INALI. The test texts (Table 4.1) were approximately related also to the areas sketched by Canger (1980) (Table 2.1). Table adapted from De la Cruz Cruz (2014)

34.74% (OaxN), as shown in **Table 5.5**. The guesser performed relatively well, still managing to provide guesses for a good percentage of the word-types (from 43.65% to 70.18%). The highest percentage of non-recognised word-types occurs for the list from Michoacan (38.36%), and the lowest for Puebla Highlands (3.51%).

In terms of analysis time, the FS model analysed the longest text (21,819 word-types for PuebH) in considerably less time (1 minute 7 seconds) than the lowest time necessary for Chachalaca to analyse 39 word-types (9 minutes 45 seconds).

Text analysed	G	HuC	HuW	HuE	OaxN	PuebN	Mich	PuebH
Number of word-types	15,608	13,182	13,452	13,656	12,930	15,265	14,028	21,819
Time of analysis (m:s)	00:10	00:18	00:10	00:10	00:09	00:17	00:16	01:07
Num. of word-types in stop-list	513	489	486	488	491	494	486	493
In stop-list (%)	3.28	3.7	361	3.57	3.78	3.23	3.46	2.25
Num. of word-types analysed by core	4,160	3,417	2,989	3,140	4,492	4,064	1,681	5,245
Analysed by core (%)	**26.65**	**25.92**	**22.21**	**22.99**	**34.74**	**26.62**	**11.98**	**24.03**
Num. of word-types given a guess	8,617	5,808	5,873	6,294	6,882	10,167	6,479	15,313
Analysed by guesser (%)	55.20	44.06	43.65	46.08	53.22	66.6	46.18	70.18
Num. of word-types not recognised	2,318	3,468	4,104	3,734	1,065	540	5,382	768
Not recognised (%)	14.85	26.3	30.5	27.34	8.23	3.53	38.36	3.51

Table 5.5: Percentages of analysed word-types from contemporary texts.

5.2.1 Discussion

5.2.1.1 Orthographical alternations and ambiguity

An extra transducer of alternation rules was composed with the CN model for each text, in order to map the orthography used in the text to the one codified in the model. Consequently, just like in the example of Canger (2002), orthographical alternations increase the number of readings an input string can have.

There are characters (or sequences of characters) in the input strings that map almost unequivocally to only one choice in the CN model, and could be encoded as mandatory mappings representing one-to-one correspondences. This ideal case occurs, for example, with <k> in OaxN, which maps directly to either <c> or <qu> in the model, according to a relatively simple context of occurrence.

For other characters, however, the correspondence was not as unequivocal as one could wish. The example is the mapping from <j> in the texts, to <h> in the CN model. This simple correspondence seems often to be the correct one, as in the following word-types from Guerrero[3]:

nejnemi ⇒ 0-ˆ*nehnemi*-0 ⇒ *Subj3Sg*+ˆ*nehnemi*+*PresSg* ***(G)***
'he walks'

nejnemij ⇒ 0-ˆ*nehnemi-h* ⇒ *Subj3Sg*+ˆ*nehnemi*+*PresPl* ***(G)***
'they walk'

However, in Guerrero a <j> will also appear where the model would expect an <hu, uh>, apparently because in speech the consonant /w/ becomes /h/ before /k/ (Maxwell & Amith, 2005:482). Thus, in the text G one finds strings like *oquincojquej*, 'they bought them'. By including an alternation rule, one could plausibly analyse the string as follows:

oquincojquej ⇒ *o*-0-*quin*-ˆ*couh-queh* ⇒ *Augment*+*Subj3Pl*+*Obj3Pl*+ˆ*cohu*+*PreteritePl* ***(G)***
'they bought them', from *cohu-a* 'to buy'

[3] Note that the intermediate strings like 0-nehnemi-0 do not appear in the analyses, and are given here just to guide the following of examples. Once the FST for CN is composed with the alternation rules for each text, the input string *nejnemi* is directly mapped to the plausible morphological analysis *Subj3Sg*+*nehnemi*+*PresSg*.

Such an alternation was not added to the rules for G since the mapping from <j> to <hu, uh> in front of <c, qu>, cannot be generalised without defining a (potentially complex) context of occurrence, as the following example shows:

oquinelojquej ⇒ *o-*0*-qui-ˆneloh-queh* ⇒ *Augment+Subj3Pl+Obj3Sg+ˆneloh+PreteritePl* ***(G)***
from *nelo-a*, 'stir up, beat', a Drop2 verb whose preterite form is built upon the Base2 *neloh-*

In this case, <j> in front of <qu> does represent <h>, and the model correctly analyses the input *oquinelojquej*, 'they beat him'.

One possible course of action could be defining optative alternation rules. The disadvantage of optative mappings is that they provide extra readings (for example, for each <j> in a string), unless more complex contexts of occurrence are defined. The optative rules included are those mapping alternations related to widely identified phonological differences between Nahuatl varieties, which are often used as isoglosses to mark dialectal boundaries (see **Figure 2.2**). This is the case for optionally mapping <t> to <tl> in PuebH, and <l> to <tl> in Mich, which yields the plausible readings:

nechtacamati ⇒ 0*-nech-ˆtlacamati-*0 ⇒ *Subj3Sg+Obj1Sg+ˆtlacamati+PresSg* ***(PuebH)***
'he obeys me', from *tlacamat-i*, 'obey'

molalis ⇒ 0*-mo-ˆtlali-z* ⇒ *Subj3Sg+Reflex3Sg+ˆtlali+FutSg* ***(Mich)***
'he will sit', from *tlali-a*, 'collocate, sit'

Although an optional mapping of <l> to <tl>, for example, could theoretically cause the readings **mo-ˆtlatli-z* and **mo-ˆlatli-z*, both readings are ruled out because only the stem *tlali-a* is listed in the lexicon of verbal stems. Thus, the undesirable effect of optional mappings on ambiguity is somewhat counterbalanced in the core transducer, as the comparisons in **Table 5.6** show. The average ambiguity for the core for each text ranges from 1.09 (G and OaxN) to 1.13 (HuE), and in each case is even slightly lower than the ambiguity produced analysing the two texts of the GS (1.18 (Text 1) and 1.24 (Text 2) in **Table 5.3**). The highest number of analyses for a string does increase for each text, although it is never higher than 9 (PuebH).

	G	HuC	HuW	HuE	OaxN	PuebN	Mich	PuebH
Ambiguity of analyses (core)								
(C) Num. of word-types analysed by core	4,160	3,417	2,989	3,140	4,492	4,064	1,874	5,245
(SC) Sum of analyses for each word-type proposed by core	4,575	3,930	3,367	3,579	4,910	4,563	1,681	5,881
Highest number of analyses for a word-type	8	5	5	5	5	8	6	9
Average Ambiguity of core analyses (SC /C)	**1.09**	**1.15**	**1.12**	**1.13**	**1.09**	**1.12**	**1.11**	**1.12**
Ambiguity of guesses (guesser)								
(G) Num. of word-types given a guess	8,617	5,808	5,873	6,294	6,882	10,167	6,479	15,313
(SG) Sum of guesses for each word-type	269,262	488,846	170,744	186,669	198,606	454,465	364,465	2,539,869
Highest number of guesses for word-type	384	3,504	497	496	272	992	56.25	12,096
Average Ambiguity of guesser analyses (SG / G)	**31.24**	**84.16**	**29.07**	**29.65**	**28.85**	**44.7**	**54.00**	**165.86**

Table 5.6: Average ambiguity generated for core and guesser analysing contemporary texts.

The guesser, in contrast, has to test all possible alternations of a hypothetical stem, and would produce a guess around both **tlatli-* and **latli-*. Moreover, and beginning with the permissive category CHICHI, the guesser will try guesses for different categories of stems, e.g. animate or inanimate nominal stems, or vowel-dropping verbal stems. The most ambiguous string analysed by the guesser was *tiitamelaucatatitaniluan* (PuebH) with 12,096 guesses, 5,184 of which were as a CHICHI nominal stem. The second most ambiguous was *quitamitaxtahuisquía* (HuC) with 3,504 guesses, of which 288 fall within the CHICHI category. In both cases, the input string is long and include at least three occurrences of the character <t>, which the alternation rules used for the texts PuebH and HuC map optionally to two consonants in the orthography of the model (either <t> or <tl>). Therefore, the guesser proposes one guess for each mapping required by the rules, i.e. one guess where a character <t> maps to <t>, and another where the same character maps to <tl>. In consequence, the ambiguity of the guesser increases considerably in comparison with the ambiguity of the core for these two texts, with 165.86 for PuebH and 84.16 for HuC.

The basal ambiguity of the guesser will multiply with each alternation rule, especially if the alternation is optional. However, the low average ambiguity of the core in the case of PuebH (1.12) and HuC (1.15) also show that the same alternation rule does not necessarily cause a dramatic increase in the ambiguity when a known stem is used as basis for the analysis of a string.

A new alternation rule might involve a need to find a balance between precision and recall. A decision to include an alternation rule has to consider whether it is more important to keep ambiguity low (i.e. to increase precision) at the expense of missing analyses (i.e. to decrease recall), or to analyse more input strings at the expense of increasing the ambiguity of the analyses. Besides, one major limitation of this research is the lack of a contemporary GS to verify the effect and suitability of tentative alternation rules. In consequence, this research chose in principle to codify only the less complex mappings (i.e. the ones not requiring a thorough definition of several contexts of occurrence) between the orthography in the texts and the orthography used in our model.

5.2.1.2 Analysis by the core

We have a low percentage of recognition by the core, less than a third for each text. The analysis proposed for the contemporary texts cannot be currently deemed as correct or incorrect. We lack a contemporary GS, and the time to manually check all the output of the model. The analyses proposed by the model for the strings in the contemporary texts are thus considered plausible. Once revised, these morphological analyses could be a GS to evaluate the model and better strategies for analysing contemporary texts.

The core model performs well recognising complex VNCs in different tenses and moods, which sometimes bear long concatenations of affixes. The following are examples of plausible analyses provided by the model from different texts, which in these cases are correct:

ohuajtlajtojquej ⇒ *Augment+Subj3Pl+DirectionHere+IndefObjTla+ˆihtoh+PreteritePl* ***(G)***
'they talked, said things (towards here)' from *ihto-a*

owalmokuepkej ⇒ *Augment+Subj3Pl+DirectionHere+Reflex3Pl+ˆcuep+PreteritePl* ***(OaxN)***
'they came back here' from *cuep-a*

xiclalican ⇒ *Subj2PlOPTATIVE+Obj3Sg+ˆtlali+OptativePl* ***(Mich)***
'sit it, put it' from *tlali-a*

xitechpalehuiqui ⇒ *Subj2SgOPTATIVE+Obj1Pl+ˆpalehui+OptTowardSg* ***(PuebH)***
'come to help me!', from *palehui-a*

monextico ⇒ *Subj3Sg+Reflex3Sg+ˆnexti+PerfTowardSg* ***(HuW)***
'he came to appear (lit. to make himself visible)' from *nexti-a*

The diverging affixes used in the texts to mark a grammatical function can cause analyses that might be incorrect. The best example is the prefix for the object second plural, which in the model is *amech-*. It is reported, for example, that in the Puebla Highlands the prefix for the object second plural is *namech-* (Robinson, 1970). Thus, in the PuebH text the model proposes the following analysis:

> *namechchihuilis* ⇒ *Subj1Sg+Obj2Pl+ˆchihuili+FutSg* ***(PuebH)***
> 'I will do it for you (Pl)'

In this case the model parses <namech> as the reduction of the subject prefix *ni-* and the object prefix *amech-*. However, taking *namech-* as the object second

plural, as Robinson (1970) would do, the correct parsing of *namechchihuilis* would be:

0-*namech-ˆchihuili-s*

thus meaning 'He will do it for you (Pl)' (i.e. *Subj3Sg+Obj2Pl+ˆchihuili+FutSg)*. It is therefore necessary to maintain that the analyses of the model for contemporary texts must be considered plausible, until a revision of all of them allows one to determine if they are correct.

5.2.1.2.1 Ambiguity

The ambiguity generated by the core analysing the contemporary texts was not notably different from the ambiguity found analysing the CN GS. Again the most ambiguous input strings tend to be RNCs using a possessive prefix, e.g. *topan*, 'on us', or *techan*, 'somebody's home'. For all texts, the most ambiguous string was not given more than 9 readings.

The word-type given more analyses by the core was <tepan> in PuebH. Alongside some expected readings (e.g. RNCs around stems pan-tli, pami-tl, pani-tl), the optional mapping of <t> to <tl>, and the unmarked vowel length yield also the following analyses:

tepan ⇒ *ˆtleˆ+Rel_pan*
from *tle-tl*, 'fire'

tepan ⇒ *ˆteˆ+Rel_pan*
from *te-tl*, 'stone'

tepan ⇒ *Temporal_tepan*
'then, after'

tepan ⇒ *Poss1Pl+ˆeˆ+Rel_pan*
from *e-tl*, 'bean'

tepan ⇒ *Poss1Pl+ˆepan+SgPoss*
from *epan-tli* 'three rows', compound stem in Wimmer's dictionary

tepan ⇒ *PossIndet+ˆˆ+Rel_pan* ***(Correct)***
'on somebody'

tepan ⇒ *PossIndet+ˆpani+SgPoss*

'somebody's banner, row'

tepan ⇒ *PossIndet+ˆpan+SgPoss*
'somebody's banner, row'

tepan ⇒ *PossIndet+ˆpami+SgPoss*
'somebody's banner, row'

The addition of alternation rules for each text did increase the ambiguity of analyses by the core. However, the advantage of the core is that it provides analyses around listed stems only, and thus the ambiguity does not increase dramatically, as compared with the ambiguity produced by the guesser.

5.2.1.2.2 Failures of the core

The core fails to analyse NCs which are very different to the NCs encoded in our model. The most extreme examples involve the second person plural object (Obj2Pl) as marked in Mich. In the model a prefix, *amech-*, marks this features, whereas in the Michoacan variety the Obj2Pl is marked with a circumfix, *an- -mitz*, that surrounds the subject prefix (Sischo & Hollenbach, 2015). The core failed to analyse, for example, *annimitzpalehuis* (an-ni-mitz-palehui-s), 'I will help you (Pl)', as the FS networks encodes only the sequence *namechpalehuiz*, i.e. ni-amech-palehui-z.

The core fails also against word-types which appear to be NCs around stems not listed in the lexicon of the model. This is the case of compound stems:

ticyectatequiutijtinemican ⇒ *??* ***(PuebH)***
'ti-c-yectatequiutijtinemi-can' from stem *yectatequiutijtinemi*, may be?

NCs around Spanish stems are expected cases of failures caused by non-listed stems. Examples are: *nolibrojwan*, 'my books' (OaxN), from Spanish *libro*, 'book'; and *quinfuerzajmacac*, 'he gave them strength' (HuE), from Spanish *fuerza*, 'strength, force'.

The core fails to analyse word-types also as a result of missing alternation rules that could be added later. In G text, for example, the model failed to analyse most second person forms of the optative mood. The prefix *xi-* appears as *x-* in Guerrero texts, even before another consonant, for example in *xkalaki*, 'enter, come in'. One has to consider, however, that in Guerrero <x> can mark also a negation, as in *xnikelnamiki* 'I do not remember it' (Mason et al., 2004).

