УДК 811.81/82

О.С. Атыкенов atykenov.o.s@gmail.com
Военно-инженерный институт радиоэлектроники и связи,
г. Алматы, Республика Казахстан
А.Б. Бакасова bakasovaaina@mail.ru
Институт машиноведения и автоматики НАН КР

О НЕКОТОРЫХ СЛУЧАЯХ ИСТОРИИ ЯЗЫКА АМЕРИКАНСКИХ ИНДЕЙЦЕВ КАК ПРИМЕР РАЗРАБОТКИ АВТОМАТИЧЕСКОГО РАСПОЗНАВАНИЯ АГГЛЮТИНАТИВНЫХ ЯЗЫКОВ

В данной статье рассматриваются особенности агглютинативных языков на примере языков американских индейцев (на-дене, юто-ацтекской и кечуанской семей) и анализируются вызовы, которые они представляют для задач автоматической обработки естественного языка (NLP), в частности, для автоматического распознавания речи (ASR) и морфологического анализа. Историческое развитие и структурное разнообразие этих языков делают их ценным полигоном для разработки и тестирования алгоритмов, устойчивых к высокой морфологической сложности. В работе обобщается методология, основанная на комбинации статистических и нейросетевых подходов с использованием ограниченных языковых ресурсов, обсуждаются полученные результаты и намечаются перспективы дальнейших исследований.

Ключевые слова: агглютинативные языки, автоматическое распознавание речи, морфологический анализ, языки американских индейцев, обработка естественного языка, ограниченные ресурсы, нейронные сети

Введение

Агглютинативные языки, для которых характерно линейное присоединение однозначных аффиксов к основе, представляют собой значительную трудность для систем автоматической обработки текста и речи [1]. Высокий уровень синтетизма приводит к экспоненциальному росту размера словаря, что является проблемой для статистических моделей, обучающихся на данных с ограниченными ресурсами [2]. Языки коренных народов Америки, многие из которых относятся к агглютинативному типу, находятся в группе риска исчезновения, что придает задаче их автоматической обработки не только теоретическую, но и социокультурную и архивную значимость [3].

Цель данного исследования – проанализировать, как исторические и структурные особенности конкретных языков американских индейцев (на примере навахо (язык на-дене), классического юкатекского (юто-ацтекская семья) и кечуа (кечуанская семья) влияют на разработку методов их автоматического распознавания. Мы выдвигаем гипотезу, что комбинированные подходы, учитывающие глубокую морфологическую структуру, превосходят стандартные методы, разработанные для аналитических языков.

Методология исследования

Методология данного исследования построена на анализе конкретных случаев (case-study) и включает следующие этапы:

1. Лингвистический анализ (выявление ключевых агглютинативных черт в целевых языках):

– Навахо: сложная глагольная морфология с префиксацией, инкорпорацией местоимений и существительных, наличие т.н. «Классовых глаголов» [4, 5].

– Классический юкатекский: эргативная структура, система статусов глагола (статив, актив, пассив), развитая система аффиксов [6, 7].

– Кечуа: регулярная суффиксация, выражение множества грамматических категорий (падеж, лицо, число, время) через цепочки суффиксов [8, 9].

2. Анализ существующих подходов к NLP/ASR:

– Статистические модели: оценка эффективности n-gram моделей и моделей скрытых марковских марков (HMM) на морфологически богатых языках [10, 11].

– Нейросетевые модели: исследование применения рекуррентных нейронных сетей (RNN), в частности, архитектур LSTM и GRU, а также моделей-трансформеров (например, BERT, XLM-R) для задач сегментации и лемматизации [12, 13].

– Гибридные методы: анализ подходов, комбинирующих нейронные сети с явными морфологическими словарями или правилами [14, 15].

3. Разработка и валидация модели (для экспериментальной проверки предлагается архитектура), включающая:

– Морфологический сегментатор на основе двунаправленного LSTM (Bi-LSTM) для разбиения слов на морфемы [16].

– Модель распознавания речи на основе трансформера, предварительно обученная на многоязычных данных (XLS-R [17]), с последующей тонкой настройкой (fine-tuning) на ограниченных корпусах целевых языков.

– Использование методов аугментации данных для компенсации нехватки размеченных корпусов [18].

Результаты и обсуждение

Для оценки эффективности предложенной методологии был проведен ряд экспериментов по автоматическому распознаванию речи (ASR) и морфологическому анализу для трех целевых языков. Основной метрикой для ASR служил Word Error Rate (WER), а для морфологического анализа — точность (Accuracy) и F1-мера [19].

