Comparing morphological complexity of Spanish, Otomi and Nahuatl

Ximena Gutierrez-Vasques
Universidad Nacional Autónoma
de México
Mexico City
xim@unam.mx

Victor Mijangos
Universidad Nacional Autónoma
de México
Mexico City
vmijangosc@ciencias.unam.mx

Abstract

We use two small parallel corpora for comparing the morphological complexity of Spanish, Otomi and Nahuatl. These are languages that belong to different linguistic families, the latter are low-resourced. We take into account two quantitative criteria, on one hand the distribution of types over tokens in a corpus, on the other, perplexity and entropy as indicators of word structure predictability. We show that a language can be complex in terms of how many different morphological word forms can produce, however, it may be less complex in terms of predictability of its internal structure of words.

1 Introduction

Morphology deals with the internal structure of words (Aronoff and Fudeman, 2011; Haspelmath and Sims, 2013). Languages of the world have different word production processes. Morphological richness vary from language to language, depending on their linguistic typology. In natural language processing (NLP), taking into account the morphological complexity inherent to each language could be important for improving or adapting the existing methods, since the amount of semantic and grammatical information encoded at the word level, may vary significantly from language to language.

Conceptualizing and quantifying linguistic complexity is not an easy task, many quantitative and qualitative dimensions must be taken into account (Miestamo, 2008). On one hand we can try to answer what is complexity in a language and which mechanisms express it, on the other hand, we can try to find out if there is a language with more complex phenomena (phonological, morphological, syntactical) than other and how can we measure it. Miestamo (2008) distinguishes between two types of complexity: the absolute, which defines complexity in terms of the number of parts of a system; and the relative, which is related to the cost and difficulty faced by language users. Some authors focuses in the absolute approach since it is less subjective. Another common complexity distinction is between global and particular. Global complexity characterizes entire languages, e.g., as easy or difficult to learn (Miestamo, 2008, p. 29), while particular complexity refers only to a level of the whole language (for example phonological complexity, morphological complexity, syntactical complexity).

We focus on morphological complexity. Many definitions of this term have been proposed (Baerman et al., 2015; Anderson, 2015; Sampson et al., 2009). From the computational linguistics perspective there has been a special interest in corpus based approaches to quantify it, i.e., methods that estimate the morphological complexity of a language directly from the production of morphological instances over a corpus. This type of approach usually represents a relatively easy and reproducible way to quantify complexity without the strict need of linguistic annotated data. The underlying intuition of corpus based methods is that morphological complexity depends on the morphological system of a language, like its inflectional and derivational processes. A very productive system will produce a lot of different word forms. This morphological richness can be captured with several statistical measures, e.g., information theory measures (Blevins, 2013) or type token relationships. For example, Bybee (2010, p. 9) affirms that "the token frequency of certain items in constructions [i.e., words] as well as the range of types [...] determines representation of the construction as well as its productivity".

Proceedings of the Workshop on Linguistic Complexity and Natural Language Processing, pages 30–37
Santa Fe, New Mexico, USA, August 25, 2018.

In this work, we are interested in using corpus based approaches; however, we would like to quantify the complexity not only by the type and token distributions over a corpus, but also by taking into account other important dimension: the predictability of a morph sequence (Montermini and Bonami, 2013). This is a preliminary work that takes as a case of study the distant languages Otomi, Nahuatl and Spanish. The general idea is to use parallel corpora, type-token relationship and some NLP strategies for measuring the predictability in statistical language models.

Additionally, most of the previous works do not analyze how the complexity changes when different types of morphological normalization procedures are applied to a language, e.g., lemmatization, stemming, morphological segmentation. This information could be useful for linguistic analysis and for measuring the impact of different word form normalization tools depending of the language. In this work, we analyze how the type-token relationship changes using different types of morphological normalization techniques.

1.1 The type-token relationship (TTR)

The type-token relationship (TTR) is the relationship that exists between the number of distinct words (types) and the total word count (tokens) within a text. This measure has been used for several purposes, e.g., as an indicator of vocabulary richness and style of an author (Herdan, 1966; Stamatatos, 2009), information flow of a text (Altmann and Altmann, 2008) and it has also been used in child language acquisition, psychiatry and literary studies (Malvern and Richards, 2002; Kao and Jurafsky, 2012).