Currently, the alternation rules for Guerrero do not map <x> in a way that would allow one to correctly distinguish between a second person optative form and a negative form, as this mapping involves defining precisely the context of the alternation. The following forms are currently recognised for G:

mitzmacacan ⇒ *Subj3Pl+Obj2Sg+^maca+OptativePl* ***(G)***
'May they give [it to] you (Sg)' from *maca* 'give'

xijistlacatican ⇒ *Subj2PlOPTATIVE+^ihiztlacati+OptativePl* ***(G)***
'Lie' or 'May you constantly lie' from *ihiztlacat-i*, 'lie'

whereas word types like *xquinnotzacan*, 'call them, invite them' will be missed. In principle, the present work chose to miss many forms for the sake of precision. So, although the core might be missing many word-types, at least one can be relatively confident that the analyses it produces are correct in most cases.

5.2.1.3 Analyses by the guesser

The guesser analyses a good percentage of the word-types missed by the core, ranging from 43.65% (HuW) to 70.18% (PuebH) of all the word-types in each text. Many word-types sent to the guesser are long strings, apparently built around compound stems not listed in the model lexicon. In the case of PuebH, for example, nearly half of the word-types sent to the guesser (7,365 of 16,081) are at least fifteen characters long; for HuW, of the 9,977 strings sent to the guesser, 5,860 are at least twelve-characters long. The length of the input strings, combined with the optional alternation rules to map <t> and <tl> seem to be the main cause of the high ambiguity produced by the guesser (**Table 5.6**) for PuebH.

The guesser provides good plausible readings based on the surrounding affixes. For *ximoyectachilican*, 'watch out, take care' (PuebH), the following guess is proposed:

Subj2PlOPTATIVE+Reflex2Pl+^yectlachili+GUESS_Drop2Tr+OptativePl

This guess suggests a hypothetical stem *yectlachili-a*, which seems related to the stems *yec-tli*, 'right, just' and *tlachiyali-a*, 'be prudent', both attested in Wimmer's dictionary. For the input *tiquinhueyichihuaca*, 'may we praise

them, let us praise them (lit. make them great)'(HuW), the guesser does recognise a plausible hypothetical stem *hueyichihu-a* (*hueyi*, 'great', *chihua*, 'make, produce'), although it fails to recognise the correct mood (OptativePl), and incorrectly recognises the affix *–ca* as a Pluperfect singular:

tiquinhueyichihuaca ⇒ *Subj2Sg+Obj3Pl+ˆhueyichihua+GUESS_Drop1TrA+PluperfectSg*
'you had made them great' **(Incorrect)**

Although the guesser covers a good percentage of the text, and make fairly good guesses, its main current shortcoming is the ambiguity generated by missing and optional alternation rules. The ambiguity is further increased by the typically long input strings. The current usability of the guesser output is reduced by the ambiguity of the guesses.

5.2.1.3.1 Failures of the guesser

Besides expected cases (like the VNCs containing the circumfix *an- -mitz* in Mich), the guesser fails to analyse mainly strings containing characters not included in the alphabet of the model. The typical example are NCs built around Spanish words, e.g. *nodios* 'my god', from Spanish *dios* 'god', and *tocuerpo*, 'our body' from Spanish *cuerpo* 'body'. Bare Spanish words, such as *hierro*, 'iron' and *caballo*, 'horse' are not given any guess either. These two types of string are not analysed mainly because they contain the characters <r, b, d>, which are not included in the restrictive alphabet of the model.

The simple inclusion of new symbols in the alphabet of the CN FS model does not seem enough to handle Spanish loans. The analysis of colonial data (Olko & Sullivan, 2013:199-201) shows not only that Nahuatl texts have long contained loanwords, but also that they exhibit a widespread "Nahuatlisation" of foreign terms. Incorporated loans, for example, have undergone phonological, and morphological adaptations. Besides, prolonged contact has resulted in "non-traditional" morphotactics, like the pluralization of inanimate nouns. In addition, Nahuatl varieties have also imported a number of Spanish prepositions and their functions (e.g. *para* 'for' and *hasta* 'until') to the point that the preposition has developed to a lexical category in Nahuatl (Olko et al., 2018:486). All these contact-induced phenomena would need to be addressed by a FS model for contemporary Nahuatl. This would require morphotactics, lexicon elements

and alternation rules not included in the FS for CN, which could, nevertheless, be used as basis for such further developments.

Other strings are not analysed because they require mappings which are not yet defined as alternation rules. It has been highlighted (**section 5.2.1.2.2**) that it is necessary to distinguish <x> as either a negation marker, or a suffix for the second person optative in G. Strings like *xchocacan*, 'may you (Pl) cry', and *xnejli*, 'it is not true' are missed by both the core and the guesser. In this case, the optative plural suffix *–can* could be used in further versions of the model as a clue to confidently map at least word-forms of the second person Optative plural in Guerrero (always marked with *x-*) to optative plurals in the FS transducer of CN (generally marked with *xi-*).

There are mappings for which no clear clue is available. It is not easy to decide, for example, whether <ú> in Mich should map to either <ohua> or <oa> in the orthography of our model. Strings like *nicpujpúa*, 'I wash it', could be mapped to *nicpohpohua*, in our model, and thus suggest the alternation <ú> for <ohua>. However, in *anquijtúa*, 'you (Pl) say it', <ú> should map to <oa>, i.e *anquihtoa*. Unlike what happens with the second person optative in G, there is no evident clue that could help us to distinguish how <ú> should be mapped in Mich. Currently, around 2% of the failures of the model for Mich are partially explained by the absence of a rule to map the character <ú>.

5.3 Discussion: the convergence between the texts and CN, and between each other

To explore the plausible overlap of the contemporary texts with CN and between each other, the present work focused on the output of the core analyser. As shown above, the guesser output can help other explorations, but is currently too ambiguous to be informative, and developing strategies to exploit it efficiently is beyond our current possibilities. The ambiguity of analyses proposed by the core transducer, in contrast, is reduced by the usage of CN known stems in the analysis of an input. Moreover, the performance of the model with the CN GS, and the provision of not including alternation rules that can cause too many inaccurate readings of the contemporary texts, should increase the reliability of the analyses proposed by our core transducer.

Although the percentage of each text analysed by the core is small in every case, it is worth exploring as it constitutes a plausible overlap between the text and CN as it is described in our model. One can conceptualise each text as a set of word-types, and thus look at the different intersections between them based on the word-types analysed by the core.

5.3.1 Intersection of each text with CN

If each text is considered a set of word-types, then those word-types analysed by the core transducer can be considered a plausible intersection between the text and CN. **Figure 5.1** represents the intersection of each text with CN, in terms of the percentage of word-types analysed by the core transducer.

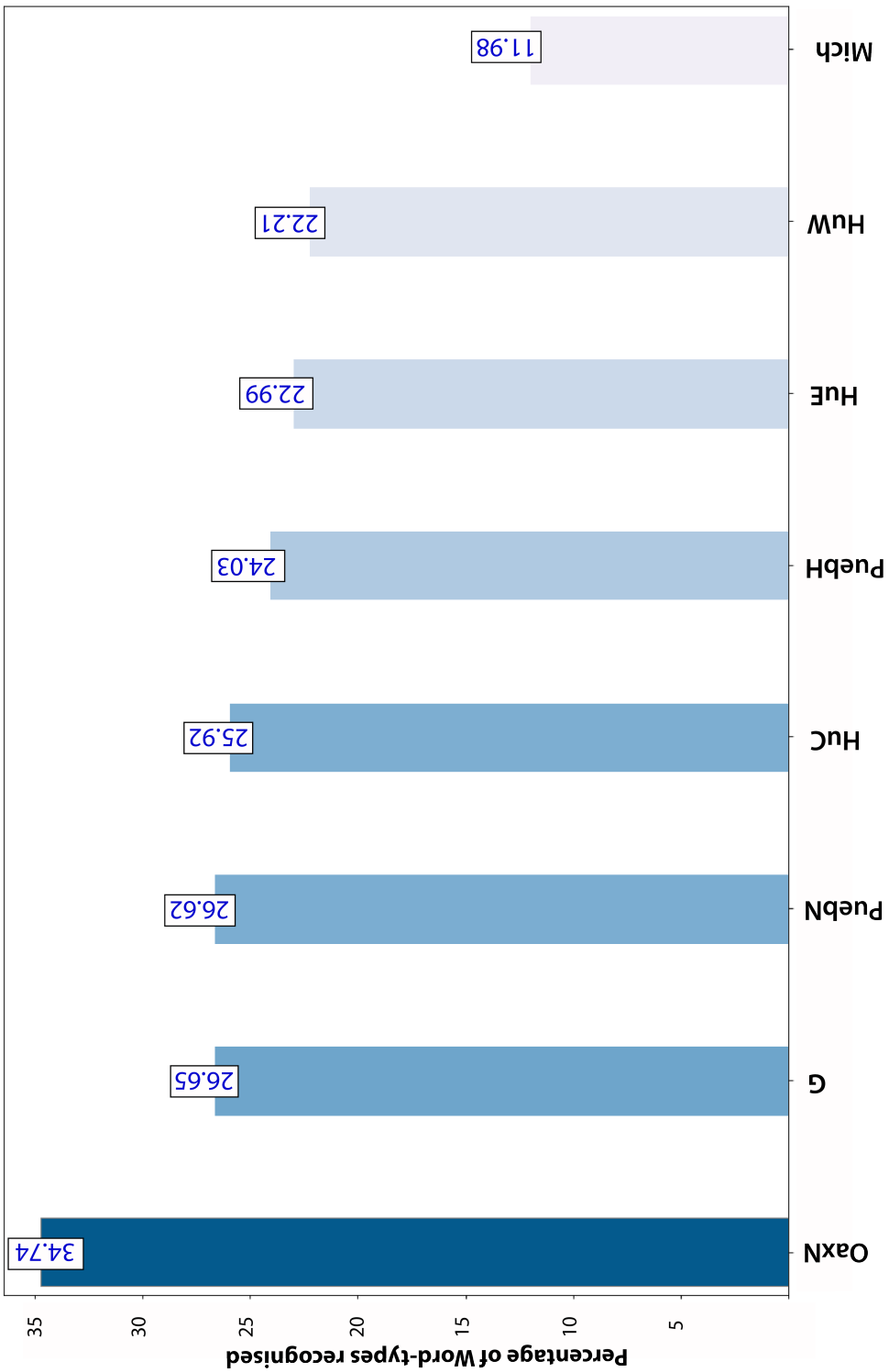


Figure 5.1 Proportional intersection of each text with CN in terms of word-types.

The bars are coloured according to a gradient of blue to further emphasise the size of the intersection of the text with CN. Note that the raw number of word-types recognised would give a slightly different perspective (see **Table 5.5**). PuebH would go up in the rank, as there were 5,425 word-types analysed for this text, whereas for OaxN the core analysed only 4,492. One has to consider, however, that the length of PuebH (21,819 word-types) is considerably higher than the length of OaxN (12,930 word-types). Therefore, the height of the bars and the intensity of the colour in **Figure 5.1** represent the intersection with CN proportionally to the length of each text. From this perspective, OaxN is the text showing the largest intersection with CN, whereas Mich is the one showing the smallest.

Intersections can also be seen in terms of word-forms and paradigmatic slots. In this section we will call *word-form* each of the morphological analyses to which a set of word-types can be plausibly mapped according to our model. All the word-types in **Table 5.7**, which diverge in orthographical terms, are considered one same word-form as they can be plausibly mapped to the same analysis in our model.

Word-types	***Word-form***
quinhuicas **(G)** *quinhuicas* **(HuC)** *quinhuicas* **(HuW)** *quinhuicas* **(HuE)** *kinwikas* **(OaxN)** *quinuicas* **(PuebN)** *quinhuicas* **(Mich)** *quinuicas* **(PuebH)**	*Subj3Sg+Obj3Pl+^huica+PresSg, 'he will bring them'* *quinhuicaz* **(CN model)**

Table 5.7: Word-types and word-forms. All the word-types which could be plausibly mapped to the same analysis by the model are considered one same word-form. The NC for 'he will bring them' is an example of how different word-types in our texts relate to a word-form in the CN model.

Analysed word-types can relate to more than one word-form each. Consider, for example, the word-type <tetat> in the text PuebH, for which the model gives the following analyses:

tetat ⇒ *ˆtleˆ+Rel_tlah*
'place of abundant fire' from *tle-tl*

tetat ⇒ *ˆteˆ+Rel_tlah*
'rocky place' from *te-tl*

tetat ⇒ *Poss1Pl+ˆeˆ+Rel_tlah*

'our place of abundance of beans' from *e-tl*

tetat ⇒ *PossIndet+ˆtlah+SgPoss*
'somebody's uncle' from *tlah-tli*

tetat ⇒ *PossIndet+ˆtah+SgPoss* ***(Correct)***
'somebody's father' form *tah-tli*

Ideally only the correct analyses for each word-type would be used. However, revising all the output of the analyser is currently beyond the possibilities of this work. Moreover, there could be cases of true homographs (*tetat*, however, reads only as '[somebody's] father' in PuebH) for which more than one analysis, i.e. more than one word-form, should be counted as correct. In consequence, to calculate intersections between texts in terms of word-forms all the analyses provided by the core for the analysed word-types were taken into account.

Departing from the sets of all plausible word-forms for each text, the number of what will be called *paradigmatic slots* was determined. By *paradigmatic slot* this work refers to a given combination of affixes that can occur around a generic stem. All word-forms in a text bearing the same combination of affixes are counted as one paradigmatic slot, as the optative second person plural forms from HuC in **Table 5.8** show.

Word-types (HuC)	***Word-forms (HuC)***
xicalaquica	*Subj2P|OPTATIVE+ˆcalaqui+OptativePl 'Enter, come in!'*
xicholoca	*Subj2P|OPTATIVE+ˆcholo+OptativePl 'Flee!'*
xihuetzica	*Subj2P|OPTATIVE+ˆhuetzi+OptativePl 'Fall!'*
xichocaca	*Subj2P|OPTATIVE+ˆchoca+OptativePl 'Cry!'*
Paradigmatic slot	
Subj2P|OPTATIVE+ˆSTEM+OptativePl	

Table 5.8: Word-types, word-forms and their relation to a paradigmatic slot.

In terms of paradigmatic slots shared with CN, the ranking of each text is the same determined using word-types analysed, except for the notable case of PuebH, which surpasses OaxN, as **Figure 5.2** shows. In the figure a gradient of green is used to help us bear in mind the degree of overlap with CN proportional to the length of each text. Thus, although PuebH plausibly shares more paradigmatic slots with CN than any other text, the paler green of PuebH should remind us that this text still ranks fifth in terms of the percentage of word-types analysed by the core.