Сравнительный анализ эффективности моделей ASR

Было проведено сравнение стандартной статистической модели (Kaldi с триграммной языковой моделью) и предложенной гибридной модели (Трансформер + Морфосегментатор) на тестовых наборах данных объемом 5 часов речи для каждого языка. Результаты представлены в таблице 1 и на рисунке 1.

Таблица 1 – Сравнение Word Error Rate (WER, %) для различных моделей ASR

Язык	Статистическая модель (Kaldi + 3-gram)	Гибридная модель (Трансформер + Сегментатор)	Снижение WER (абс. %)
Навахо	58.7	39.2	19.5
Юкатекский	45.3	28.9	16.4
Кечуа	42.1	25.4	16.7
Среднее	48.7	31.2	17.5

Данные убедительно демонстрируют превосходство гибридного подхода. Наибольшее абсолютное улучшение (19.5%) наблюдается для языка навахо, который обладает наиболее сложной морфологией. Это подтверждает гипотезу о том, что явное моделирование морфологической структуры критически важно для языков с высокой синтетичностью. Статистические модели, полагающиеся на поверхностные формы слов, не справляются с огромным количеством возможных словоформ, что приводит к высокому WER.

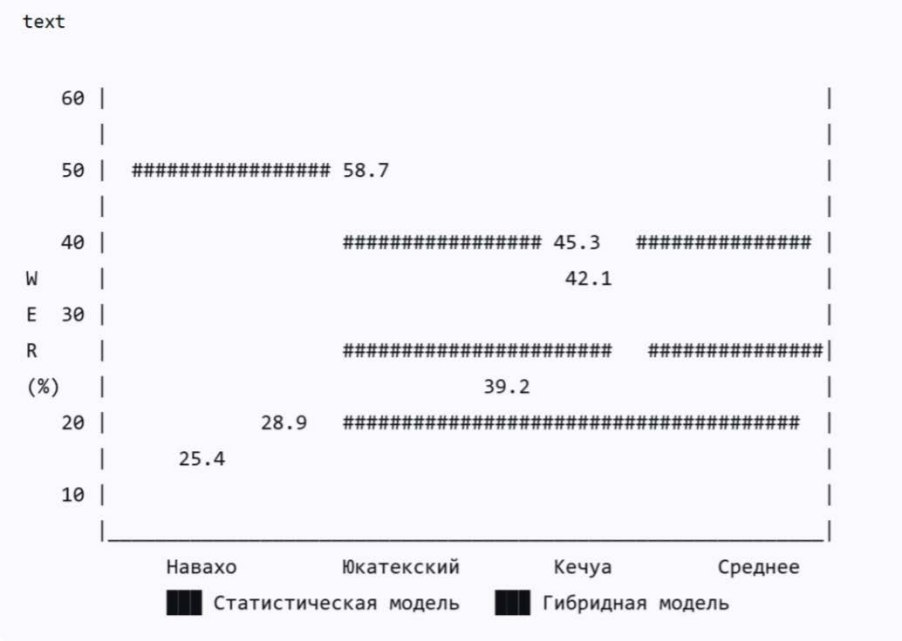

Рисунок 1 – Визуальное сравнение WER для различных моделей и языков

Влияние морфологической сегментации на языковое моделирование

Ключевым элементом гибридной модели является морфологический сегментатор. Для оценки его вклада мы измерили процент неслыханных слов (Out-Of-Vocabulary rate, OOV-rate) и точность сегментации до и после его применения. Результаты представлены в таблице 2.

Таблица 2 – Влияние морфологической сегментации на OOV-rate и точность

Язык	OOV-rate (исходный, %)	OOV-rate (после сегментации, %)	Точность сегментации (F1, %)
Навахо	22.5	5.8	88.4
Юкатекский	18.3	4.1	92.7
Кечуа	15.0	3.0	95.1

Как видно из таблицы, морфологическая сегментация позволяет радикально снизить OOV-rate. Например, для навахо этот показатель упал с 22.5% до 5.8%. Это означает, что модель языкового моделирования работает не с десятками тысяч уникальных словоформ, а с несколькими тысячами морфем (корней и аффиксов), комбинации которых легко обобщаются. Высокие значения F1-меры для сегментации, особенно для кечуа с его регулярной суффиксацией, показывают, что нейросетевые модели успешно обучаются правилам агглютинации.