TTR has proven to be a simple, yet effective, way to quantify the morphological complexity of a language. This is why it has been used to estimate morphological complexity using relatively small corpora (Kettunen, 2014). It has also shown a high correlation with other types of complexity measures like entropy and paradigm-based approaches that are based on typological information databases (Bentz et al., 2016)

It is important to notice that the value of TTR is affected by the type and length of the texts. However, one natural way to make TTRs comparable between languages is to use a parallel corpus, since the same meaning and functions are, more or less, expressed in the two languages. When TTR is measured over a parallel corpus, it provides a useful way to compare typological and morphological characteristics of languages. Kelih (2010) works with parallel texts of the Slavic language family to analyze morphological and typological features of the languages, i.e., he uses TTR for comparing the morphological productivity and the degree of syntheticity and analycity between the languages. Along the same line, Mayer et al. (2014) automatically extract typological features of the languages, e.g., morphological synthesis degree, by using TTR.

There exist several models that have been developed to examine the relationship between the types and tokens within a text (Mitchell, 2015). The most common one is the ratio $\frac{types}{tokens}$ and it is the one that we use in this work.

1.2 Entropy and Perplexity

In NLP, statistical language models are a useful tool for calculating the probability of any sequence of words in a language. These models need a corpus as training data, they are usually based on n-grams, and more recently, in neural representations of words.

Information theory based measures can be used to estimate the predictiveness of these models, i.e., perplexity and entropy. Perplexity is a common measure for the complexity of n-grams models in NLP (Brown et al., 1992). Perplexity is based in Shannon's entropy (Shannon et al., 1951) as the perplexity of a model μ is defined by the equation $2^{H(\mu)}$, where $H(\mu)$ es the entropy of the model (or random variable). Shannon's entropy had been used for measuring complexity of different systems. In linguistics, entropy is commonly used to measure the complexity of morphological systems (Blevins, 2013; Ackerman and Malouf, 2013; Baerman, 2012). Higher values of perplexity and entropy mean less predictability.

Perplexity depends on how the model is represented (this includes the size of the data). In this work, we compare two different models for calculating the entropy and perplexity: a typical bigram model

adapted to a morph level(Brown et al., 1992); and our proposal based on using the word as a context instead of ngrams.

We rely in parallel corpora to compare the measures across languages, since the same meaning and functions are shared in the two languages.

Bigram model. This model takes into consideration bigrams (Brown et al., 1992) as context for determining the joint probabilities of the sub-strings. Here the bigrams are sequences of two morphs in the text (whether they belong to the same word or not). This is a typical statistical language model but instead of using sequences of words, we use morphological segmented texts. In addition, we use a Laplacian (or add one) smoothing for the conditional probabilities (Chen and Goodman, 1999).

Word level. The word level representation takes the whole word as context for the determination of joint probabilities. Therefore, the frequency of co-occurrence is different from zero only if the sub-word units (morphs) are part of the same word. For example, if xby is a word with a prefix x and a suffix y, the co-occurrence of x with b will be different from zero as both morphs are part of the word xby. Similarly, the co-occurrence of y with b will be different from zero. Conversely, if two morphs are sub-strings of different words, its co-occurrence will be zero. To calculate the conditional probabilities we use and add one estimator defined as:

$$p(x|y) = \frac{fr(x,y)+1}{fr(x,y)+V} \tag{1}$$

Where V is the number of types and $fr(\cdot)$ is the frequency of co-occurrence function.

2 Experimental setting

2.1 The corpus

We work with two language pairs that are spoken in the same country (Mexico) but they are typologically distant languages: Spanish (Indo-European)-Nahuatl (Uto-Aztecan) and Spanish-Otomi (Oto-Manguean). Both, Nahuatl and Otomi are low-resource languages that face scarcity of digital parallel and monolingual corpora.