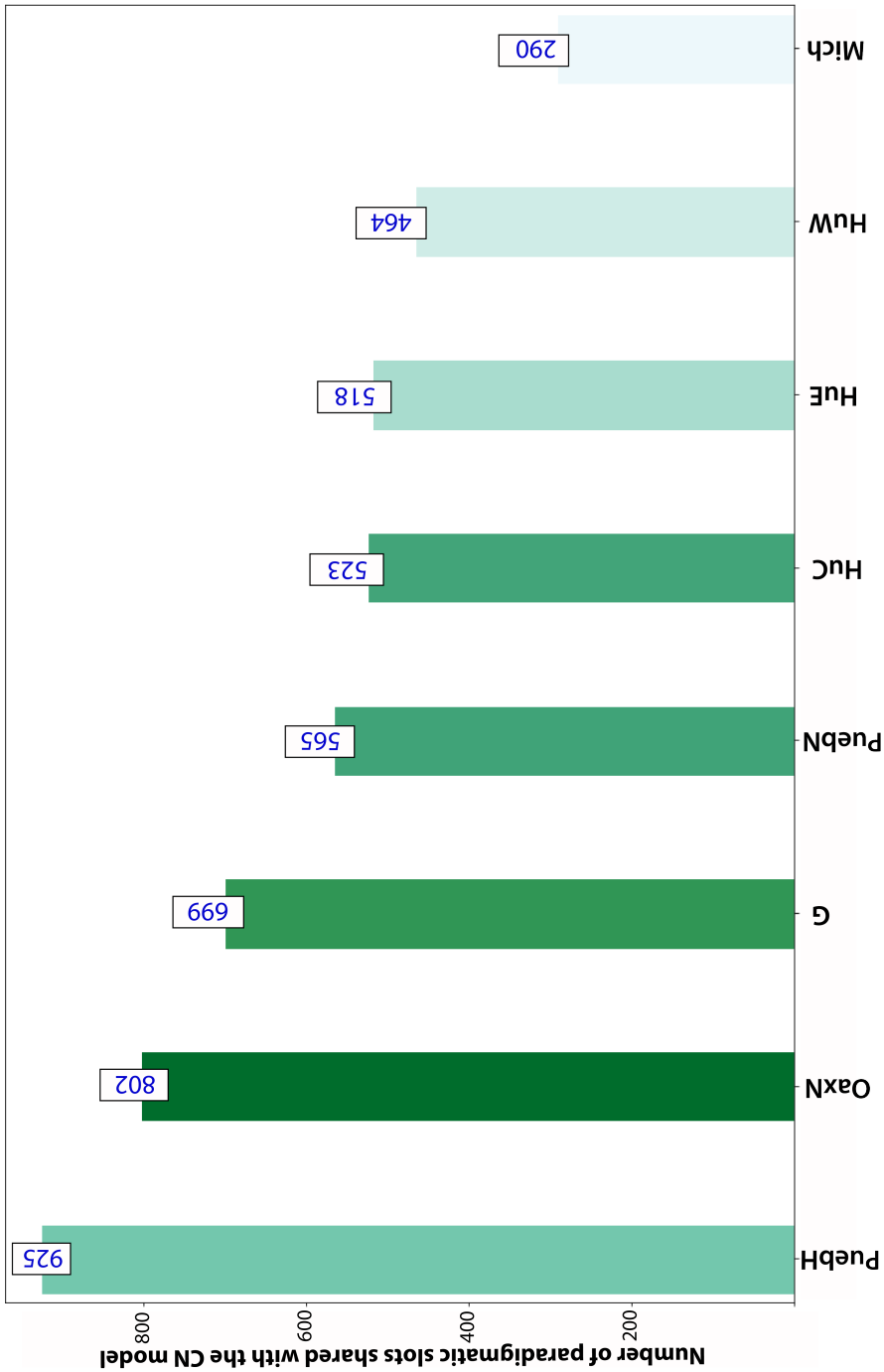


Figure 5.2 Intersection of texts with CN in terms of paradigmatic slots. The gradient of green indicates the degree of overlap with CN proportional to the length of each text: the paler the green, the less percentage of word types of the text were analysed by the CN model

5.3.2 Intersections between texts

The sets of word-forms and paradigmatic slots determined for each text were used to calculate the different intersections between them. The total number of intersection possibilities between a group of eight texts equals to 247. To list them all or represent them graphically becomes complex and space-consuming. Although there are algorithms (Lex et al., 2014:e.g.) to graphically represent intersections of more than four sets or more, here is shown as a manner of example only the size of the intersections between groups of two texts in terms of word-forms (**Figure 5.3**) and paradigmatic slots (**Figure 5.4**).

The potential value of calculating these intersections between pairs of texts is not immediately visible due to the limitations of a simple bar plot. However, each bar in these plots gives us a first idea of the plausible overlap between two texts. Based on the size of their intersection in **Figure 5.3**, it is visible that the three texts from the Huasteca regions (HuC, HuW, HuE), for example, have more word-forms in common with each other than with any other of the texts.

The perspective improves when the intersections are further seen in terms of paradigmatic slots, i.e. in terms of the structural similarity of the NCs recognised. The divergence between the Huasteca group and other texts does not seem as wide as when one takes into account NCs sharing not only the same paradigmatic slot but also the same stem. In **Figure 5.4** the overlap between HuE and PuebH (406) appears almost as large as the overlap between HuW and HuE (408), despite HuE and PuebH having considerably fewer word-forms in common (1,063) than HuW and HuE (2,700) (**Figure 5.3**). This could suggest that although the NCs in HuE and PuebH texts might seem very different by just looking at their common word-forms, HuE text has almost as many paradigmatic slots in common with PuebH as it has with HuW. Looking at intersections between texts in terms of word-forms and paradigmatic slots might thus reveal relations between the texts that are not immediately visible. Moreover, any intersection between the outputs of our core model for each text have also implicit a plausible intersection of the text with CN.

From all the possible intersections between the texts, the one bearing more interest for this work was the intersection of all the eight texts, discussed in the following sections.

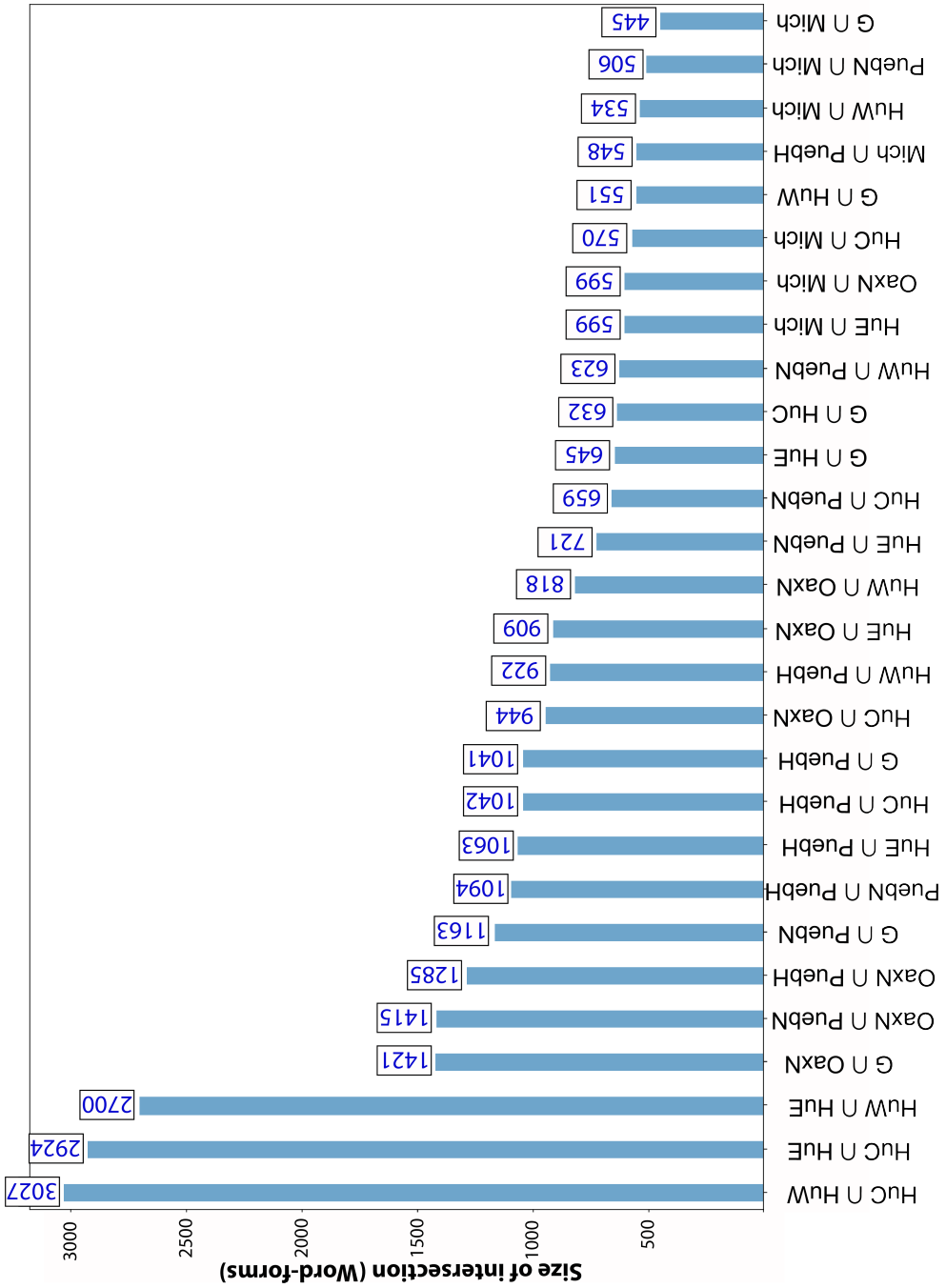


Figure 5.3 Intersections of pair of texts in terms of word-forms.

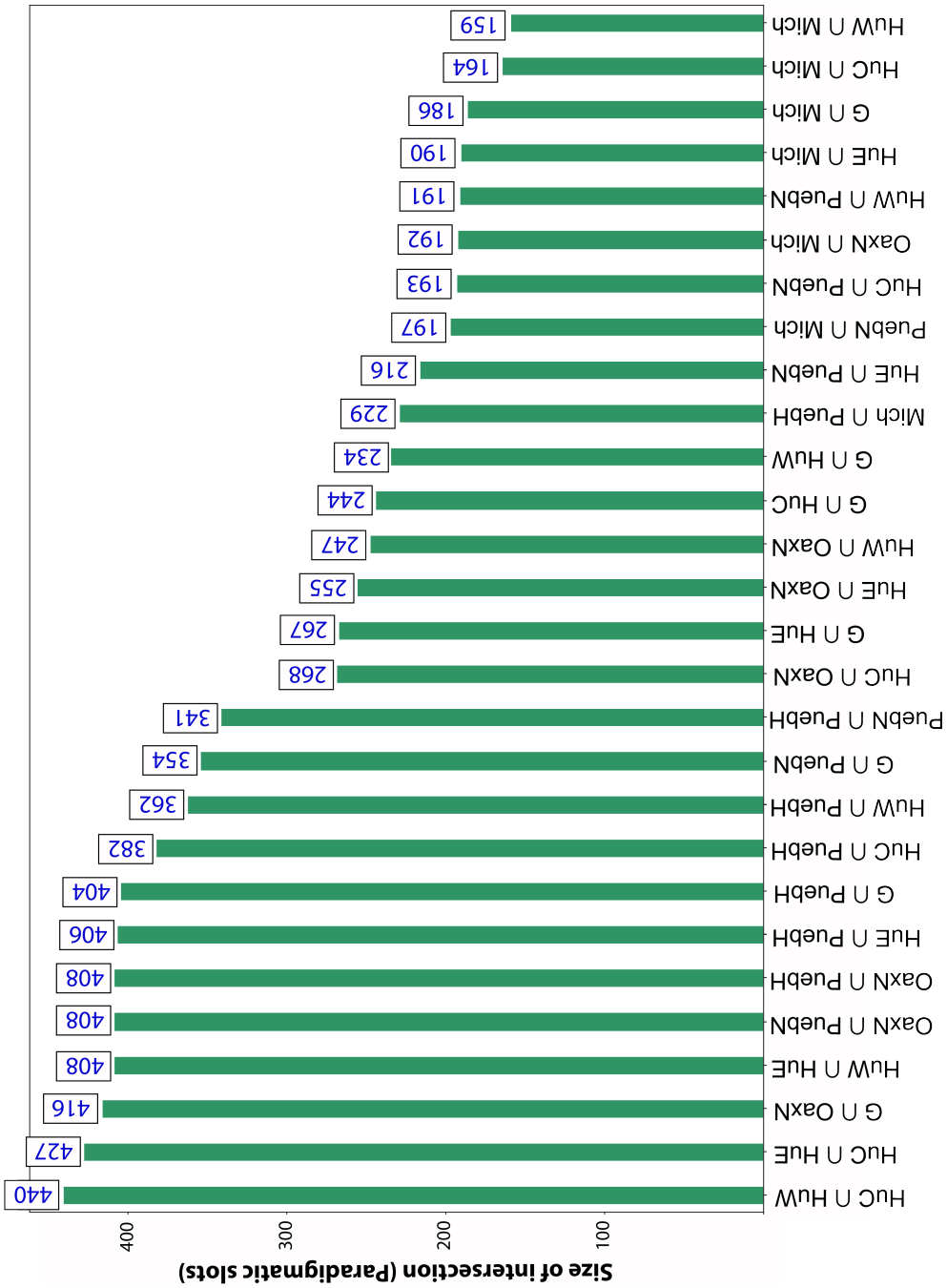


Figure 5.4 Intersections of pair of texts in terms of paradigmatic slots.

5.3.3 Word forms common to all texts

The set of word-forms common to all the texts is estimated as the intersection of all the sets of word-forms. 111 word forms are found under different spellings across all texts, but are still morphologically similar enough to be mapped to the same analysis in CN using relatively simple alternation rules.

Most of the word-forms found across all texts are morphologically simple NCs (none of them bear more than three affixes), built around simple (short) stems, for example:

Subj1Sg+Obj3Sg+ˆitta+PresSg ***(All texts)***
'I see it' *(is this the meaning of this NC in all occurrences in all texts?)*

Subj1Sg+Obj3Pl+ˆtitlani+FutSg ***(All texts)***
'I will send them/I will use them' *(...?)*

Subj3Sg+ˆpehua+PresSg ***(All texts)***
'It begins' *(...?)*

Subj3Sg+Obj1Pl+ˆmaca+FutSg ***(All texts)***
'He/She will give us (it)' *(...?)*

The set of 111 shared word-forms is small taking into account that texts are between 11,000 and 21,000 word-types long. One has to consider, however, that some texts are not as rich as others in terms of the word-forms they contain. The possibility of finding one identical word-form in all the texts might be reduced by one single text not containing a given word-form. For example, the word-form:

Subj3Sg+IndefObjTe+ˆmictia+PresSg
'he kills somebody' (All texts except **Mich**)

is found in all texts, except for Mich. The precise word-form above was not counted as common to all texts. Mich, however, clearly contains similar word-forms built around *micti-a*, for example:

Subj3Sg+Obj2Sg+ˆmictia+PresSg ***(Mich)***
'he kills you(Sg)'

5.3.4 Paradigmatic slots common to all texts

The paradigmatic slots common to all the texts is determined as the intersection of all the sets of paradigmatic slots. 83 paradigmatic slots occurring around different stems are found across all texts. Most of the paradigmatic slots found along all texts are morphologically simple, formed by less than 4 affixes, for example:

Subj1Sg+Obj3Pl+ˆSTEM+FutSg

Subj2Sg+Obj3Pl+ˆSTEM+FutSg

Subj3Sg+Reflex3Sg+ˆSTEM+FutSg

Subj3Sg+Reflex3Sg+ˆSTEM+PresSg

Subj3Sg+Obj1Pl+ˆSTEM+PreteriteSg

Subj3Sg+Obj3Sg+ˆSTEM+PreteriteSg

Subj3Sg+ˆSTEM+PreteriteSg

Subj1Sg+Obj3Sg+ˆSTEM+OptAwaySg

Subj3Sg+ˆSTEM+OptTowardSg

Poss1Sg+ˆSTEM+SgPoss

Poss2Sg+ˆSTEM+SgPoss

Poss3Sg+ˆSTEM+SgPoss

Of the possible paradigmatic slots for VNCs, for example, those used for singular forms were found across all texts using simple alternation rules. In other words, in our texts at least VNCs for the three singular grammatical persons seem structurally similar in the written form to the point of being recognisable with the CN model despite divergent spellings across all the texts.