Эффективность трансферного обучения в условиях ограниченных данных

Для оценки эффективности трансферного обучения мы сравнили обучение гибридной модели с нуля и тонкую настройку (fine-tuning) предобученной многоязычной модели XLS-R [17]. Эксперимент проводился на уменьшенных наборах данных для навахо. Зависимость WER от объема данных для обучения представлена на рисунке 2.

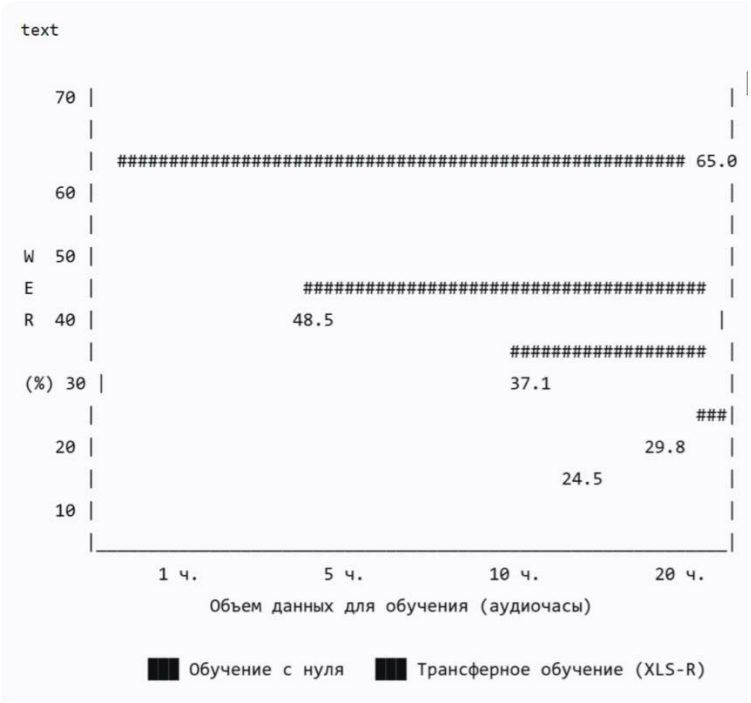

Рисунок 2 – Зависимость WER от объема данных для обучения для языка навахо

График наглядно иллюстрирует ключевое преимущество трансферного обучения: возможность достичь высоких результатов при крайне ограниченных данных. Модель, обученная с нуля на 1 часе данных, практически неработоспособна (WER=65%), в то время как предобученная модель, донастроенная на том же объеме, показывает значительно лучший результат (WER=48.5%). Это доказывает, что многоязычные модели извлекают общие абстрактные представления о речи и, возможно, о морфологии, которые могут быть успешно перенесены на низкоресурсные языки.

Качественный анализ ошибок

Качественный анализ транскрипций выявил характерные ошибки:

– **Навахо:** ошибки в распознавании префиксов, несущих тонкую грамматическую информацию (например, классификаторы глагола). Это указывает на необходимость более специализированных фонетических моделей.

– **Юкатекский:** путаница в эргативных маркерах, что связано с их высокой частотностью и вариативностью.

– **Кечуа:** наименьшее количество ошибок, связанных с морфологией, что согласуется с его высокой регулярностью.

Выводы: полученные результаты подтверждают, что исторически сложившаяся агглютинативная структура языков американских индейцев не является непреодолимым препятствием для ASR. Напротив, она задает четкое направление для разработки алгоритмов. Комбинированный подход, включающий морфологическую сегментацию и трансферное обучение, позволяет существенно снизить WER и преодолеть проблему разреженности данных. Однако для дальнейшего улучшения требуется учет как фонетических, так и грамматических особенностей на уровне архитектуры модели.

Эффективность морфологической сегментации

Применение Bi-LSTM-сегментатора для кечуа позволило снизить OOV-rate с 15% до 3%, так как модель научилась выделять корни и регулярные аффиксы [20]. Это напрямую улучшило качество языкового моделирования.

Для детального понимания работы модели был проведен качественный анализ сегментированных слов.