Nahuatl is an indigenous language with agglutinative and polysynthethic morphological phenomena. It can agglutinate many different prefixes and suffixes to build complex words. Spanish also has rich morphology, but it mainly uses suffixes and it can have a fusional behavior, where morphemes can be fused or overlaid into a single one that encodes several grammatical meanings. Regarding to Otomi, its morphology also has a fusional tendency, and it is head-marking. Otomi morphology is usually considered quite complex (Palancar, 2012) as it exhibits different phenomena like stem alternation, inflectional class changes and suprasegmental variation, just to mention some.

Since we are dealing with low resource languages that have a lot of dialectal and orthographic variation, it is difficult to obtain a standard big parallel corpus. We work with two different parallel corpora, i.e., Spanish-Nahuatl and Spanish-Otomi. Therefore the complexity comparisons are always in reference to Spanish.

We used a Spanish-Nahuatl parallel corpus created by Gutierrez-Vasques et al. (2016). However, we used only a subset since the whole corpus is not homogeneous, i.e., it comprises several Nahuatl dialects, sources, periods of time and it lacks of a general orthographic normalization. We chose the texts that had a more or less systematic writing. On the other hand, we used a Spanish-Otomi parallel corpus (Lastra, 1992) conformed by 38 texts transcribed from speech. This corpus was obtained in San Andrés Cuexcontitlan. It is principally composed by narrative texts, but also counts with dialogues and elicited data. Table 1 shows the size of the parallel corpora used for the experiments.

2.2 Morphological analysis tools

We used different morphological analysis tools, in order to explore the morphological complexity variation among languages and between the different types of morphological representations. We performed lemmatization for Spanish language, and morphological segmentation for all languages.

Parallel Corpus	Tokens	Types
Spanish-Nahuatl		
Spanish (ES)	118364	13233
Nahuatl (NA)	81850	21207
Spanish-Otomi		
Spanish (ES)	8267	2516
Otomi (OT)	6791	3381

Table 1: Size of the parallel corpus

In NLP, morphology is usually tackled by building morphological analysis (taggers) tools. And more commonly, lemmatization and stemming methods are used to reduce the morphological variation by converting words forms to a standard form, i.e., a lemma or a stem. However, most of these technologies are focused in a reduced set of languages. For languages like English, with plenty of resources and relatively poor morphology, morphological processing may be considered solved.

However, this is not the case for all the languages. Specially for languages with rich morphological phenomena where it is not enough to remove inflectional endings in order to obtain a stem.

Lemmatization and stemming aim to remove inflectional endings. Spanish has available tools to perform this task. We used the tool Freeling[1]. Regarding to morphological segmentation, we used semi-supervised statistical segmentation models obtained with the tool Morfessor (Virpioja et al., 2013). In particular, we used the same segmentation models reported in Gutierrez-Vasques (2017) for Spanish and Nahuatl. As for Otomi, we used manual morphological segmentation of the corpus, provided by a specialist.

2.3 Complexity measures

We calculated the type-token relationship for every language in each parallel corpus. Table 2 shows the TTR of the texts without any processing (ES, NA) and with the different types of morphological processing: morphological segmentation (ES_{morph}, NA_{morph}), lemmatization (ES_{lemma}). In a similar way, Table 3 shows the TTR values for the Spanish-Otomi corpus. It is worth mentioning that the TTR values are only comparable within the same parallel corpus.

	Tokens	Types	TTR (%)
ES	118364	13233	11.17
NA	81850	21207	**25.90**
ES_{morph}	189888	4369	2.30
NA_{morph}	175744	2191	**1.24**
ES_{lemma}	118364	7599	6.42

Table 2: TTR for Nahuatl-Spanish corpus

	Tokens	Types	TTR (%)
ES	8267	2516	30.43
OT	6791	3381	**49.78**
ES_{morph}	14422	1072	7.43
OT_{morph}	13895	1788	**1.28**
ES_{lemma}	8502	1020	8.33

Table 3: TTR for Otomi-Spanish corpus

We also calculate the perplexity and complexity for the different languages. Since we are focusing on morphological complexity, we took only the segmented data for computing the entropy and the perplexity. We do not use the lemmatized or non segmented data since this would be equivalent to measuring the combinatorial complexity between words, i.e. syntax. In this sense, the entropy and

[1] http://nlp.lsi.upc.edu/freeling/

perplexity reflects the predictability of the morphs sequences. Tables 4 and 5 shows the perplexity and entropy in each language pair.

