# A Resource for Studying Chatino Verbal Morphology

**Hilaria Cruz[1], Antonios Anastasopoulos[2], and Gregory Stump[3]**
[1]University of Louisville, 2211 South Brook, Louisville KY 40292, USA,
[2]Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh PA 15213, USA,
[3]University of Kentucky, 13964 West 144th Court, Olathe KS 66062, USA
hilaria.cruz@louisville.edu, aanastas@cs.cmu.edu, gregorystump76@gmail.com

**Abstract**

We present the first resource focusing on the verbal inflectional morphology of San Juan Quiahije Chatino, a tonal mesoamerican language spoken in Mexico. We provide a collection of complete inflection tables of 198 lemmata, with morphological tags based on the UniMorph schema. We also provide baseline results on three core NLP tasks: morphological analysis, lemmatization, and morphological inflection.

**Keywords:** Chatino, Endangered Languages, Morphology

## 1. Introduction

The recent years have seen unprecedented forward steps for Natural Language Processing (NLP) over almost every NLP subtask, relying on the advent of large data collections that can be leveraged to train deep neural networks. However, this progress has solely been observed in languages with significant data resources, while low-resource languages are left behind.

The situation for endangered languages is usually even worse, as the focus of the scientific community mostly relies in *language documentation*. The typical endangered language documentation process typically includes the creation of language resources in the form of word lists, audio and video recordings, notes, or grammar fragments, with the created resources then stored into large online linguistics archives. This process is often hindered by the so-called Transcription Bottleneck, but recent advances (Ćavar et al., 2016; Adams et al., 2018) provide promising directions towards a solution for this issue.

However, language documentation and linguistic description, although extremely important itself, does not meaningfully contribute to *language conservation*, which aims to ensure that the language stays in use. We believe that a major avenue towards continual language use is by further creating language technologies for endangered languages, essentially elevating them to the same level as high-resource, economically or politically stronger languages.

The majority of the world's languages are categorized as synthetic, meaning that they have rich morphology, be it fusional, agglutinative, polysynthetic, or a mixture thereof. As Natural Language Processing (NLP) keeps expanding its frontiers to encompass more and more languages, modeling of the grammatical functions that guide language generation is of utmost importance. It follows, then, that the next crucial step for expanding NLP research on endangered languages is creating benchmarks for standard NLP tasks in such languages.

With this work we take a small first step towards this direction. We present a resource that allows for benchmarking two NLP tasks in San Juan Quiahije Chatino, an endangered language spoken in southern Mexico: morphological analysis and morphological inflection, with a focus on the verb morphology of the language.

We first briefly discuss the Chatino language and the intricacies of its verb morphology (§2), then describe the resource (§3), and finally present baseline results on both the morphological analysis and the inflection tasks using state-of-the-art neural models (§4). We make our resource publicly available online[1].

| | All | Train | Dev | Test |
|---|---|---|---|---|
| Paradigms | 198 | | | |
| Verb Classes | 29 | | | |
| Forms | 4716 | 3774 | 471 | 471 |

Table 1: Basic Statistics of our resource.

## 2. The Chatino Language

Chatino is a group of languages spoken in Oaxaca, Mexico. Together with the Zapotec language group, the Chatino languages form the Zapotecan branch of the Otomanguean language family. There are three main Chatino languages: Zenzontepec Chatino (ZEN, ISO 639-2 code czn), Tataltepec Chatino (TAT, cta), and Eastern Chatino (ISO 639-2 ctp, cya, ctz, and cly) (E.Cruz 2011 and Campbell 2011). San Juan Quiahije Chatino (SJQ), the language of the focus of this study, belongs to Eastern Chatino, and is used by about 3000 speakers.