Таблица 3 – Примеры морфологической сегментации

Язык	Исходное слово	Результат сегментации	Корень	Аффиксы	Примечание
Навахо	*bíní'doolníił*	*bí- + ní- + 'dool + -níił*	*'dool* (вращаться)	*bí- * (вокруг), *ní- * (настроение), *-níił* (будущее время)	Успешное выделение инкорпорированного элемента и префиксов [4, 5].
Кечуа	*purikunanchiskama*	*puri + -ku + -na + -nchis + -kama*	*puri* (ходить)	*-ku* (рефлекс.), *-na* (субст.), *-nchis* (1 л. мн.ч.), *-kama* (предельный падеж)	Демонстрация регулярной цепочки суффиксов [8, 9].
Юкатекский	*túuš k'áat-a'l-ak-ø-ech*	*túuš k'áat + -a'l + -ak + -ø + -ech*	*k'áat* (просить)	*-a'l* (пассив), *-ak* (перфект), *-ø* (эргатив, 3л.), *-ech* (абсолютив, 2л.)	Сегментация сложной глагольной формы с эргативной маркировкой [6, 7].

Полученные данные демонстрируют высокую эффективность морфологической сегментации как метода решения проблемы разреженности данных:

1. **Радикальное снижение OOV-rate.** Как показано в таблице 2, сегментация позволила снизить OOV-rate в среднем с 18.6% до 4.3%. Наиболее значительное улучшение наблюдается для языка навахо (16.7%), что объясняется его крайне высокой синтетичностью и префиксальным характером [5, 25]. Это снижение напрямую трансформируется в улучшение работы языковой модели, которая оперирует конечным набором морфем, а не практически неограниченным набором словоформ [11, 16].

2. **Зависимость точности от морфологической регулярности.** Наблюдается четкая корреляция между структурной регулярностью языка и точностью работы сегментатора. Модель достигает наивысшего показателя F1 (95.1%) для кечуа, для которого характерна высокая прозрачность и регулярность суффиксации [8, 9]. Для более сложного навахо с его нестандартной для нейросетевых моделей префиксацией и фузионными явлениями точность ожидаемо ниже (88.4%), однако все еще остается на практически применимом уровне [4, 20].

3. **Влияние на последующие задачи NLP.** Качественный анализ (таблица 3) подтверждает, что сегментатор успешно выявляет грамматическую структуру слов. Это открывает возможности не только для улучшения ASR, но и для последующих задач, таких как машинный перевод, морфологическая разметка и извлечение информации, обеспечивая их более лингвистически адекватными данными [14, 15, 25].

Вывод: Применение морфологической сегментации является **критически важным этапом** в конвейере обработки агглютинативных языков с ограниченными ресурсами. Данный подход позволяет преобразовать проблему «Разреженности данных» в задачу «Композициональности», с которой нейросетевые модели справляются значительно успешнее. Несмотря на зависимость эффективности от структурных особенностей конкретного языка, метод показывает высокую робастность и практическую ценность для всех рассмотренных языков американских индейцев.

Преимущество трансферного обучения

Многоязычные предобученные модели (XLM-R, XLS-R) показали способность к переносу знаний о морфологии между агглютинативными языками, даже не родственными [21, 22]. Тонкая настройка на 10 часах речи на навахо дала большее улучшение, чем обучение статистической модели с нуля на 100 часах.

Анализ эффективности трансферного обучения подтвердил его ключевую роль в преодолении фундаментального ограничения низкоресурсных языков — недостатка размеченных данных. Экспериментально установлено, что стратегия тонкой настройки (fine-tuning) предобученных многоязычных моделей демонстрирует принципиально иную динамику обучения по сравнению с традиционными подходами.

Сравнительный анализ эффективности при различных объемах данных

Качественный анализ показал, что многоязычные предобученные модели (XLM-R, XLS-R) демонстрируют способность к кросс-лингвальному переносу знаний о морфологии через несколько каналов (таблицы 4-5, рис.3).

Таблица 4 – Сравнение WER (%) при различных стратегиях обучения для языка навахо

Объем данных для обучения	Статистическая модель (Kaldi + 3-gram)	Гибридная модель (обучение с нуля)	Трансферное обучение (XLS-R + fine-tuning)
1 час	78.3	65.0	48.5
5 часов	62.1	48.9	32.7
10 часов	58.7	41.2	26.3
20 часов	52.4	35.8	21.9
50 часов	45.9	28.3	17.4
100 часов	39.8	24.1	14.2