In contrast, paradigmatic slots containing the prefix for the subject second plural (Subj2Pl) were not common to all texts. The Subj2Pl seems indeed the grammatical feature which tends to be marked more differently across Nahuatl varieties, as the extreme case of the circumfix *an- - mitz* in Mich illustrates.

Some common paradigmatic slots for the present, future and preterite tenses are consistently found in all texts, but again only for singular persons. In the case of the preterite, however, the current tagging style of the model does not mark the class of verb (vowel-dropping or not dropping) around which a paradigmatic

slot for the preterite was found. It cannot be claimed, therefore, that common paradigmatic slots for the preterite singular make VNCs for the preterite singular similar for all classes of verbs. It seems, however, that at least preterite singular VNCs around class NODROP1 of verbal stems (see **Table 4.3**) could be recognisably similar across all texts, judging from the preterite forms around *itt-a*, 'to see':

oquitac ⇒ *Augment+Subj3Sg+Obj3Sg+ˆitta+PreteriteSg* ***(G)***
'He/she saw it'

oquitac ⇒ *Augment+Subj3Sg+Obj3Sg+ˆitta+PreteriteSg* ***(G)***
'He/she saw it'

nechitac ⇒ *Subj3Sg+Obj1Sg+ˆitta+PreteriteSg* ***(HuC)***
'He/she saw me'

nimitzitac ⇒ *Subj1Sg+Obj2Sg+ˆitta+PreteriteSg* ***(HuW)***
'I saw you (Sg)'

tinechitac ⇒ *Subj2Sg+Obj1Sg+ˆitta+PreteriteSg* ***(HuE)***
'You (Sg) saw me'

omitzitak ⇒ *Augment+Subj3Sg+Obj2Sg+ˆitta+PreteriteSg* ***(OaxN)***
'He/she saw you (Sg)'

onimitzitac ⇒ *Augment+Subj1Sg+Obj2Sg+ˆitta+PreteriteSg* ***(PuebN)***
'I saw you (Sg)'

tinechitac ⇒ *Subj2Sg+Obj1Sg+ˆitta+PreteriteSg* ***(PuebH)***
'You (Sg) saw me'

techitac ⇒ *Subj3Sg+Obj1Pl+ˆitta+PreteriteSg* ***(Mich)***
'He/she saw us'

Olko & Sullivan have noted this common formation of the preterite of this class of verbs, illustrating it with the VNC titechittac, 'you (Sg.) saw us'. Such example, they emphasise "could be taken from a colonial manuscript or from everyday speech in any modern variety of Nahuatl" (Olko & Sullivan, 2014:375). The analyses of the texts in the present study thus confirm this observation in written data as well, although also shows that the Augment[4] *o-* is not used by all texts.

[4]The antecessive prefix o- is called *Augment* by Launey (2011), and this last term is the one used in the labels of the model. The appearance of the antecessive prefix in the past form

For the optative mood common paradigmatic slots for the second-person singular and plural were found across all texts using simple alternation rules. A series of circumstances combine to aid the identification of the optative plural in all texts by the model. For example, the suffix for the optative plural *–can* is hardly ever left unmarked, albeit sometimes with the variant *–ca* (HuC, HuW, HuE). Moreover, the co-appearance of *–ca(n)* and the optative prefix for the second person *xi-* was relatively easy to code as an alternation rule for each text in the Huasteca group. Thus, at least one plural optative in each text[5] was correctly analysed by the model. The word-forms in **Table 5.9** show that paradigmatic slots containing the second person plural of the optative have similarly enough orthographic realisations across all the texts.

Examples in **Table 5.9** suggest that despite one divergent spelling calquing the pronunciation of the word (namely, the dropping of the final <n> of *-can* in the Huasteca group), the second plural optative is similar enough across all texts to be mapped to a same morphological analysis in the CN model using relatively simple alternation rules. Below the orthographic divergence, therefore, all the texts seem to converge in the use of very similar paradigmatic slots for the optative plural, and for the singular persons of the present, future and preterite. The tags used currently in the model, however, do not allow for the identification of other classes of verbs for which similar paradigmatic slots of the preterite singular were identified.

was one of the criteria used by Canger (1980) to trace an issoglose (Figure 2.2). According to Olko, its usage was optional also in colonial Nahuatl, most probably depending on meaning (antecessive function) than on regional features, although regional tendencies are also possible (Personal communication, January 2019).

[5]Even in text G where, as it has been noted in **section 5.2.1.2.2**, *x-* might also be an incorporated negation in NCs, the model was able to correctly identify one plural optative in the sentence *ma ca xijistlacatican*, 'May you not repeatedly lie'. The negative of the optative is marked outside the VNC by the invariant sequence *ma ca*.

Word-types	***Word-forms***
xijistlacatican ***(G)***	*Subj2Pl OPTATIVE+^ihiztlacati+OptativePl* *'May you [not] lie repeatedly'*
xijchihuaca ***(HuC)***	*Subj2Pl OPTATIVE+Obj3Sg+^chihua+OptativePl* *'Do it'*
xiquilpica ***(HuW)***	*Subj2Pl OPTATIVE+Obj3Sg+^ilpi+OptativePl* *'Bind / tie him'*
xihuetzica ***(HuE)***	*Subj2Pl OPTATIVE+^huetzi+OptativePl* *'may you (Pl.) [not] fall'*
xikchiwakan ***(OaxN)***	*Subj2Pl OPTATIVE+Obj3Sg+^chihua+OptativePl* *'Do it'*
xicmictican ***(PuebN)***	*Subj2Pl OPTATIVE+Obj3Sg+^micti+OptativePl* *'Kill him'*
xiclasacan ***(Mich)***	*Subj2Pl OPTATIVE+Obj3Sg+^tlaza+OptativePl* *'Throw him [out]'*
xicneltogacan ***(PuebH)***	*Subj2Pl OPTATIVE+Obj3Sg+^neltoca+OptativePl* *'Believe it'*
Paradigmatic slots	
Subj2Pl OPTATIVE+^STEM+OptativePl and *Subj2Pl OPTATIVE+Obj3Sg+^STEM+OptativePl*	

Table 5.9: Examples of word-types from across the texts related to the same paradigmatic slots.

5.4 A graphic representation of connections between texts

The intersections between all the texts are small in comparison to the word-types in each text. Considering that the intersections were estimated using only a percentage of the word-types, namely those that were analysed by the core, the intersections shown above are a small percentage of an already small percentage of each text. The value of this small percentage, however, is that it has already implicitly an overlap with CN as represented by a model which can be expanded and refined.

The representation of intersections as quantities in bar plots alone do not make important connections evident. The quantification of different intersections, however, can be used to graphically represent the connections and distance between texts as a *force atlas diagram* (Jacomy et al., 2014). Force atlas diagrams are a data visualisation technique used to represent networks as groups of nodes that are as close to each other as the number of connections they share. In the case of our texts, each text could be seen as a node in a network of texts, with the distance between them being proportional to the number of connections (i.e. the size of the intersections) between them. The study of Deumert (2002) on the emergence and diffusion of standard Afrikaans shows the explanatory potential of network analysis for studying the development of linguistic conventions. Whereas Deumert focused on networks of cliques, the focus of this research is on texts and the linguistic connections they share.

5.4.1 Connections based on word forms

The connections between all the texts estimated in terms of their shared word forms are represented in **Figure 5.5**. The size and colour intensity of each node depends on the percentage of word-types of the text which were recognised by our core (**Table 5.5**). The sizes of intersections between pairs of sets of word-forms, represented in figure **Figure 5.3**, were used to estimate the graphical distance between nodes. The distance between nodes is inversely proportional to the number of plausible word-forms shared between them: the more shared word-forms, the closer two nodes are. In contrast, the width of the connecting

lines, or edges, is directly proportional to the number of word-forms shared: the more shared word-forms, the broader the edge connecting two nodes.

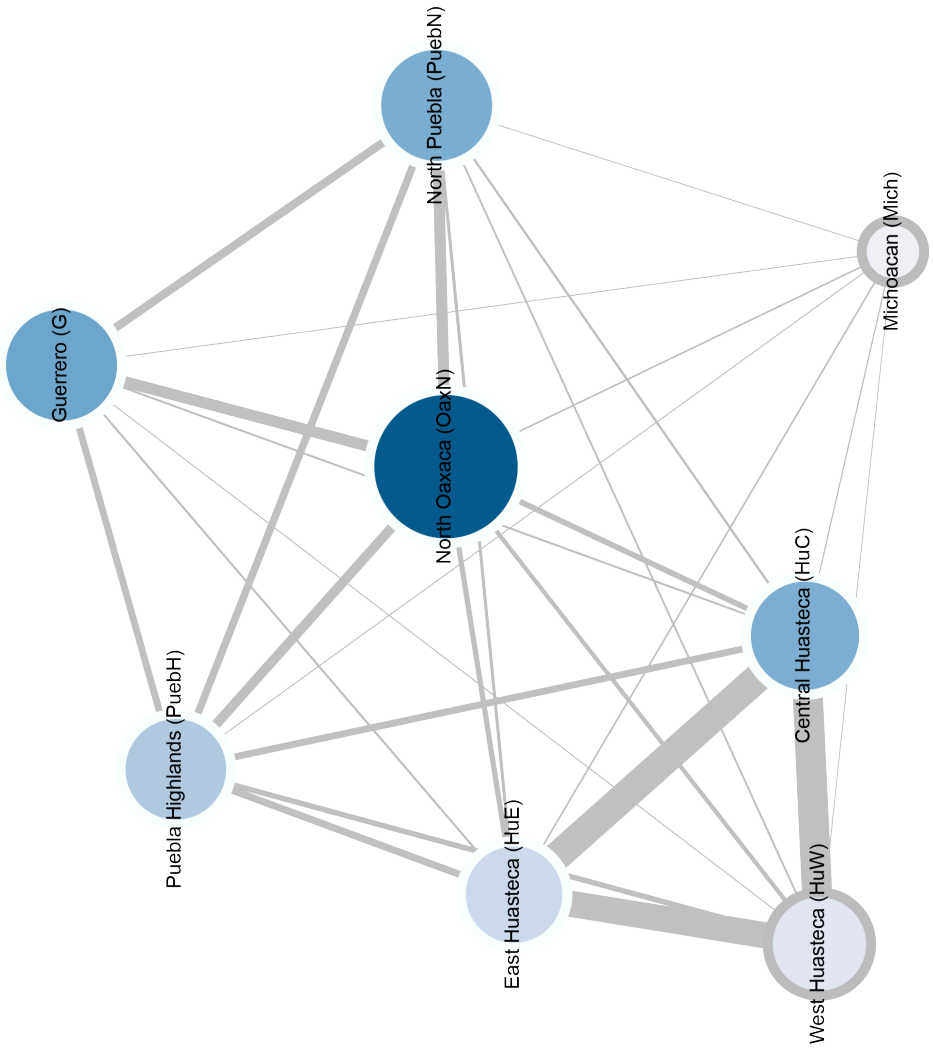


Figure 5.5 Force atlas diagram of the connections between the texts in terms of word forms.

Unlike a bar plot, a force atlas diagram allows us to see that in terms of word-forms OaxN is at the centre of the network formed by all the texts. This 'centrality' depends on the number of shared word-forms with other texts. The arrangement of the nodes in the diagram is not necessarily or exclusively geographical, of course, despite what the close connection between the Huasteca group might suggest. OaxN is put at the centre of the network by the intensity of the 'forces' with which other texts attract each other. In such a network, except for HuW and Mich, all other texts seem roughly equidistant to the centre occupied by OaxN. This suggests that most texts might share with OaxN a roughly similar number of word-forms.

5.4.2 Connections based on paradigmatic slots

The perspective about the 'centrality' of a text in the network changes when connections are made in terms of paradigmatic slots. In **Figure 5.6** the size and colour intensity of each node depends again on the percentage of word-types of the text which were recognised by our core transducer. The graphical distance between nodes in **Figure 5.6**, however, is determined using the number of common paradigmatic slots as connecting criterion **Figure 5.4**. The text from PuebH moves to the centre of the network, despite the percentage of text analysed by the core for this text being proportionally smaller than the percentage analysed for OaxN.

In terms of paradigmatic slots, most texts seem roughly equidistant to PuebH, with the exception of Mich. This suggests that most texts share with PuebH a roughly similar number of paradigmatic slots.

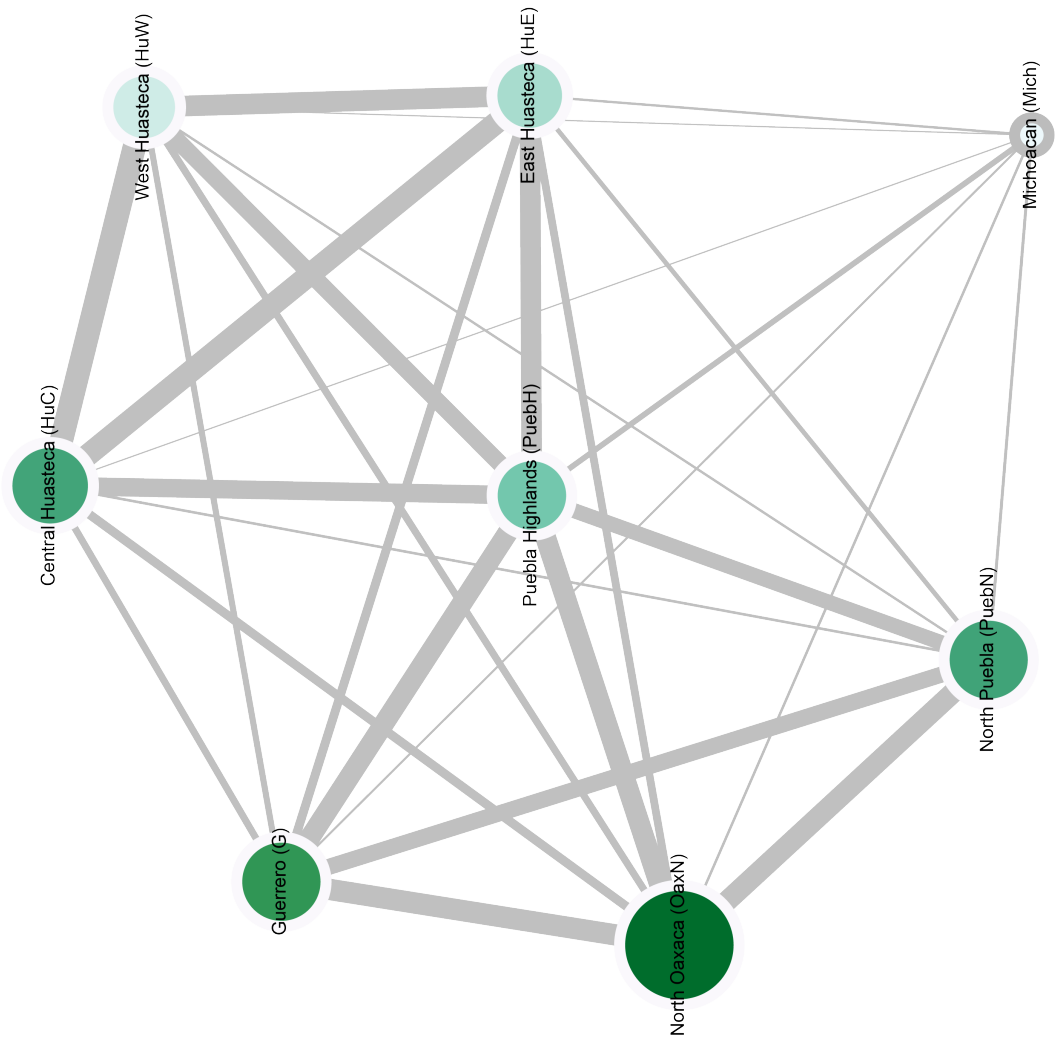


Figure 5.6 Force atlas diagram of the connections between the texts in terms of paradigmatic slots.