**Typology and Writing System** Eastern Chatino languages , including SJQ Chatino, are intensively tonal (Cruz, 2004; Cruz and Woodbury, 2014). Tones mark both lexical and grammatical distinctions in Eastern Chatino languages. In SJQ Chatino, there are eleven tones. Three different systems for representing tone distinctions are employed in the literature: the S-H-M-L system of (Cruz, 2004); the numeral system of (Cruz, 2014); and the alphabetic system of (Cruz and Woodbury, 2014). The correspondences among these three systems are given in Table 2. For present purposes, we will use numeral representations of the second

[1] https://github.com/antonisa/chatino_inflection_paradigms

| Tone description | S-H-M-L (Cruz 2011) | Numeral (Cruz 2014) | Alphabetic (Cruz & Woodbury 2013) |
|---|---|---|---|
| high | H | 1 | E |
| high-superhigh | HS | 10 | D |
| high-low | HL | 14 | B |
| mid | M | 2 | C |
| mid-superhigh | MS | 20 | H |
| mid-high | MH | 32 | I |
| mid-low | ML | 24 | J |
| low | L | 4 | A |
| low-superhigh | LS | 40 | M |
| low-high | LH | 42 | G |
| low-mid | LM | 3 | F |

Table 2: Three alternative systems for representing the SJQ Chatino tones.

sort. The number 1 represents a high pitch, 4 represents a low pitch, and double digits represent contour tones.

**Verb Morphology** SJQ Chatino verb inflection distinguishes four aspect/mood categories: completive (*'I did'*), progressive (*'I am doing'*), habitual (*'I habitually do'*) and potential (*'I might do'*). In each of these categories, verbs inflect for three persons (first, second, third) and two numbers (singular, plural) and distinguish inclusive and exclusive categories of the first person plural (*'we including you'* vs *'we excluding you'*). Verbs can be classified into dozens of different conjugation classes. Each conjugation class involves its own tone pattern; each tone pattern is based on a series of three person/number (PN) triplets. A PN triplet [X, Y, Z] consists of three tones: tone X is employed in the third person singular as well as in all plural forms; tone Y is employed in the second person singular, and tone Z, in the third person singular. Thus, a verb's membership in a particular conjugation class entails the assignment of one tone triplet to completive forms, another to progressive forms, and a third to habitual and potential forms. The paradigm of the verb lyu1 *'fall'* in Table 3 illustrates: the conjugation class to which this verb belongs entails the assignment of the triplet [1, 42, 20] to the completive, [1, 42, 32] to the progressive, and [20, 42, 32] to the habitual and potential. Verbs in other conjugation classes exhibit other triplet series.[2]

## 3. The Resource

We provide a hand-curated collection of complete inflection tables for 198 lemmata. The morphological tags follow the guidelines of the UniMorph schema (Sylak-Glassman, 2016; Kirov et al., 2018), in order to allow for the potential of cross-lingual transfer learning, and they are tagged with respect to:

- Person: first (1), second (2), and third (3)
- Number: singular (SG) ad plural (PL)
- Inclusivity (only applicable to first person plural verbs: inclusive (INCL) and exclusive (EXCL)
- Aspect/mood: completive (CPL), progressive (PROG), potential (POT), and habitual (HAB).

Two examples of complete inflection tables for the verbs *ndyu2* 'fell from above' and *lyu1* 'fall' are shown in Table 3. Note how the first verb has the same PN triplet for all four aspect/mood categories, while the second paradigm is more representative in that it involves three different triplets (one for the completive, another for the progressive, and another for the habitual/potential). This variety is at the core of why the SJQ verb morphology is particularly interesting, and a challenging testcase for modern NLP systems.

In total, we end up with 4716 groupings (triplets) of a lemma, a tag-set, and a form; we split these groupings randomly into a training set (3774 groupings), a development set (471 groupings), and test set (471 groupings). Basic statistics of the corpus are outlined in Table 1 1. Compared to all the other languages from the Unimorph project, this puts SJQ Chatino in the low- to mid-resource category, but nonetheless it is more than enough for benchmarking purposes.[3]

## 4. Baseline Results

**Inflectional realization** Inflectional realization defines the inflected forms of a lexeme/lemma. As a computational task, often referred to as simply "morphological inflection," inflectional realization is framed as a mapping from the pairing of a lemma with a set of morphological tags to the corresponding word form. For example, the inflectional realization of SJQ Chatino verb forms entails a mapping of the pairing of the lemma lyu1 'fall' with the tag-set 1;SG;PROG to the word form nlyon32.