	Word level	**Bigram model**
	ES-NA	
NA_{morph}	214.166	1069.973
ES_{morph}	1222.956	2089.774
	ES-OT	
ES_{morph}	208.582	855.1766
OT_{morph}	473.830	1315.006

Table 4: Perplexity obtained in the different parallel corpora

	Word level	**Bigram model**
	ES-NA	
NA_{morph}	0.697	0.906
ES_{morph}	0.848	0.911
	ES-OT	
ES_{morph}	0.765	0.967
OT_{morph}	0.843	0.984

Table 5: Entropy obtained in the different parallel corpora

3 Results analysis

3.1 TTR as a measure of morphological complexity

When no morphological processing is applied, Nahuatl has a lot higher TTR value than Spanish, i.e., a greater proportion of different word forms (types). In spite of Nahuatl having fewer tokens because of its agglutinative nature, it has a lot more types than Spanish. This suggests that Nahuatl has a highly productive system that can generate a great number of different morphological forms. In other words, it is more likely to find a repeated word in Spanish than in a Nahuatl corpus. In the case of Otomi-Spanish, Otomi also has a bigger complexity compared to Spanish in terms of TTR. Even though both Otomi and Spanish show fusional patterns in its inflection, Otomi also count with a lot of derivational processes and shows regular stem alternations.

In every case, morphological segmentation induced the smallest values of TTR for all languages. Suggesting that greater reduction of the morphological complexity is achieved when the words are split into morphs, making it more likely to find a repeated item. For instance, when Nahuatl was morphologically segmented, TTR had a dramatic decrease (from 26.22 to 1.23). This TTR reduction could be the result of eliminating the combinatorial variety of the agglutinative and polysynthetical morphology of the language. Therefore, when we segment the text we break this agglutination, leading to significantly less diverse units.

In the case of Otomi language, a similar trend can be observed. Otomi seems to be morphologically more complex than Spanish in terms of TTR, i.e., more diverse types or word forms. When morphological segmentation is applied, TTR decreases and Otomi language has a lower TTR compared to Spanish. Even though Otomi is not a polysynthetic language like Nahuatl, these results suggest that Otomi has also a great combinatory potential of its morphs, i.e, when Otomi gets morphologically segmented we obtain less diverse types, these morphs may be recurrent in the text but they can be combined in many several ways within the Otomi word structure. Linguistic studies have shown that Otomi language can concatenate several affixes, specially in derivative processes (Lastra, 1992).

It has brought to our attention that Spanish has a higher TTR than Nahuatl and Otomi, only when the languages are morphologically segmented. It seems that the morphs inventory is bigger in Spanish, we conjecture this is related to the fact that Spanish has more suppletion or "irregular" forms phenomena (Boyé and Hofherr, 2006).

3.2 Predictability

The predictability of the internal structure of word is other dimension of complexity. It reflects the difficulty of producing novel words given a set of lexical items (stems, suffixes or morphs). First of all, as a general overview, we can see that word level models have the lower perplexity and entropy (Tables 4 and 5). We believe that this type of models capture better the morphological structure, since they take into account the possible combinations of morphs within a word and not outside the bounds of it (like the bigram model).

It is interesting to compare the TTR and the predictability measures for each language. In the case of Nahuatl, TTR shows that there is a lot of complexity at lexical level (many different word forms, few repetitions), however, this contrasts with the predictability of the elements that conform a lexical item: the combination of morphs within a word is more predictable than Spanish, since it obtains lower values of Perplexity and entropy. The combinatorial structure of Nahuatl morphology shows less uncertainty than Spanish one, despite the fact that Nahuatl is capable of producing many more different types in the corpus due to its agglutinative and polysynthetic nature.