One possible explanation for this 'centrality' of PuebH in the network is that PuebH has a richer pool of paradigmatic slots than OaxN. It is true that PuebH has the most ambiguous word-type analysed by the core, but ambiguous word-types do not seem the main reason for the centrality of PuebH in the network. PuebH is also the longest text in terms of word-types and paradigmatic slots. This in turn may cause occurrences of some paradigmatic slots not found in OaxN to be found in PuebH. Take for example, the directional conjugation of motion towards (Launey, 2011:227-9). Word-forms of the perfective (PerfToward) are plausibly found in both OaxN and PuebH, for example:

nicalaquico ⇒ *Subj1Sg+ˆcalaqui+PerfTowardSg* ***(PuebH)***
'I came to enter'

huetzicoj ⇒ *Subj3Pl+ˆhuetzi+PerfTowardPl* ***(PuebH)***
'they came to fall'

owetziko ⇒ *Augment+Subj3Sg+ˆhuetzi+PerfTowardSg* ***(OaxN)***
'he came to fall'

oajsikoj ⇒ *Augment+Subj3Pl+ˆahci+PerfTowardPl* ***(OaxN)***
'they came to reach'

However, only in PuebH did one finds plausible word-forms of the imperfective (ImperfToward), for example:

huetziquiu ⇒ *Subj3Sg+ˆhuetzi+ImperfTowardSg* ***(PuebH)***
'he will come to fall'

nechitaquiu ⇒ *Subj3Sg+Obj1Sg+ˆitta+ImperfTowardSg* ***(PuebH)***
'he will come to see me'

The larger number of word-types in PuebH might cause more diverse types of word-forms to be found in it, and thus this text might constitute a larger pool of more diverse paradigmatic slots. However, if the number of connections for paradigmatic slots really depends primarily on the length of the text, there is a confounding factor here (text length) and perhaps the word-form analysis would be more reliable to locate a 'central' text in our network. Therefore, the OaxN text seems the text that have more plausible connection points to most other texts and CN.

Chapter 6

Conclusions

Language Revitalisation (LR) is a contemporary concern which can be labelled as a type of Language Planning (LP), i.e. the intervention and nurturing of language according to cultural and political ideologies. As other LP enterprises, like the development of a standard in Norway, LR is influenced by political and cultural ideologies. Whereas the influence of national romanticism influenced the standardisation enterprise in Norway, LR seems largely influenced by a commitment to authenticity, manifested in the emphasis on differences between varieties of a continuum, and the desire to preserve this distinctiveness. Assumptions about authenticity among LR actors might also echo a tendency in sociolinguistics to promote the spoken language as the 'real', 'authentic' language.

Progressing towards the strong side of LR would ideally involve an increasing regional, literate usage across a continuum X. When the question of literacy is approached in LR, however, common assumptions about authenticity encourage the idea that the written usage must be 'designed' based on the spoken language. These assumptions, combined with the emphasis on the differences and value of 'traditional' local usages, might hinder the progress towards the strong side of LR, for example by developing a set of orthographies that aim to accommodate to the distinctiveness of the language used in small communities along a linguistic continuum. In addition, a distinctive written codification might sharpen a sense of difference between speakers of local varieties, and a sense of separation from the continuum. A commitment to authenticity thus not only posits a challenge to promoting literacy across a continuum of varieties X, but might also contribute to the reification as separated languages of varieties with few users.

One way to counter the tendency to division over literacy, which is common in LR enterprises, might be to study the overlap between the texts produced by a community of practice of writers along the continuum X. The example of languages like English shows that written conventions can be achieved by a community of practice of writers, and not only nor necessarily carefully designed based on the spoken language. The study of convergence in written texts complements authenticity-based LR with the investigation of an authentication practice occurring despite the lack of initial uniformity. Studying the contemporary written practice to support the development of literacy seems rare in LR, but should be attempted as texts become available.

Nahuatl is a cluster of linguistic varieties including a contemporary dialect continuum and Classical Nahuatl (CN), a corpus of written texts from the sixteenth and seventeenth centuries reflecting mainly the sociolect of the Nahuatl elites in central Mexico at that time.

An exclusive focus of LR on the diversity and value of 'traditional' local usages might not only divide the Nahuatl continuum into separate 'languages', but also divide the potential community of users into 'authentic' and 'unauthentic'. So far, official decisions about how to promote literacy for Nahuatl rely mainly on studies of its spoken diversity. INALI, the official LR promoter for indigenous languages in Mexico, draws upon sociolinguistic and dialectology studies of the indigenous continua to 'design' guidelines for written usage, for example, orthographies. For Nahuatl LR, one pitfall of this strategy is the potential need for designing as many written standards as there are recognised varieties in the Nahuatl continuum. On the other hand, CN seems a valuable asset to LR and a valuable antecedent to support contemporary literacy. Due to a certain sociolinguistic wariness, however, José Antonio Flores Farfán, one influential Nahuatl activist, is cautious of using CN in LR, as comparisons with CN are said to encourage a negative perception of contemporary varieties or of written practices deviating from CN. His position against the formation of a unified Nahuatl written standard aims to preserve and celebrate linguistic diversity. The same activist is wary of an intelligentsia which might use CN as linguistic capital, and overtly favours the language and the interest of the Nahuatl 'common man'. Nahuatl LR approaches too focused on the spoken language of the 'common man' —which reminds of the concerns of national romanticism influencing LP in Norway— somewhat neglect the study of the contemporary written practice, and the contributions and visions of Nahuatl intellectuals to LR. As it happened

with the Norwegian case, a promotion of the 'common' man and his language as only legitimate model of Nahuatl LR could lead Nahuatl activists to dismiss the vision and usage of language of a Nahuatl intelligentsia as not indigenous or local enough; an undesirable result, as LR calls for complementing rather than confronting endeavours.

The practice of contemporary literacy —and its connection with CN— by Nahuatl speakers seems valuable to promote the authentication of new forms of language and community that can bring Nahuatl to the strong side of LR. The view of Olko & Sullivan holds that CN is a valuable asset for LR because it provides examples of previous written usage, and can boost the pride and interest of Nahuatl users in accessing their heritage. Although Olko & Sullivan recognise the value of spoken diversity, they aim to promote a unified Nahuatl orthography, and the formation of Nahua scholars in an academic programme at the Instituto de Docencia e Investigación Etnológica de Zacatecas (IDIEZ). The Universidad Veracruzana Intercultural (UVI) is other promising space where the academic use of Nahuatl is encouraged. Institutions like IDIEZ and UVI are thus producing generations of proud Nahuatl speakers, confident enough to conduct scholarly work not only *on* Nahuatl, but also to write *in* Nahuatl. Although the promotion of a unified orthography might not be accepted by all Nahuatl writers, the practice of writing is important, however heterogeneous, for it might plant the seeds of future conventions; this authentication practice, occurring despite the lack of initial consensus, must be investigated to complement authenticity-based approaches to Nahuatl LR.

The study of convergence in the written practice of Nahuatl must be carried out in a context of ideological and orthographical heterogeneity. This dissertation has tested a methodology to extensively investigate points of convergence between eight contemporary Nahuatl texts and of them with CN. The present work has attempted to locate in these texts the commonalities that might link them with the CN corpus by focusing on the identification of nuclear clauses (NCs): morphosyntactic constructions which seem to be a significant element across the Nahuatl continuum. For this investigation, a FS model of CN has been used to attempt a morphological analysis of the word-types found in eight contemporary text. The word-types that could be plausibly analysed as either NCs or particles by our model, were proposed as points of convergence between the texts and CN.

6.1 Possibility of investigating convergence between texts using a FS approach

Our analysis illustrates that the characteristics of a FS model make it a relevant tool to investigate the overlap between contemporary writing and an older set of written examples. A FS model is not only a compact, efficient codification of an older set of examples (understood as a large set of plausible word-forms according to an available grammatical description); the FS formalism allows for mapping orthographically divergent words to the plausible morphological analysis they plausibly share, without modifying the test text or the core FS model.

A FS model seems a good initial approach to the analysis of Nahuatl texts. As for CN texts, the FS model of CN proved to be faster and more accurate to provide morphological analyses than an available procedural application.

As for contemporary texts, the possibility of composing alternation rules for each text proved to be an advantage of a FS approach. The same FS model provided plausible morphological analyses as CN NCs for word-types in every text, despite the orthographical divergence, and with considerable speed. The analysis of NCs incorporating Spanish loans, however, will require more than the inclusion of new symbols in the alphabet of the model. Incorporated loans have undergone phonological, and morphological adaptations requiring the modelling of morphotactics, lexicon elements and alternation rules not included in the FS for CN.

One main disadvantage of our FS approach to explore convergence is the ambiguity obtained in the analyses. Ambiguity escalates considerably when analyses are performed around hypothetical stems and involve at the same time optional alternation rules. In our study of Nahuatl texts, the use of <t> in PuebH, for example, could map in our CN model to either <t> or to the digraph <tl>. Thus, the ambiguity of analysis for one PuebH word-type rises significantly, and this is most notable in the performance of the guesser. When a potentially recognisable NC is built around a stem not listed in the lexicon of the core module, the effects of optional alternation rules dramatically increase the number of plausible analyses around hypothetical stems (guesses).

High ambiguity is currently the main hindrance to taking advantage of the guesses for a considerable percentage of word-types of each text. Not being able to identify the correct guess for many word-types implies missing many of the paradigmatic slots used in each text, and also many stems not recorded in Wimmer's dictionary of CN. It will be necessary to tackle ambiguity in order to identify many unknown stems attempted as neologisms in the authentication practice of writing across a continuum, in addition to stems used locally, derivatives, and compounds not listed in the model used in the analysis.

Another disadvantage of our approach to studying convergence is the potential complexity of alternation rules required to work with texts from all across a continuum. In the Nahuatl case, linking more NCs across different texts may gradually require many complex alternation rules, as result of the current practice of using writing mostly as a representation of sounds. Even if the texts are written with the same sets of characters, the plausible recognition of an affix, for example, is difficult if the texts represent a NC mostly as a sequence of sounds. The suffix to mark the optative plural of VNCs, for example, has the orthographical realisation <-can> in our model. In the Huasteca group, the same grammatical feature is marked as <-ca>, i.e. as a transcription of the sequence /ka/. Thus, although the Huasteca texts use the same characters that our model to represent the consonant /k/ and the vowel /a/, <-ca> can be recognised as a marker of the optative plural at the cost of either increased ambiguity, or a complex alternation rule requiring the co-occurrence of the prefix for the second person optative plural <xi->. Using the same set of characters across texts from different varieties is only a partial aid to explore convergence if the texts use writing mostly as a representation of sounds.

A FS approach thus will need to tackle high ambiguity and employ complex alternation rules to explore convergence below the heterogeneous practices of writing. This seems to be the cost of being inclusive, and not invasive of what writers might consider authentic or necessary to mark in their texts. Of course, ambiguity might be a concern mainly from the perspective of the automatic analysis of texts, and not necessarily a hindrance for the readers of the texts. It will be useful, therefore, to study whether and how actual readers identify the correct reading of ambiguous word-types occurring in texts from across the continuum.

6.2 Too small a set of points of convergence?

The test texts in our study seem to converge, below their orthographic divergence, in the similar written representation of VNCs of the optative plural; of the singular persons of the present, future; and of the singular persons of the preterite of verbs from the class NODROP1. These are some first plausible points of convergence to emphasise, as their divergent spellings still make relatively evident the stems and paradigmatic slots shared by these set of NCs.

The sets of common word-forms and common paradigmatic slots shared by all texts seems very small, considering only the number of word-types in each text. Each text shares with all other texts a small subset of word-forms or paradigmatic slots, determined from an already small subset of the text (word-types recognised by the core transducer of the CN model).

The plausible points of convergence between all the texts of this exploratory study seem few, but one has to bear in mind that the FS model still needs adjustments to improve its recall and precision for CN texts, and these adjustments will likely have an impact on the analysis of contemporary texts. In addition, more complex alternation rules may also allow to find other plausible points of overlap currently missed. Moreover, our current incapability to disambiguate the analyses produced by the guesser also hinders the identification of other shared paradigmatic slots which plausibly occur around stems not listed in our core transducer. Finally, the degree of convergence between all texts might also reflect the richness of certain texts: some texts contain less diverse word-types than others, and therefore could offer fewer examples of NCs around which plausible convergence could be determined. Mich and PuebH for example, are the texts in our sample with the most word-tokens (254,756 and 275,368 respectively). Mich, however, has fewer word-types (14,028) repeating more, whereas PuebH has more word-types (21,819) repeating less. This might partially explain why PuebH plausibly shares more connections with other texts than Mich. The plausible convergence determined for all the texts in a given set thus depends on a number or variables; these must be addressed to improve the accuracy of subsequent studies, and to determine more points of convergence with increased certainty.

This small overlap between all texts is potentially valuable nonetheless because it has implicit already a plausible overlap with CN as described by our

model. The common word-forms and paradigmatic slots shared by all the texts in a study of convergence are a set of written examples potentially usable across different contemporary varieties of the continuum; equally important, these examples link to the CN corpus. The overlap between texts thus points to vocabulary and paradigmatic slots plausibly shared across the Nahuatl cluster. Further work could study whether and how users from across the continuum understand the overlapping examples in the same way, and how much their interpretation depends on orthographical uniformity.

In the case of the test texts, the overlap with CN is not necessarily an indication of 'puristic' practices reverting to CN as a higher, ideal variety. Arguably, the overall concern of the biblical texts in our study is to get its message across their readers, and not calquing CN or constructing an idealised linguistic model. Due to their objectives, these kinds of texts are likely to employ a style, vocabulary and phonemic orthography closer to the language of 'lay' users, and thus the plausible points of convergence of this texts with CN is more relevant: if a NC in these texts can be plausibly analysed according to a CN description, then common users should not find it too difficult to recognise it in a CN text.

The points of convergence between different texts and the CN corpus might present a diminished version of the Nahuatl cluster. However, this set of commonalities is to be taken only as a departing set of written examples to support the development of literacy across a continuum by linking it with CN, not as an argument to underestimate the divergence between the spoken varieties the texts allegedly represent. The plausible overlaps found in studies like the present one can also serve as point of comparison when determining the convergence in texts of different genres or from more informal practices like texting or blogging in social networks. In more extensive studies, the sets of overlaps between a number texts could be registered as sets of written examples potentially recognisable across diverse localities.