Morphological inflection has been thoroughly studied in monolingual high resource settings, especially through the recent SIGMORPHON challenges (Cotterell et al., 2016; Cotterell et al., 2017; Cotterell et al., 2018), with the latest iteration focusing more on low-resource settings, utilizing cross-lingual transfer (McCarthy et al., 2019). We use the guidelines of the state-of-the-art approach of (Anastasopoulos and Neubig, 2019) that achieved the highest inflection accuracy in the latest SIGMORPHON 2019 morphological inflection shared task. Our models are implemented in DyNet (Neubig et al., 2017).

Inflection results are outlined in Table 4. In the 'standard' setting we simply train on the pre-defined training set, achieving an exact-match accuracy of 60% over the test set. Interestingly, the data augmentation approach of (Anastasopoulos and Neubig, 2019) that hallucinates new training paradigms based on character level alignments does not heed significant improvements in accuracy (only 2 percentage points increase, cf. with more than 15 percentage points increases in other languages). These results indicate that automatic morphological inflection for low-resource tonal languages like SJQ Chatino poses a particularly challenging setting, which perhaps requires explicit handling of tone information by the model.

[2] A more thorough introduction into Chatino verbal morphology will appear at (Cruz and Stump, 2020).

[3] In future work we will investigate whether more controlled training-development-test splits such that the splits is non-random but rather across whole lemmata or even across whole verb classes results in different generalization issues.

| | Aspect: | CPL | PROG | HAB | POT |
|---|---|---|---|---|---|
| ndyu2 'fell from above' | | | | | |
| | PN triple: | 2-1-40 | 2-1-40 | 2-1-40 | |
| Singular | 1 | ndyon40 | ndyon40 | ndyon40 | tyon40 |
| | 2 | ndyu1 | ndyu1 | ndyu1 | tyu1 |
| | 3 | ndyu2 | ndyu2 | ndyu2 | tyu2 |
| Plural | 1 inclusive | ndyon2on1 | ndyon2on1 | ndyon2on1 | ntyon2on1 |
| | 1 exclusive | ndyu2 wa42 | ndyu2 wa42 | ndyu2 wa42 | ntyu2 wa42 |
| | 2 | ndyu2 wan1 | ndyu2 wan1 | ndyu2 wan1 | ntyu2 wan1 |
| | 3 | ndyu2 renq1 | ndyu2 renq1 | ndyu2 renq1 | ntyu2 renq1 |
| lyu1 'to fall' | | | | | |
| | PN triple: | 1-42-20 | 1-42-32 | 20-42-32 | |
| Singular | 1 | lyon20 | nlyon32 | nlyon32 | klyon32 |
| | 2 | lyu42 | nlyu42 | nlyu42 | klyu42 |
| | 3 | lyu1 | nlyu1 | nlyu20 | klyu20 |
| Plural | 1 inclusive | lyon1on1 | nlyon1on1 | nlyon20on32 | klyon20on32 |
| | 1 exclusive | lyu1 wa42 | nlyu1 wa42 | nlyu20 wa42 | klyu20 wa42 |
| | 2 | lyu1 wan24 | lyu1 wan24 | nlyu20 wan24 | klyu20 wan24 |
| | 3 | lyu1 renq24 | lyu1 renq24 | nlyu20 renq24 | klyu20 renq24 |

Table 3: Complete inflection paradigms for two lemmata: one with a single PN triple across all aspects (top), and one with three different PN triples (bottom).

| Setting | Accuracy | Levenshtein distance |
|---|---|---|
| standard | 60% | 0.92 |
| +hallucinated data | 62% | 1.02 |

Table 4: Morphological Inflection Results

| Setting | Exact Match Accuracy |
|---|---|
| standard | 67% |

Table 5: Morphological Analysis Results

**Morphological Analysis** Morphological analysis is the task of creating a morphosyntactic description for a given word. It can be framed in a context-agnostic manner (as in our case) or within a given context, as for instance for the SIGMORPHON 2019 second shared task (McCarthy et al., 2019). We trained an encoder-decoder model that receives the form as character-level input, encodes it with a BiLSTM encoder, and then an attention enhanced decoder (Bahdanau et al., 2014) outputs the corresponding sequence of morphological tags, implemented in DyNet. The baseline results are shown in Table 5. The exact-match accuracy of 67% is lower than the average accuracy that context-aware systems can achieve, and it highlights the challenge that the complexity of the tonal system of SJQ Chatino can pose.