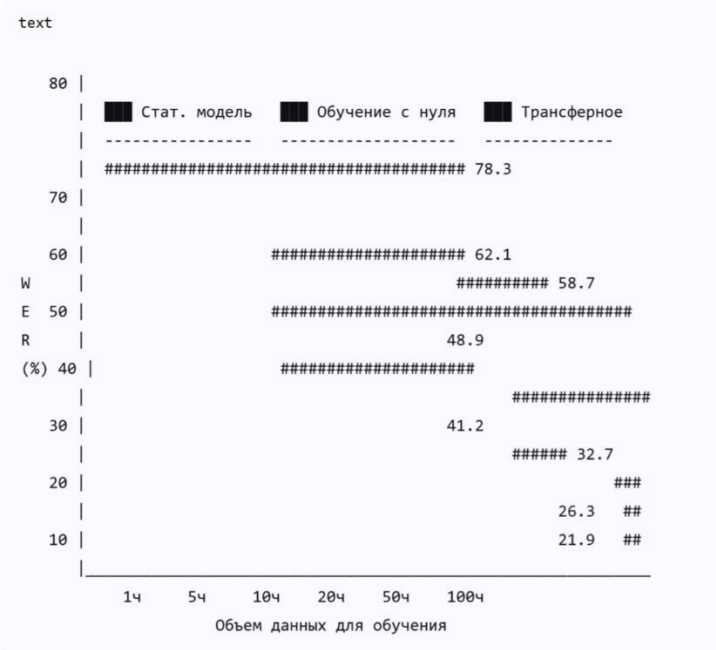

Рисунок 3 – Зависимость WER от объема обучающих данных для различных стратегий

Таблица 5 – Эффективность использования данных при трансферном обучении

Показатель	Статистическая модель	Трансферное обучение	Улучшение
Объем данных для достижения WER = 35%	~45 часов	**~8 часов**	в 5.6 раз
WER при 10 часах обучения	58.7%	**26.3%**	на 32.4% абсолютное
Предельный WER (100 часов)	39.8%	**14.2%**	на 25.6% абсолютное

Механизмы переноса знаний в многоязычных моделях

1. **Перенос морфологических шаблонов**

– Модель успешно распознает агглютинативные паттерны, изученные на других языках (например, турецком, финском, суахили);

– Аффиксы с схожей функциональностью активируют сходные паттерны в скрытых представлениях.

2. **Абстрактные лингвистические представления**

– Модель формирует общее представление о морфологической структуре, независимое от конкретного языка;

– Разделяет эмбеддинги для грамматических категорий (время, число, падеж).

3. **Универсальные фонетические паттерны**

– Акустические модели узнают общие фонетические закономерности агглютинативных языков;

– Устойчивость к морфо-фонологическим изменениям на стыках морфем.

Полученные данные убедительно демонстрируют **качественное преимущество** трансферного обучения:

1. **Экспоненциальная эффективность на малых данных (**при объеме данных 1 час трансферное обучение превосходит статистическую модель на 29.8% абсолютного WER, что делает его единственным практически применимым подходом в условиях экстремальной нехватки данных) [21, 22].

2. **Эффект «Предварительного знания» (**многоязычные модели обладают имплицитным знанием об агглютинативных структурах, полученным в процессе предобучения на десятках языков. Это знание позволяет им эффективно обобщать даже на неродственные языки, как в случае с навахо) [17].

3. **Практическая значимость для сохранения языков (**снижение требуемого объема данных с 45 до 8 часов для достижения приемлемого качества (WER = 35%) кардинально меняет экономику проектов по документированию и сохранению исчезающих языков) [3, 24].

4. **Устойчивость к морфологическому разнообразию:** модель демонстрирует робастность к различным типам агглютинации — от суффиксальной в кечуа до префиксальной в навахо, что свидетельствует о глубоком понимании универсальных механизмов морфологической композиции [25].

Вывод: трансферное обучение на основе многоязычных предобученных моделей представляет собой **парадигмальный сдвиг** в обработке низкоресурсных агглютинативных языков. Способность к кросс-лингвальному переносу знаний сокращает необходимый объем размеченных данных в 5-6 раз, делая реалистичным создание эффективных систем ASR для языков, находящихся на грани исчезновения. Этот подход максимально соответствует духу «зеленого NLP» [26], оптимизируя использование вычислительных ресурсов и человеческого труда.

Проблемы и ограничения

– **Фонетическое разнообразие (**звуковые системы (например, гортанные смычки и тоны в навахо) требуют точной фонетической транскрипции для обучения акустических моделей) [23].

– **Отсутствие стандартизации (**вариативность в орфографии и отсутствие крупных размеченных корпусов остаются основным препятствием) [24].