6.3 An additional perspective for LR: seeing overlaps and connections in the written practice

The central perspective of this study was to consider the texts of different varieties as pertaining to a linguistic cluster, and in terms of their overlaps or interconnections, and not only in terms of their differences. Nahuatl users might stand better chances to revitalise their language as a cluster, rather than deconstructing it into a collection of isolated varieties. If Nahuatl written practice is going to be seen as occurring in a linguistic cluster and not only within a contemporary dialect continuum, it is useful to locate points of convergence not only between contemporary texts, but also with the CN corpus.

Estimating the overlaps between different texts allows for their exploration as nodes of an expanding network of written samples connecting language users. From this perspective, written samples from across a continuum can be revalorised as much more than isolated products from isolated localities.

Within its current limitations, the exploration of convergence allowed this study to visualise the plausible connections that exists between the test texts and the CN corpus. The overlap estimated between pairs of texts helped represent them in a network using a force atlas diagram, where the texts are arranged not in geographical or hierarchical terms, but in terms of the connections they have between each other. For the promotion of trans-local literacy, this is an important perspective complementing the delineation of isoglosses. Whereas isoglosses offer a geographical estimation of dialect regions based on spoken data, the force atlas suggests a text or a number of texts as an abstract 'centre', a 'heat zone' containing written examples used in most of the other texts. Studies on convergence can use a force atlas to identify the texts to which many others are connected in terms of NCs, paradigmatic slots, etc., and could therefore constitute a pool of common examples usable across the continuum.

The force atlas representing our test texts in a network indicate that the texts from OaxN and PuebH are the ones with which most texts could have more commonalities. OaxN is the most central one in terms of word-forms, i.e. it contains many word-forms which are also found in most of the other texts.

PuebH is the most central one in terms of paradigmatic slots, meaning that many paradigmatic slots found in most other texts are also found in PuebH.

The 'centrality' of these two texts cannot be used to propose as literacy model the local, traditional varieties they allegedly represent. Most likely, these texts reflect a language difficult to link exactly with a single locality. In addition, if one keeps in mind that writing is an authentication practice, it is possible that neither of the texts contains only traditional vocabulary; the texts contain many neologisms, as suggested by the fails of the core and guesser transducers. Numerous word-types in these texts are NCs containing a long, possibly compound stem, or built around Spanish borrowings. The centrality of a text neither suggests that the text is better, or more purist than others because it is more similar to CN. PuebH, for example, ranks fifth in terms of the percentage of word-types from it that could be analysed by the core transducer –and has thus fewer points in common with our CN model. The relevance PuebH acquires in the analyses of paradigmatic-slots suggests that the largest text in a network could get the most central position. The OaxN might be more interesting because, unlike PuebH, its centrality does not seem an artefact of text length. Moreover, from all the texts, OaxN employs the orthography which appears more divergent from the model of CN —it employs, for example, <k, w> instead of the digraphs <qu, hu, uh>, and <j> for <h>— and is nevertheless the text for which more NCs were plausibly recognised. OaxN seems a potential source of examples of NCs whose morphology make them not only close to CN NCs in many cases, but might be also usable in the written practice along the Nahuatl continuum. Orthographic conventions need not to emerge at the expense of local features, although giving priority to the representation of morphemic structures rather than sounds seems necessary to bring together writers from all the Nahuatl continuum.

6.4 Identifying important texts and not only classifying varieties

The area to which the text OaxN is related illustrates the complexity involved in the classification and delimitation of Nahuatl varieties, as it was discussed in chapter 2. The repopulation of the area, reported by inhabitants of the zone, might partially explain the complex situation in the zone where the sates of

Puebla and Oaxaca meet, and offers an interesting opportunity for dialectological studies.

More importantly for the question of Nahuatl literacy and LR, the complex linguistic situation in the zone related to the OaxN text shows how difficult and arbitrary it can be to differentiate Nahuatl varieties, and the pitfalls of using this differentiation as departing point to design a number of written standards for the Nahuatl continuum. To support the practice of writing for Nahuatl LR, therefore, it is also necessary to locate texts that can provide examples usable across most of the continuum. Nahuatl LR could benefit from considering contemporary writing as an expanding network where certain texts share many connections with most others, and not as a calque of a linguistic usage neatly confined to a specific locality. The importance of certain texts in the whole network of all contemporary texts will likely change as more extensive studies integrate more texts, more plausibly shared paradigmatic slots and stems are identified, and ambiguity in the analysis of NCs is tackled. The interconnections, and the representation of centralities in this work provide a first hint, a first point of comparison for further, more extensive and deeper explorations of commonalities between written samples of the contemporary continuum, comprising more genres, authors, and regions.

The investigation of convergence between samples of written texts offers LR enterprises a way to identify commonalities in written data to build upon in the authentication practice of writing. How relevant or acceptable such commonalities might be for the users of a language seems largely influenced by assumptions about authenticity which –consciously or unconsciously, for good or for bad– guide their individual actions. Maybe no LR project will ever achieve 'ideological clarification'; hopefully, the need and enthusiasm to communicate using a language will eventually be more important than subjective judgements about language usage.

References

Aburto, P. & Mason, D. (2005). Vocabulario comparativo del náhuatl de Guerrero y del náhuatl de Tlamacazapa.

Adelaar, W. F. & Muysken, P. (2004). *The languages of the Andes*. Cambridge: Cambridge University Press.

Adkins, M. (2013). Will the real Breton please stand up?: Language revitalization and the problem of authentic language. *International Journal of the Sociology of Language*, 233, 55–70.

Aduriz, I., Urkia, M., Alegria, I., & others (1997). A spelling corrector for Basque based on morphology. *Literary and Linguistic Computing*, 12(1), 31–38.

Alegria, I., Aranzabe, M., Ezeiza, N., Ezeiza, A., & Urizar, R. (2002). Using Finite State technology in Natural Language Processing for Basque. In B. W. Watson & D. Wood (Eds.), *Implementation and Application of Automata*, volume 2494 of *Lecture Notes in Computer Science* (pp. 1–12). Heidelberg: Springer-Verlag.

Alegria, I., Artola, X., Sarasola, K., & Urkia, M. (1996). Automatic morphological analysis of Basque. *Literary and Linguistic Computing*, 11(4), 193–203.

Amery, R. (2001). Language Planning and Language Revival. *Current Issues in Language Planning*, 2(2-3), 141–221.

Amith, J. D. (2002). What is in a word? The whys and what-fors of a Nahuatl dictionary. In W. Frawley, K. C. Hill, & P. Munro (Eds.), *Making dictionaries: preserving indigenous languages of the Americas* (pp. 219–258). Berkeley, CA: University of California Press.

Anderson, B. (2006). *Imagined communities: reflections on the origin and spread of Nationalism*. London: Verso, revised ed edition.

Andrews, J. R. (1975). *Introduction to Classical Nahuatl*. Austin, Texas: University of Texas Press.

Andrews, J. R. (2003). *Introduction to Classical Nahuatl*. Norman, Oklahoma: University of Oklahoma Press, revised ed edition.

Appadurai, A. (1990). Disjuncture and difference in the global cultural economy. *Theory, Culture and Society*, 7(2), 295–310.

AVELI (2010). *Catálogo de las lenguas indígenas y sus variantes lingüísticas del Estado de Veracruz*. Xalapa: Academia Veracruzana de las Lenguas Indígenas.

Axelson, E. (2013). HFST: using weights.

Ayac (2014). Aesop's Fables at 'Nahuatlahtolli' [Weblog].

Azurmendi, M. J., Bachoc, E., & Zabaleta, F. (2001). Reversing Language Shift: the case of Basque. In *Can threatened languages be saved? : reversing language shift, revisited: a 21st century perspective* (pp. 2234–259). Clevendon: Multilingual Matters.

Beesley, K. R. (2004). Morphological Analysis and Generation: A First-Step in Natural Language Processing. In *SALTMIL Workshop at LREC 2004: First Steps in Language Documentation for Minority Languages* (pp. 1–6). Lisbon.

Beesley, K. R. & Karttunen, L. (2003). *Finite State Morphology*. Stanford, California: Center for the Study of Language and Information.

Beijering, K., Gooskens, C., & Heeringa, W. (2008). Modeling intelligibility and perceived linguistic distances by means of the Levenshtein algorithm. In M. Van Koppen & B. Botma (Eds.), *Linguistics in the Netherlands 2008* (pp. 13–24). Amsterdam: John Benjamins.

Bell, J. (2013). Language attitudes and language revival/survival. *Journal of Multilingual and Multicultural Development*, 34(4), 399–410.

Beller, R. & Beller, P. (1979). Huasteca Nahuatl. In R. W. Langacker (Ed.), *Studies in Uto-Aztecan Grammar vol. 2 : Modern Aztec Grammatical Sketches* (pp. 199–306). Dallas: SIL.

Bermel, N. (2007). *Linguistic Authority, Language Ideology and Metaphor: the Czech Orthography Wars*. Number 17 in Language, Power and Social Process. Berlin: Mouton de Gruyter.

Bowers, D., Arppe, A., Lachler, J., Moshagen, S. N., & Trosterud, T. (2017). A Morphological Parser for Odawa (forthcomming 2017). In *Proceedings of the ComputEL-2: 2nd Workshop on Computational Methods for Endangered Languages* Manoa, Hawai.

Brambila Rojo, O. F. (2004). *Hacia una didáctica del náhuatl: la formación de profesores en la enseñanza de la fonología*. PhD thesis, UNAM.

Brockway, E. (1979). North Puebla Nahuatl. In R. W. Langacker (Ed.), *Studies in Uto-Aztecan Grammar vol. 2 : Modern Aztec Grammatical Sketches* (pp. 141–189). Dallas: Summer Institute of Linguistics.

Bucholtz, M. (2003). Sociolinguistic nostalgia and the authentication of identity. *Journal of Sociolinguistics2*, 7(3), 398–416.

Bull, T. (2005). Special linguistic developments in 19th century Norway. In O. Bandle (Ed.), *The Nordic languages: an international handbook of the history of the North Germanic languages, v. 2* (pp. 1468–1475). Berlin: Walter de Gruyter.

Campbell, L. (1985). *The Pipil language of El Salvador*. Berlin: Mouton.

Canger, U. (1980). *Five Studies inspired by Nahuatl verbs in -oa*. Copenhagen: The Linguistic Circle of Copenhagen.

Canger, U. (1988). Nahuatl Dialectology: A Survey and Some Suggestions. *International Journal of American Linguistics*, 54(1), pp. 28–72.

Canger, U. (1994). A book in an unwritten language. *International Journal of Linguistics*, 27(1), 79–89.

Canger, U. (2002). An interactive dictionary and text corpus for sixteenth and seventeenth century Nahuatl. In W. Frawley, K. C. Hill, & P. Munro (Eds.), *Making dictionaries: preserving indigenous languages of the Americas* (pp. 195–218). Berkeley, CA: University of California Press.

Canger, U. (2011). El nauatl urbano de Tlatelolco/Tenochtitlan, resultado de convergencia entre dialectos. Con un esbozo brevísimo de la historia de los dialectos. *Estudios de Cultura Náhuatl*, 42, 246–258.

Carochi, H. (2001). *Grammar of the Mexican Language with an explanation of its adverbs (1645): edited by James Lockhart*. Stanford, California: Stanford University Press.

Coronel-Molina, S. M. (2008). Language Ideologies of the High Academy of the Quechua Language in Cuzco, Peru. *Latin American and Caribbean Ethnic Studies*, 3(3), 319–340.

Costa, J. & Gasquet-Cyrus, M. (2013). What is language revitalization really about? competing language revitalization movements in Provence. In M. C. Jones & S. Ogilvie (Eds.), *Keeping Languages Alive: Documentation, Pedagogy, and Revitalization* (pp. 212–224). Cambridge: Cambridge University Press.

Coulmas, F. (1989). *The writing systems of the world*. Oxford: Basil Blackwell.

Coulmas, F. (2003). *Writing systems : an introduction to their linguistic analysis*. Cambridge: Cambridge University Press.

Coulmas, F. (2013). *Writing and Society: An Introduction*. Cambridge: Cambridge University Press.

Creutz, M. & Lagus, K. (2005). *Unsupervised morpheme segmentation and morphology induction from text corpora using morfessor 1.0: Technical Report A81.* Technical report, Helsinki University of Technology.

Crowley, T. (2000). The Consequences of Vernacular (II)literacy in the Pacific. *Current Issues in Language Planning*, 1(3), 368–388.

Crowley, T. (2005). Competing Agendas in Indigenous-Language Renewal: Initial Vernacular Education in Vanuatu. *International Journal of the Sociology of Language*, 2005(172), 31–49.

De la Cruz Cruz, V. (2014). La escritura náhuatl y los procesos de su revitalización. *Contributions in New World Archaeology*, 7, 199–210.

De Wolf, P. (2003). *Diccionario español-náhuatl [Spanish-Nahuatl Dictionary].* Mexico: UNAM: Fideicomiso Teixidor: UABCS.

Deumert, A. (2002). Standardization and social networks: the emergence and diffusion of standard Afrikaans. In A. R. Linn & N. McLelland (Eds.), *Standardization: Studies from the Germanic Languages* (pp. 1–25). Amsterdam: John Benjamins.

Deumert, A. (2010). Imbodela zamakhumsha – Reflections on standardization and destandardization. *Multilingua - Journal of Cross-Cultural and Interlanguage Communication*, 29(3-4), 243–264.

Deumert, A. & Lexander, K. V. (2013). Texting Africa: Writing as performance. *Journal of Sociolinguistics*, 17(4), 522–546.

DiGiacomo, S. M. (1999). Language ideological debates in an Olympic city: Barcelona 1992-1996. In J. Blommaert (Ed.), *Language ideological debates* (pp. 105–142). Berlin: Mouton de Gruyter.

Ferguson, C. A. (1959). Diglossia. *Word*, 15(2), 325–340.

Fishman, J. (1990). What is reversing language shift (RLS) and how can it succeed? *Journal of Multilingual and Multicultural Development*, 11(1-2), 5–36.

Fishman, J. (1991). *Reversing language shift: theoretical and empirical foundations of assistance to threatened languages.* Clevendon: Multilingual Matters.

Fishman, J. (2001a). *Can threatened languages be saved? : reversing language shift, revisited : a 21st century perspective.* Clevendon: Multilingual Matters.

Fishman, J. (2001b). From theory to practice (and vice versa): review, reconsideration and reiteration. In J. Fishman (Ed.), *Can threatened languages be saved? : reversing language shift, revisited: a 21st century perspective* (pp. 451–483). Clevendon: Multilingual Matters.

Fishman, J. (2004). Ethnicity and supra-ethnicity in corpus planning: the hidden status agenda in corpus planning. *Nations and Nationalism*, 10, 79–94.

Flores Farfán, J. A. (1999). *Cuatreros Somos y Toindioma Hablamos. Contactos y Conflictos entre el náhuatl y el español en el Sur de México [Contact and conflict between Nahuatl and Spanish in Southern Mexico]*. Mexico: CIESAS.

Flores Farfán, J. A. (2000). Transferencias náhuatl-español en el Balsas (Guerrero, México): reflexiones sobre el desplazamiento y la resistencia lingüística en el náhuatl moderno [Transferences Nahuatl-Spanish: some thoughts on the displacement and linguistic resistance in modern Na. *Amerindia*, 23, 87–106.