**Lemmatization** Lemmatization is the task of retrieving the underlying lemma from which an inflected form was derived. Although in some languages the lemma is distinct from all forms, in SJQ Chatino the lemma is defined as the completive third-person singular form. As a computational task, lemmatization entails producing the lemma given an inflected form (and possibly, given a set of morphological tags describing the input form). Popular approaches tackle it as a character-level edit sequence generation task (Chrupała, 2006), or as a character-level sequence-to-sequence task (Bergmanis and Goldwater, 2018). For our baseline lemmatization systems we follow the latter approach. We trained a character level encoder-decoder model, similar to the above-mentioned inflection system, implemented in DyNet.

| Input | Accuracy | Levenshtein distance |
|---|---|---|
| form (no tags) | 89% | 0.27 |
| form + tags | 90% | 0.21 |

Table 6: Lemmatization Results.

The baseline results, with and without providing gold morphological tags along with the inflected form as input, are outlined in Table 6. We find that automatic lemmatization in SJQ Chatino achieves fairly high accuracy even with our simple baseline models (89% accuracy, $0.27$ average Levenshtein distance) and that providing the gold morphological tags provides a performance boost indicated by small improvements on both metrics. It it worth noting, though, that these results are also well below the $94 - -95\%$ average accuracy and $0.13$ average Levenshtein distance that lemmatization models achieved over 107 treebanks in 66 languages for the SIGMORPHON 2019 shared task (McCarthy et al., 2019).

## 5. Related Work

Our work builds and expands upon previous work on Indigenous languages of the Americas specifically focusing on the complexity of their morphology. Among other works

similar to ours, (Cox et al., 2016) focused on the morphology of Dene verbs, (Moeller et al., 2018) on Arapaho verbs, (Cardenas and Zeman, 2018) on Shipibo-Konibo, and (Chen and Schwartz, 2018) on Saint Lawrence Island and Central Siberian Yupik. (Sylak-Glassman et al., 2016) describe an approach for elicit complete inflection paradigms, with experiments in languages like Nahuatl. Our resource is the first one for SJQ Chatino, but it also provides an exciting new data point in the computational study of morphological analysis, lemmatization, and inflection, as it is the first one in a tonal language with explicit tonal markings in the writing system. In a similar vein, the Oto-Manguean Inflectional Class Database[4] (Palancar and Feist, 2015) provides a valuable resource for studying the verbal morphology of Oto-Manguean languages (including a couple of other Chatino variants: Yaitepec and Zenzotepec Chatino) but not in a format suitable for computational experiments.

## 6. Conclusion

We presented a resource of 198 complete inflectional paradigms in San Juan Quiahije Chatino, which will facilitate research in computational morphological analysis and inflection for low-resource tonal languages and languages of Mesoamerica. We also provide strong baseline results on computational morphological analysis, lemmatization, and inflection realization, using character-level neural encoder-decoder systems.

For future work, while we will keep expanding our resource to include more paradigms, we will also follow the community guidelines in extending our resource to include morphological analysis and inflection examples *in context*.

## 7. Acknowledgements

Part of this work was done during the Workshop on Language Technology for Language Documentation and Revitalization.[5] This material is based upon work generously supported by the National Science Foundation under grant 1761548.

## 8. Bibliographical References

Adams, O., Cohn, T., Neubig, G., Cruz, H., Bird, S., and Michaud, A. (2018). Evaluation phonemic transcription of low-resource tonal languages for language documentation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Anastasopoulos, A. and Neubig, G. (2019). Pushing the limits of low-resource morphological inflection. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 983–995, Hong Kong, China, November. Association for Computational Linguistics.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv:1409.0473.