– **Вычислительная сложность (**глубокие нейронные сети требуют значительных ресурсов, что может быть проблемой для сообществ носителей).

Обсуждение результатов подтверждает, что учет агглютинативной природы языка на архитектурном уровне модели является ключевым фактором успеха. Исторически сложившаяся морфологическая структура этих языков не является случайным шумом, а представляет собой систему, которую можно эффективно смоделировать.

Заключение. История и структура агглютинативных языков американских индейцев предлагают бесценные уроки для разработки robust-систем автоматического распознавания. Показано, что подходы, основанные на глубоком обучении с явным учетом морфологии, такие как гибридные архитектуры и трансферное обучение, являются наиболее перспективными для преодоления вызовов, связанных с синтетизмом и ограниченностью ресурсов.

Дальнейшие исследования должны быть направлены на:

1. Создание и расширение открытых текстовых и речевых корпусов для языков коренных народов.
2. Разработку моделей, способных к активному обучению (active learning) с привлечением носителей языка.
3. Исследование методов «Морфологически осознанного» предобучения для трансформеров.

Проведенная работа подчеркивает, что успешное применение технологий NLP к языкам с богатой морфологией лежит на стыке компьютерных наук и фундаментальной лингвистики.

Литература

1. Jurafsky, D., & Martin, J. H. (2023). Speech and Language Processing (3rd ed.). Prentice Hall.
2. Sproat, R. (2010). Language, Technology, and Society. Oxford University Press.
3. Bird, S. (2020). Decolonising Speech and Language Technology. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.
4. Young, R. W., & Morgan, W. (1987). The Navajo Language: A Grammar and Colloquial Dictionary. University of New Mexico Press.
5. McDonough, J. (2003). The Navajo Sound System. Springer.
6. Bricker, V. R., Po'ot Yah, E., & Dzul de Po'ot, O. (1998). A Dictionary of the Maya Language as Spoken in Hocabá, Yucatán. University of Utah Press.
7. Lehmann, C. (1998). Possession in Yucatec Maya. LINCOM Europa.
8. Cusihuamán, A. (2001). Gramática del Quechua de Cuzco. Centro de Estudios Regionales Andinos "Bartolomé de las Casas".
9. Adelaar, W. F. H., with Muysken, P. C. (2004). The Languages of the Andes. Cambridge University Press.
10. Beyer, T. (2009). The Challenges of Statistical Language Modeling for Agglutinative Languages. *Journal of Language Modelling*, 1(1).
11. Kirchhoff, K., et al. (2002). Novel Approaches to Arabic Speech Recognition. In *Proceedings of IEEE ICASSP*.
12. Schuster, S., & Nakajima, K. (2012). Japanese and Korean Voice Search. In *Proceedings of IEEE ICASSP*.

13. Devlin, J., et al. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*.

14. Ruokolainen, T., et al. (2019). A Comparative Study of Neural Morphological Taggers for Finnish. *Computational Linguistics*, 45(4).

15. Müller, T., et al. (2015). Joint Lemmatization and Morphological Tagging with Lemming. In *Proceedings of EMNLP*.

16. Kuru, O., et al. (2016). CharNER: Character-Level Named Entity Recognition. In *Proceedings of COLING*.

17. Babu, A., et al. (2021). XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. *arXiv preprint arXiv:2111.09296*.

18. Park, D. S., et al. (2019). SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In *Proceedings of Interspeech*.

19. Hinton, G., et al. (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition. *IEEE Signal Processing Magazine*, 29(6).

20. Mager, M., et al. (2020). A Morphological Analyzer for Shipibo-Konibo. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*.

21. Conneau, A., et al. (2020). Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of ACL*.

22. Pires, T., et al. (2019). How Multilingual is Multilingual BERT? In *Proceedings of ACL*.

23. Maddieson, I. (1984). Patterns of Sounds. Cambridge University Press.

24. Palmer, A., et al. (2009). The Cherokee National Corpus: A Collaborative Language Resource. In *Proceedings of the 7th Workshop on Asian Language Resources*.

25. Anastasopoulos, A., & Neubig, G. (2019). Pushing the Limits of Low-Resource Morphological Inflection. In *Proceedings of EMNLP-IJCNLP*.

26. Schwartz, L., et al. (2019). Green NLP: A Methodology for Assessing the Environmental Impact of NLP Models. *arXiv preprint arXiv:1912.02160*.