Flores Farfán, J. A. (2002). The use of Multimedia and the Arts in Language Revitalization, Maintenance and Development: the case of the Balsas Nahuas of Guerrero, Mexico. In B. J. Burnaby & J. A. Reyhner (Eds.), *Indigenous Languages Across the Community* (pp. 225–236). Flagstaff, Arizona: North Arizona University.

Flores Farfán, J. A. (2003). Efectos del contacto náhuatl-español en la región del Balsas [Effects of the Nahuatl-Spanish contact in the Balsas region]. *Estudios de Cultura Náhuatl*, 34, 331–348.

Flores Farfán, J. A. (2004a). Classical Nahuatl: outlining its sociolinguistic complexity. In T. Stolz (Ed.), *Alte Sprachen: Diversitas Linguarum 8*, Diversitas Linguarum 8 (pp. 167–178). Bochum: Universitätsverlag.

Flores Farfán, J. A. (2004b). Notes on Nahuatl typological change. *Sprachtypologie und Universalienforschung - STUF*, 57(1), 85–97.

Flores Farfán, J. A. (2005). Variations and languages ideologies in Mesoamerican Languages. In R. Muhr (Ed.), *Standard Variation and Languages Ideologies in Different Language Cultures around the World* (pp. 311–330). Frankfurt: Peter Lang.

Flores Farfán, J. A. (2007). Desarrollando buenas prácticas en la revitalización de lenguas [Developing good practices in language revitalization]. In M. Schrader-Kniffki & L. M. García (Eds.), *Romania en interacción: entre historia, contacto y política. Ensayos en homenaje a Klaus Zimmermann* (pp. 675–689). Madrid: Iberoamericana.

Flores Farfán, J. A. (2009). *Variación, ideologías y purismo lingüístico: el caso del mexicano o náhuatl.* Mexico: CIESAS.

Flores Farfán, J. A. (2010a). Hacia una historia sociolingüística mesoamericana: explorando el náhuatl clásico [Towards a mesoamerican sociolinguistic history: exploring classical Nahuatl]. In R. Barriga Villanueva & P. Martín Butragueño

(Eds.), *Historia Sociolingüística de México* (pp. 185–206). Mexico: El Colegio de México.

Flores Farfán, J. A. (2010b). Sociolinguistics in Mexico: defining new agendas. In J. M. Ball (Ed.), *The Routledge Handbook of Sociolinguistics around the World* (pp. 34–41). Oxon: Routledge.

Flores Farfán, J. A. (2011a). Discurso ritual y conversacional en el Nahuatl del Alto Balsas, Guerrero. *Estudios de Cultura Nahuatl*, 42, 267–284.

Flores Farfán, J. A. (2011b). Keeping the fire alive: a decade of language revitalization in Mexico. *International Journal of the Sociology of Language*, 2011(212), 189–209.

Flores Farfán, J. A. (2012). Another look at Nahuatl-Spanish contact morphology. In H. Otsuka, C. Stroht, & A. Urdze (Eds.), *More morphologies: Diversitas Linguarum 35* (pp. 33–52). Bochum: Universitätsverlag.

Flores Farfán, J. A. (2013). El potencial de las artes y los medios audiovisuales en la revitalización lingüística [The potential of arts and audiovisual media in language revitalization]. *RLA. Revista de lingüística teórica y aplicada*, 51(1), 33–52.

Flores Farfán, J. A. (2017). On language regimes in the Americas: Mexicano illustrations. *International Journal of the Sociology of Language*, 246, 59–84.

Francis, N. (1999a). Bilingualism, Writing, and Metalinguistic Awareness: Oral-Literate Interactions between First and Second Languages. *Applied Psycholinguistics*, 20(4), 533–561.

Francis, N. (1999b). Self-Correction Patterns and Metalinguistic Awareness: A Proposed Typology for Studying Text-Processing Strategies of Proficient Readers. *Journal of Research in Reading*, 22(3), 304–310.

Francis, N. (2000a). An examination of written expression in bilingual students' 'non-academic' language: assessment of sense of story structure and interlinguistic transfer. *International Journal of Applied Linguistics*, 10(2), 187–214.

Francis, N. (2000b). Rincones de Lectura Comes to San Isidro: New Contexts for Biliteracy and Language Maintenance. *Language, Culture and Curriculum*, 13(1), 31–50.

Francis, N. & Navarrete Gomez, P. R. (2003). Language Interaction in Nahuatl Discourse: The Influence of Spanish in Child and Adult Narratives. *Language, Culture and Curriculum2*, 16(1), 1–17.

Francis, N. & Nieto Andrade, R. (2006). Stories for Language Revitalization in Nahuatl and Chichimeca. In G. Cantoni (Ed.), *Stabilizing Indigenous Lan-*

guages (pp. 146–154). Flagstaff, Arizona: Northern Arizona University, 2 edition.

Gal, S. & Irvine, J. T. (1995). The boundaries of languages and disciplines: How ideologies construct differences. *Social Research*, 62(4), 967–1001.

Galarza, J. (1992). *In amoxtli in tlacatl: el libro, el hombre: códices y vivencias [Books, men: codex and life stories]*. Mexico: Tava.

Gammelgaard, K. (2002). Approaching the Rise of Spoken Standard Language: The Case of Polish, Czech, and Slovak, 1800-1918. *The Slavonic and East European Review*, 80(4), 601–623.

Grenoble, L. A. & Whaley, L. J. (2006). *Saving languages: an introduction to language revitalization*. Cambridge: Cambridge University Press.

Gutierrez Vasques, X. (2015). Bilingual lexicon extraction for a distant language pair using a small parallel corpus. In *Proceedings of NAACL-HLT 2015: Student Research Workshop* (pp. 154–160). Denver: Association for Computational Linguistics.

Gutierrez-Vasques, X., Sierra, G., & Hernandez Pompa, I. (2016). Axolotl: a Web Accessible Parallel Corpus for Spanish-Nahuatl. In N. C. C. Chair), K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* Paris, France: European Language Resources Association (ELRA).

Hasler, A. (1996). *El náhuatl de Tehuacan-Zongolica*. Mexico: CIESAS.

Hasler Hangert, A. (2001). *Gramática Moderna del náhuatl de Tehuacan – Zongolica*. Universidad Veracruzana.

Haspelmath, M. & Sims, A. D. (2013). *Understanding Morphology*. London: Routledge, 2 edition.

Haugen, E. (1966). Dialect, language, nation. *American Antropologist*, 68(4), 922–935.

Hernández, N. (2015). Garibay, León-Portilla y los escritores nahuas en la revista. *Estudios de Cultura Náhuatl*, 50, 65–73.

Hill, J. H. & Hill, K. C. (1980). Mixed Grammar, Purist Grammar, and Language Attitudes in Modern Nahuatl. *Language in Society*, 9(3), 321–348.

Hill, J. H. & Hill, K. C. (1986). *Speaking Mexicano: dynamics of sycretic language in central Mexico*. Tucson: University of Arizona Press.

Hill, J. H. & Hill, K. C. (2004). Word Order Type Change and the Penetration of Spanish de in Modern Nahuatl. *Sprachtypologie und Universalienforschung - STUF*, 57(1), 23–48.

Hinskens, F., Auer, P., & Kerswill, P. (2005). The study of dialect convergence and divergence: conceptual and methodological considerations. In P. Auer, F. Hinskens, & P. Kerswill (Eds.), *Dialect Change: Convergence and Divergence in European Languages* (pp. 1–48). Cambridge: Cambridge University Press.

Hinton, L. (2001). Language Revitalization: an overview. In L. Hinton & K. Hale (Eds.), *The green book of language revitalization in practice* (pp. 3–18). San Diego, CA: Academic Press.

Hinton, L. (2014). Orthography wars. In M. Cahill & K. Rice (Eds.), *Developing orthographies for unwritten languages* (pp. 139–168). Dallas: SIL International.

Hinton, L. & Ahlers, J. (1999). The Issue of "Authenticity" in California Language Restoration. *Anthropology & Education Quarterly*, 30(1), 56–67.

Hobsbawm, E. J. (2002). *Nations and Nationalism since 1780: programme, myth, reality*. Cambridge: Cambridge University Press, 2nd edition.

Hornberger, N. H. (1995). Five vowels or three? Linguistics and Politics in Quechua Language Planning in Peru. In J. W. Tollefson (Ed.), *Power and Inequality in Language Education*. Cambridge: Cambridge University Press.

Hornberger, N. H. & King, K. (1998). Authenticity and Unification in Quechua Language Planning. *Language, Culture and Curriculum*, 11(3), 390–410.

Hornberger, N. H. & King, K. A. (2001). Reversing Quechua language shift in South America. In J. Fishman (Ed.), *Can threatened languages be saved? : reversing language shift, revisited: a 21st century perspective* (pp. 166–194). Clevendon: Multilingual Matters.

Hornsby, M. (2005). Néo-breton and questions of authenticity. *Estudios de sociolingüística: Linguas, sociedades e culturas*, 6(2), 191–218.

Hornsby, M. & Quentel, G. (2013). Contested varieties and competing authenticities: neologisms in revitalized Breton. *International Journal of the Sociology of Language*, 233, 71–86.

Hroch, M. (2007). National romanticism. In B. Trencsényi & M. Kopeček (Eds.), *Discourses of Collective Identity in Central and Southeast Europe 1770–1945, volume II* (pp. 4–18). Budapest: Central European University Press.

Hualde, J. A. & Zuazo, K. (2007). The Standardization of the Basque Language. *Language Problems and Language Planning*, 31(2), 143–168.

Hulden, M. (2009). Foma: a finite-state compiler and library. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 29–32).

Hulden, M. & Samih, Y. (2012). Conversion of Procedural Morphologies to Finite-State Morphologies: a case of study of Arabic. In *Proceedings of the 10th International Workshop on Finite State Methods and Natural Language Processing* (pp. 70–74). Donostia-San Sebastian.

Iheanetu, O. & Adeyeye, M. (2013). Finite state representation of reduplication processes in Igbo. In *IEEE AFRICON Conference* (pp. 1–6).

INALI (2009). *Programa de revitalización, fortalecimiento y desarrollo de las lenguas indígenas nacionales 2008-2012*. Mexico: INALI.

INALI (2010). *Catálogo de las Lenguas Indígenas Nacionales [Catalogue of the National Indigenous Languages]*. Mexico: INALI.

INALI (2012). *Normalización lingüística: informe de rendición de cuentas 2006-2012 [Linguistic norm-creation: report 2006-2012]*. Mexico: Instituto Nacional de Lenguas Indígenas.

INALI (2015). Las 364 variantes de las lenguas indígenas nacionales, con algún riesgo de desaparecer.

INEGI (2005). *Perfil sociodemográfico de la población hablante de Náhuatl [Sociodemographic profile of Nahuatl-speakers]*. Mexico: INEGI.

INEGI (2010). *Censo de Población y Vivienda 2010*. Technical report, INEGI, Mexico.

Jacomy, M., Venturini, T., Heymann, S., & Bastian, M. (2014). ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi Software. *PLoS ONE*, 9(6), e98679.

Jaffe, A. (2003). Misrecognition Unmasked? Polynomic Language, Expert Statuses and Orthogrpahic Practices in Corsican Schools. *Pragmatics*, 13(3-4), 515–537.

Jahr, E. H. (2014). *Language Planning as a Sociolinguistic Experiment: the Case of Modern Norwegian*. Edinburgh: Edinburgh University Press.

Jany, C. (2010). Orthography design for Chuxnabán Mixe. *Language Documentation and Conservation*, 4, 231–253.

Johnson, S. (2005). *Spelling trouble? : language, ideology and the reform of German orthography*. Clevendon: Multilingual Matters.

Johnson, S. & Braber, N. (2008). *Exploring the German Language*. Cambridge: Cambridge University Press, 2 edition.

Jones, M. C. (1998). *Language obsolescense and revitalization: linguistic change in two sociolinguistically contrasting communities*. Oxford: Oxford University Press.

Jurafsky, D. & Martin, J. H. (2009). *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition.* Upper Saddle River, New Jersey: Pearson-Prentice Hall, 2 edition.

Kaplan, R. B. & Baldauf, R. B. (1997). *Language planning: from practice to theory.* Clevendon: Multilingual Matters.

Karan, E. (2014). Standardization: What's the hurry? In M. Cahill & K. Rice (Eds.), *Developing orthographies for unwritten languages* (pp. 107–138). Dallas: SIL International.

Karttunen, F. (1992). *An analytical dictionary of Nahuatl.* Norman, Oklahoma: The University of Oklahoma Press.

Karttunen, F. & Amsler, R. A. (1983). Computer-assisted compilation of a Nahuatl dictionary. *Computers and the Humanities*, 17(4), 175–184.

Kerswill, P. (2007). Standard and non-standard English. In D. Britain (Ed.), *Language in the British Isles* (pp. 34–51). Cambridge: Cambridge University Press.

Koskenniemi, K. (1983). *Two-level morphology: a general computational model for word-form recognition and production.* PhD thesis, University of Helsinki.

Koskenniemi, K. (1986). Compilation of automata from morphological two-level rules. In F. Karlsson (Ed.), *Papers from the Fifth Scandinavian Conference on Computational Linguistics* (pp. 178–181).

Kroskrity, P. V. (2009). Language Renewal as Sites of Language Ideological Struggle: The Need for Ideological Clarification. In J. Reyhner & L. Lockard (Eds.), *Indigenous Language Revitalization: Encouragement, Guidance and Lessons Learned* (pp. 71–83). Flagstaff, Arizona: Northern Arizona University.

Lagos, C., Espinoza, M., & Rojas, D. (2013). Mapudungun according to its speakers: Mapuche intellectuals and the influence of standard language ideology. *Current Issues in Language Planning*, 14(3-4), 403–418.

Lastra de Suárez, Y. (1986). *Las áreas dialectales del náhuatl moderno [The dialectal areas of modern Nahuatl].* Mexico: UNAM.

Lastra de Suárez, Y. & Horcasitas, F. (1976). El nahuatl en el Distrito Federal. *Anales de Antropología*, 13, 103–136.

Lastra de Suárez, Y. & Horcasitas, F. (1978). El nahuatl en el norte y el occidente del Estado de México. *Anales de Antropología*, 15, 185–250.

Lastra de Suárez, Y. & Horcasitas, F. (1979). El nahuatl en el estado de Tlaxcala. *Anales de Antropología*, 16, 275–323.

Launey, M. (1979). *Introduction à la langue et à la littérature aztèques t. I. (grammaire)*. Paris: L'Harmattan.

Launey, M. (2011). *An introduction to Classical Nahuatl*. Cambridge: Cambridge University Press.

Lehmann, C. (2018). Variation and standardization of Yucatec Maya. *Cuadernos de Lingüística de El Colegio de México*, 5(1), 331–387.

León-Portilla, M. (1986). Yancuic Tlahtolli: Palabra nueva. Una antología de la literatura náhuatl contemporánea. *Estudios de Cultura Náhuatl*, 18, 123–169.

Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., & Pfister, H. (2014). UpSet: Visualization of Intersecting Sets. *IEEE Transactions on Visualization and Computer Graphics*, 20(12), 1983–1992.

Linden, K., Silfverberg, M., & Pirinen, T. (2009). HFST Tools for Morphology - An Efficient Open-Source Package for Construction of Morphological Analyzers. In *Proceedings of the Workshop on Systems and Frameworks for Computational Morphology 2009* Zürich.

Linn, A. R. (2010). Voices from above – voices from below. Who is talking and who is listening in Norwegian language politics? *Current Issues in Language Planning*, 11(2), 114–129.