Bergmanis, T. and Goldwater, S. (2018). Context sensitive neural lemmatization with Lematus. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1391–1400, New Orleans, Louisiana. Association for Computational Linguistics.

Cardenas, R. and Zeman, D. (2018). A morphological analyzer for shipibo-konibo. In *Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 131–139.

Ćavar, M., Ćavar, D., and Cruz, H. (2016). Endangered language documentation: Bootstrapping a chatino speech corpus, forced aligner, asr. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4004–4011.

Chen, E. and Schwartz, L. (2018). A morphological analyzer for st. lawrence island / central siberian yupik. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Chrupała, G. (2006). Simple data-driven context-sensitive lemmatization.

Cotterell, R., Kirov, C., Sylak-Glassman, J., Yarowsky, D., Eisner, J., and Hulden, M. (2016). The SIGMORPHON 2016 shared task—morphological reinflection. In *Proc. SIGMORPHON*.

Cotterell, R., Kirov, C., Sylak-Glassman, J., Walther, G., Vylomova, E., Xia, P., Faruqui, M., Kübler, S., Yarowsky, D., Eisner, J., and Hulden, M. (2017). CoNLL-SIGMORPHON 2017 shared task: Universal morphological reinflection in 52 languages. In *Proc. CoNLL SIGMORPHON*.

Cotterell, R., Kirov, C., Sylak-Glassman, J., Walther, G., Vylomova, E., McCarthy, A. D., Kann, K., Mielke, S., Nicolai, G., Silfverberg, M., Yarowsky, D., Eisner, J., and Hulden, M. (2018). The CoNLL–SIGMORPHON 2018 shared task: Universal morphological reinflection. In *Proc. CoNLL–SIGMORPHON*.

Cox, C., Hulden, M., Silfverberg, M., Lachler, J., Rice, S., Moshagen, S. N., Trosterud, T., and Arppe, A. (2016). Computational modeling of the verb in dene languages–the case of tsuut'ina. In *Dene Languages Conference*.

Cruz, H. and Stump, G. (2020). The complex exponence relations of tonal inflection in sjq chatino verbs.

Cruz, E. and Woodbury, A. C. (2014). Finding a way into a family of tone languages: The story and methods of the chatino language documentation project. *Language Documentation & Conservation*, 8:490–524.

Cruz, E. (2004). The phonological patterns and orthography of san juan quiahije chatino. *University of Texas Masters Thesis. Austin*.

Cruz, H. (2014). *Linguistic poetic and rhetoric of Eastern Chatino of San Juan Quiahije*. Ph.D. thesis, University of Texas.

Kirov, C., Cotterell, R., Sylak-Glassman, J., Walther, G., Vylomova, E., Xia, P., Faruqui, M., Mielke, S. J., McCarthy, A., Kübler, S., et al. (2018). Unimorph 2.0: Uni-

[4] http://www.oto-manguean.surrey.ac.uk/

[5] https://sites.google.com/view/ltldr/home

versal morphology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

McCarthy, A. D., Vylomova, E., Wu, S., Malaviya, C., Wolf-Sonkin, L., Nicolai, G., Kirov, C., Silfverberg, M., Mielke, S. J., Heinz, J., et al. (2019). The sigmorphon 2019 shared task: Morphological analysis in context and cross-lingual transfer for inflection. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–244.

Moeller, S., Kazeminejad, G., Cowell, A., and Hulden, M. (2018). A neural morphological analyzer for arapaho verbs learned from a finite state transducer. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*, pages 12–20.

Neubig, G., Dyer, C., Goldberg, Y., Matthews, A., Ammar, W., Anastasopoulos, A., Ballesteros, M., Chiang, D., Clothiaux, D., Cohn, T., et al. (2017). Dynet: The dynamic neural network toolkit. arXiv:1701.03980.

Palancar, E. and Feist, T. (2015). Oto-manguean inflectional class database.

Sylak-Glassman, J., Kirov, C., and Yarowsky, D. (2016). Remote elicitation of inflectional paradigms to seed morphological analysis in low-resource languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3116–3120.

Sylak-Glassman, J. (2016). The composition and use of the universal morphological feature schema (unimorph schema). *Johns Hopkins University*.