The case of Otomi language is different, since it seems that it is not only complex in terms of TTR but also in terms of predictability. It obtains higher entropy and perplexity than Spanish. We conjecture this is related to several phenomena. For instance, Otomi and Nahuatl allow a large number of morphs combinations to modify a stem (inflectional and derivational). However, Otomi shows phenomena that is not easy to predict; for example, it has a complex system of inflectional classes, stem alternations and prefix changes. Moreover, tones and prosody plays an important role in the morphology of Otomi verbs (Palancar, 2004; Palancar, 2016). Also, we mentioned before that many of the affixes concatenations in Otomi take place in derivative processes. Derivation tends to be less predictable than inflection phenomena (derivation is less frequent and less regular), and this could be an additional reason of why the entropy values of this language are high.

4 Conclusions

In this work we used corpus based measures like TTR, entropy and perplexity for exploring the morphological complexity of three languages, using two small parallel corpora. We use TTR as a measure of morphological productivity of a language, and we use the entropy and perplexity calculated over a sequence of morphs, as a measure of predictability.

There may be a common believe that polysynthetical languages are far more complex than analytic ones. However, it is important to take into account the many factors that lay a role in the complexity of the system. We stressed out that morphological complexity has several dimensions that must be taken into account (Baerman et al., 2015).

While some agglutinative polysynthetical languages, like Nahuatl, could be considered complex by the number of morphemes the combinations and the information than can be encoded in a single word; the sequence of these elements may be more predictable than fusional languages like Spanish.

Languages like Otomi, showed high complexity in the two dimensions that we focused in this work (this is consistent with qualitative perspectives (Palancar, 2016)).

These two dimensions of complexity are valid and complementary. Measures like TTR reflect the amount of information that words can encode in a language, languages that have a high TTR have the potential of encoding a lot of functions at the word level, therefore, they produce many different word forms. Perplexity and entropy measured over a sequence of morphs reflect the predictability or degree of uncertainty of these combinations. The higher the entropy (hence, the perplexity), the higher the uncertainty in the combinations of morphs.

This was a preliminary work. Deeper linguistic analysis, more corpora and more languages are needed. However, we believe that quantitative measures extracted from parallel corpora can complement and deepen the study of linguistic complexity. Efforts are currently being made (Bane, 2008). However, more studies are needed, especially for low resources languages.

4.1 Future work

Languages of the world have a wide range of functions that can be codified at the world level. Therefore, it would be interesting to consider the study of more complexity dimensions in our work. Popular quantitative approaches are successful in reflecting how many morphs can be combined into a single word. However, it is also important to take into account how complex the format of a word can be, i.e., not only how many elements can be combined but also what type of elements. For example, Dahl (2009) argues that when a phoneme is added to a word, this process is not as complex as adding a tone.

Another interesting dimension is the complexity of the morphology in terms of acquisition (of native and L2 speakers). Miestamo (2008) points out that this typo of complexity should be made on the basis of psycho-linguistics analysis in both processing and acquisition.

Finally, one important factor that influences language complexity is culture. In many languages, pragmatics nuances are produced via morphological processes. For instance, languages like Nahuatl have a complex honorific or reverential system that is expressed using different types of affixes. Spanish expresses this type of phenomena with morphosyntactic processes. It is a challenging task to be able to quantify all these factors that play a role in the complexity of a language.

Acknowledgements

This work was supported by the Mexican Council of Science and Technology (CONACYT), fund 2016-01-2225, and CB-2016/408885. We also thank the reviewers for their valuable comments and to our friend Morrisé P. Martinez for his unconditional support.

References

Farrell Ackerman and Robert Malouf. 2013. Morphological organization: The low conditional entropy conjecture. *Language*, 89(3):429–464.

Vivien Altmann and Gabriel Altmann. 2008. Anleitung zu quantitativen textanalysen. *Methoden und Anwendungen*.

Stephen R Anderson. 2015. Dimensions of morphological complexity. *Understanding and measuring morphological complexity*, pages 11–26.

Mark Aronoff and Kirsten Fudeman. 2011. *What is morphology?*, volume 8. John Wiley & Sons.