Linn, A. R. (2013). Vernaculars and the idea of a standard language. In K. Allan (Ed.), *The OxFord handbook of the History of Linguistics* (pp. 379–395). Oxford: Oxford University Press.

Linn, A. R. (2014). Parallel languages in the history of language ideology in Norway and the lesson for Nordic higher education. In A. K. Hultgren, F. Gregersen, & J. Thøgersen (Eds.), *English in Nordic Universities: Ideologies and Practices* (pp. 27–52). Amsterdam: John Benjamins.

Lockhart, J. (2001). *Nahuatl as written: lessons in older written Nahuatl, with copious examples and texts*. Stanford, California: Stanford University Press.

López-Goñi, I. (2003). Ikastola in the twentieth century: an alternative for schooling in the Basque Country. *History of Education*, 32(6), 661–676.

MacSwan, J. (1997). *A Minimalist Approach to Intrasentential Code Switching: Spanish-Nahuatl Bilingualism in Central Mexico (PhD Thesis)*. PhD thesis, University of California, Los Angeles.

Manning, C. D. & Schütze, H. (1999). *Foundations of statistical language processing*. Cambridge, Mass.: The MIT Press.

Marcellesi, J. B., Bulot, T., & Blanchet, P. (2003). *Sociolinguistique: Épistémologie, Langues Régionales, Polynomie*. Paris: L'Harmattan.

Marquis, J. & Sallabank, J. (2013). Speakers and language revitalization: A case study of Guernésiais (Guernsey). In M. C. Jones & S. Ogilvie (Eds.), *Keeping Languages Alive: Documentation, Pedagogy, and Revitalization* (pp. 169–180). Cambridge: Cambridge University Press.

Mason, D., Nelson de Mason, M. E., & Aburto, P. (2004). *Puede hablar el náhuatl: náhuatl de Guerrero y español.* Mexico: SIL, 2 edition.

Matras, Y. (2015). Transnational policy and 'authenticity' discourses on Romani language and identity. *Language in Society*, 44(03), 295–316.

Matthews, P. H. P. H. (1991). *Morphology.* Cambridge: Cambridge University Press.

Maxwell, M. & Amith, J. D. (2005). Language Documentation: The Nahuatl Grammar. In A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing*, volume 3406 of *Lecture Notes in Computer Science* (pp. 474–485). Heidelberg: Springer.

McDonough, K. S. (2010). *Indigenous experience in Mexico: Readings in the Nahua intellectual tradition (PhD Thesis).* PhD thesis, University of Minnesota.

McDonough, K. S. (2014). *The Learned Ones : Nahua Intellectuals in Postconquest Mexico.* University of Arizona Press.

McEnery, T. & Hardie, A. (2012). *Corpus linguistics: method, theory and practice.* Cambridge: Cambridge University Press.

Meza Patiño, I. (2012). *Matinahuatlahtolzalocan: aprendamos el idioma nahuatl.* Mexico: El Ajolote.

Milroy, J. (2001). Language ideologies and the consequences of standardization. *Journal of Sociolinguistics*, 5(4), 530–555.

Milroy, J. (2007). The ideology of the standard language. In C. Llamas, L. Mullany, & P. Stockwell (Eds.), *The Routledge companion to sociolinguistics* (pp. 133–139). London: Routledge.

Milroy, J. (2012). *Authority in language : investigating standard English.* Routledge linguistics classics. London: Routledge, 4th ed. edition.

Milroy, J. & Milroy, L. (1999). *Authority in language: investigating standard English.* London: Routledge, third edition.

Mühlhäusler, P. (1996). *Linguistic ecology.* London: Routledge.

Munro, P. (2014). Breaking rules for orthography development. In M. Cahill & K. Rice (Eds.), *Developing orthographies for unwritten languages* (pp. 169–189). Dallas: SIL International.

Ó Murchadha, N. P. (2016). The Efficacy of Unitary and Polynomic Models of Codification in Minority Language Contexts: Ideological, Pragmatic and Pedagogical Issues in the Codification of Irish. *Journal of Multilingual and Multicultural Development*, 37(2), 199–215.

Olko, J., Borges, R., & Sullivan, J. (2018). Convergence as the driving force of typological change in Nahuatl. *STUF - Language Typology and Universals*, 71(3), 467–507.

Olko, J. & Sullivan, J. (2013). Empire, Colony and Globalization: a Brief History of the Nahuatl Language. *Colloquia Humanistica*, (2), 181–216.

Olko, J. & Sullivan, J. (2014). Toward a Comprehensive Model for Nahuatl Language Research and Revitalization. In H. Leung & others (Eds.), *Proceedings of the Fortieth Annual Meeting of the Berkeley Linguistics Society* (pp. 369–397). Berkeley, CA: Berkeley Linguistics Society.

Olko, J. & Sullivan, J. (2016a). Bridging divides: A proposal for integrating the teaching, research and revitalization of Nahuatl. In V. Ferreira & P. Bouda (Eds.), *Language Documentation and Conservation in Europe* (pp. 159–184). Honolulu: University of Hawaii Press.

Olko, J. & Sullivan, J. (2016b). Bridging Gaps and Empowering Speakers: an Inclusive, Partnership-Based Approach to Nahuatl Research and Revitalization. In *Integral Strategies for Language Revitalization2* (pp. 345–383). Warsaw: University of Warsaw.

O'Rourke, B. (2015). Language Revitalisation Models in Minority Language Contexts: Tensions between Ideologies of Authenticity and Anonymity. *Anthropological Journal of European Cultures*, 24(1), 63–82.

Özerk, K. & Todal, J. (2013). Written Language Shift among Norwegian Youth. *International Electronic Journal of Elementary Education*, 5(3), 285–300.

Pecos, R. & Blum-Martinez, R. (2001). The key to cultural survival: language planning and revitalization in the Pueblo de Cochiti. In L. Hinton & K. Hale (Eds.), *The green book of language revitalization in practice2* (pp. 75–83). San Diego, CA: Academic Press.

Penny, R. (2000). *Variation and change in Spanish*. Cambridge: Cambridge University Press.

Pharao Hansen, M. (2014). Nahuatl Dictionaries Online at "Nahuatl Scholar" [Weblog].

Pharao Hansen, M. (2016). *Nahuatl Nation: Language Revitalization and Indigenous Resurgence in 21st Century Mexico (PhD Thesis)*. PhD thesis, Brown University.

Pirinen, T. & Linden, K. (2010). Finite-State Spell-Checking with Weighted Language and Error Models. Building and Evaluating Spell-Checkers with Wikipedia as Corpus. In *Proceedings of LREC 2010 Workshop on Creation and use of basic lexical resources for less-resourced languages* Malta.

Pirkola, A. (2001). Morphological typology of languages for IR. *Journal of Documentation*, 57(3).

Prado Bernardo, A. (2018). Nahuatl of Northern Oaxaca.

Rastle, K. (2018). The place of morphology in learning to read in English (In press). *Cortex*.

Reyhner, J. & Lockard, L., Eds. (2009). *Language Revitalization: encouragement, guidance and lessons learned*. Flagstaff, Arizona: Northern Arizona University.

Rios, A. (2011). Spell checking an agglutinative language: Quechua. In *Fifth Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics* (pp. 51–55). Poznan.

Rios, A. & Castro Mamani, R. (2014). Morphological Disambiguation and Text Normalization for Southern Quechua Varieties. In M. Zampieri & others (Eds.), *Proceedings of VarDial: Applying NLP Tools to Similar Languages, Varieties and Dialects* (pp. 39–47). Dublin.

Robinson, F. (1970). *Gramática inductiva mexica: náhuatl de la Sierra de Puebla*. Mexico: SIL.

Rolstad, K. (2001). Language Death in Central Mexico: the Decline of Nahuatl and the New Bilingual Maintenance Programs. *The Billingual Review*, 26(1), 3–18.

Sallabank, J. (2010). Standardisation, prescription and polynomie: can Guernsey follow the Corsican model? *Current Issues in Language Planning*, 11(4), 311–330.

Sampson, G. (2015). *Writing systems: a linguistic introduction*. Sheffield: Equinox, 2 edition.

Scannell, K. (2014). Corpus building for minority languages. Online.

Schieffelin, B. B. & Charlier Doucet, R. (1998). The "Real" Haitian Creole: Ideology, Metalinguistics, and Orthographic Choice. In B. B. Schieffelin, K. A. Woolard, & P. V. Kroskrity (Eds.), *Language ideologies : practice and theory* (pp. 285–316). Oxford: Oxford University Press.

Schwartz, S. (2017). Writing Chiwere: Orthography, literacy, and language revitalization. *Language and Communication*, IN PRESS, 1–13.

Sebba, M. (1997). *Contact Languages: pidgins and creoles*. London: Macmillan.

Sebba, M. (2000). Orthography as literacy: how Manx was reduced to writing. In N. Ostler & B. Rudes (Eds.), *Endangered languages and literacy* (pp. 63–70). Bath: The Foundation for endangered languages.

Sebba, M. (2007). *Spelling and society : the culture and politics of orthography around the world*. Cambridge University Press.

Sénéchal, M. & Kearnan, K. (2007). The role of morphology in reading and spelling. *Advances in child development and behavior*, 35, 297–325.

Shaalan, K. & Attia, M. (2012). Handling unknown words in Arabic FST morphology. In *Proceedings of the 10th International Workshop on Finite State Methods and Natural Language Processing* (pp. 20–24). Donostia-San Sebastian: Association for Computational Linguistics.

Shulist, S. (2016). "Graduated authenticity": Multilingualism, revitalization, and identity in the Northwest Amazon. *Language and Communication*, 47, 112–123.

Silverstein, M. (1979). Language structure and linguistic ideology. In P. Clyne, W. Hanks, & C. Hofbauer (Eds.), *The elements: a parasession on linguistic units and levels* (pp. 193–247). Chicago: Chicago Linguistic Society.

Simons, G. F. & Fennig, C. D., Eds. (2018). *Ethnologue: Languages of the World*. Dallas: SIL International, 21st edition.

Sischo, W. R. (1979). Michoacán Nahual. In R. W. Langacker (Ed.), *Studies in Uto-Aztecan Grammar vol. 2 : Modern Aztec Grammatical Sketches* (pp. 307–380). Dallas: Summer Institute of Linguistics.

Sischo, W. R. & Hollenbach, E. d. (2015). Gramática breve del Náhuatl de Michoacán.

Smit, P., Virpioja, S., Grönroos, S.-A., & Kurimo, M. (2014). Morfessor 2.0: Toolkit for statistical morphological segmentation. In *Proceedings of Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 21–24). Gothenburg, Sweden.

Sullivan, J., De la Cruz Cruz, E., & De la Cruz De la Cruz, A. (2016). *Tlahtolxitlauhcayotl, Chicontepec, Veracruz [Dictionary of Chicontepec, Veracruz]*. Warsaw: University of Warsaw: IDIEZ.

Sullivan, T. D. (1992). *Compendio de la gramática náhuatl*. Mexico: UNAM.

Svartvik, J. & Leech, G. (2006). *English: one tongue, many languages*. New York: Palgrave-Macmillan.

Thouvenot, M. (2008a). CEN: Tout en un.

Thouvenot, M. (2008b). Grand Dictionnaire du Nahuatl. http://www.sup-infor.com/navigation.htm.

Thouvenot, M. (2011). La normalización gráfica del Códice Florentino. In P. Máynez & J. R. Romero Galván (Eds.), *Segundo coloquio El universo de Sahagún, pasado y presente, 2008* (pp. 159–176). Mexico: UNAM.

Thouvenot, M. & De Pury, S. (2008). Chachalaca: Analyseur morphologique du nahuatl.

Trudgill, P. & Chambers, J. K. (1991). English dialect grammar. In P. Trudgill & J. K. Chambers (Eds.), *Dialects of English: studies in grammatical variation* (pp. 1–3). London: Longman.

Tsunoda, T. (2005). *Language Endangerment and Language Revitalization.* Berlin: Mouton de Gruyter.

Tuggy, D. (1979). Tetelcingo Nahuatl. In R. W. Langacker (Ed.), *Studies in Uto-Aztecan Grammar vol. 2 : Modern Aztec Grammatical Sketches* (pp. 1–140). Dallas: Summer Institute of Linguistics.

Tuggy, D. (2001). *Lecciones para un curso del Náhuatl Moderno (nawatl de Orizaba o de la Sierra de Zongolica).* Mexico: Summer Institute of Linguistics.

Tulloch, S. (2006). Preserving Dialects of an Endangered Language. *Current Issues in Language Planning*, 7(2-3), 269–286.

Universidad Veracruzana (2015). *Minuta de la tercera sesión ordinaria del consejo consultivo general [Minute of the third ordinary session of the general council].* Technical report, Universidad Veracruzana, Xalapa.

Upward, C. & Davidson, G. (2011). *The history of English spelling.* Oxford: Wiley-Blackwell.

Urla, J., Amorrortu, E., Ortega, A., Goirigolzarri, J., & Uranga, B. (2016). Authenticity and linguistic variety among new speakers of Basque. In *Language Documentation and Conservation in Europe* (pp. 1–12). Honolulu: University of Hawaii Press.

Vandenbussche, W. (2002). Dutch Orthography in Lower, Middle and Upper Class Documents in 19th Century Flanders. In A. R. Linn & N. McLelland (Eds.), *Standardization: Studies from the Germanic Languages*, Current Issues in Linguistic Theory v. 235. Amsterdam: John Benjamins.

Virpioja, S., Turunen, V. T., Spiegler, S., Kohonen, O., & Kurimo, M. (2011). Empirical Comparison of Evaluation Methods for Unsupervised Learning of Morphology. *Traitement Automatique des Langues (TAL)*, 52(2), 45–90.

Wardhaugh, R. (1999). *Proper English: myths and misunderstanding about language.* Oxford: Blackwell.

Watts, R. (1999). The ideology of dialect in Switzerland. In J. Blommaert (Ed.), *Language ideological debates* (pp. 67–104). Berlin: Mouton de Gruyter.

Weinberg, M. & De Korne, H. (2016). Who can speak Lenape in Pennsylvania? Authentication and language learning in an endangered language community of practice. *Language and Communication*, 47, 124–134.

Wilson, J. (2011). Types of dialect accommodation in first-generation contact between adult speakers of mutually intelligible but regionally different varieties. *Multilingua - Journal of Cross-Cultural and Interlanguage Communication*, 30(2), 177–220.

Wimmer, A. (2006). Dictionnaire de la Langue Nahuatl Classique.

Wood, S., Ed. (2016). *Nahuatl Dictionary*. Wired Humanities Project.

Woolard, K. A. (1998). Introduction: language ideology as a field of inquiry. In B. B. Schieffelin, K. A. Woolard, & P. V. Kroskrity (Eds.), *Language ideologies : practice and theory* (pp. 3–47). Oxford: Oxford University Press.

Wright-Carr, D. C. (2007). *Lectura del náhuatl: fundamentos para la traducción de los textos en náhuatl del periodo novohispano*. Mexico: INALI.

Yavorska, G. (2010). The impact of ideologies on the standardization of modern Ukrainian. *International Journal of the Sociology of Language*, (201), 163–197.

Zuckermann, G. & Walsh, M. (2011). Stop, Revive, Survive: Lessons from the Hebrew Revival Applicable to the Reclamation, Maintenance and Empowerment of Aboriginal Languages and Cultures. *Australian Journal of Linguistics*, 31(1), 111–127.