Matthew Baerman, Dunstan Brown, and Greville G Corbett. 2015. *Understanding and measuring morphological complexity*. Oxford University Press, USA.

Matthew Baerman. 2012. Paradigmatic chaos in nuer. *Language*, 88(3):467–494.

Max Bane. 2008. Quantifying and measuring morphological complexity. In *Proceedings of the 26th west coast conference on formal linguistics*, pages 69–76. Somerville, MA, USA: Cascadilla Proceedings Project.

Christian Bentz, Tatjana Soldatova, Alexander Koplenig, and Tanja Samardžić. 2016. A comparison between morphological complexity measures: typological data vs. language corpora.

James P Blevins. 2013. The information-theoretic turn. *Psihologija*, 46(4):355–375.

Gilles Boyé and Patricia Hofherr. 2006. The structure of allomorphy in spanish verbal inflection. *Cuadernos de Lingüística del Instituto Universitario Ortega y Gasset*, 13:9–24.

Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. 1992. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479.

Joan Bybee. 2010. *Language, usage and cognition*. Cambridge University Press.

Stanley F Chen and Joshua Goodman. 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–394.

Östen Dahl. 2009. *Testing the assumption of complexity invariance: The case of Elfdalian and Swedish*. na.

Ximena Gutierrez-Vasques, Gerardo Sierra, and Isaac Hernandez Pompa. 2016. Axolotl: a web accessible parallel corpus for spanish-nahuatl. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

Ximena Gutierrez-Vasques. 2017. Exploring bilingual lexicon extraction for Spanish-Nahuatl. In *ACL Workshop in Women and Underrepresenting Minorities in Natural Language Processing*.

Martin Haspelmath and Andrea Sims. 2013. *Understanding morphology*. Routledge.

Gustav Herdan. 1966. *The advanced theory of language as choice and chance*. Springer-Verlag New York.

Justine Kao and Dan Jurafsky. 2012. A computational analysis of style, affect, and imagery in contemporary poetry. In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, pages 8–17.

Emmerich Kelih. 2010. The type-token relationship in slavic parallel texts. *Glottometrics*, 20:1–11.

Kimmo Kettunen. 2014. Can type-token ratio be used to show morphological complexity of languages? *Journal of Quantitative Linguistics*, 21(3):223–245.

Yolanda Lastra. 1992. *El otomí de Toluca*. IIA, UNAM.

David Malvern and Brian Richards. 2002. Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language testing*, 19(1):85–104.

Thomas Mayer, Bernhard Wälchli, Christian Rohrdantz, and Michael Hund. 2014. From the extraction of continuous features in parallel texts to visual analytics of heterogeneous areal-typological datasets. *Language Processing and Grammars. The role of functionally oriented computational models*, pages 13–38.

Matti Miestamo. 2008. Grammatical complexity in a cross-linguistic perspective. *Language complexity: Typology, contact, change*, pages 23–41.

David Mitchell. 2015. Type-token models: a comparative study. *Journal of Quantitative Linguistics*, 22(1):1–21.

Fabio Montermini and Olivier Bonami. 2013. Stem spaces and predictability in verbal inflection. *Lingue e linguaggio*, 12(2):171–190.

Enrique L Palancar. 2004. Verbal morphology and prosody in otomi. *International journal of American linguistics*, 70(3):251–278.

Enrique L Palancar. 2012. The conjugation classes of tilapa otomi: An approach from canonical typology.

Enrique L Palancar. 2016. A typology of tone and inflection: A view from the oto-manguean languages of mexico. *Tone and inflection: New facts and new perspectives*, pages 109–139.

Geoffrey Sampson, David Gil, and Peter Trudgill. 2009. *Language complexity as an evolving variable*, volume 13. Oxford University Press.

Claude E Shannon, Warren Weaver, and Arthur W Burks. 1951. The mathematical theory of communication.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556.

Sami Virpioja, Peter Smit, Stig-Arne Grönroos, Mikko Kurimo, et al. 2013. Morfessor 2.0: Python implementation and extensions for morfessor baseline.