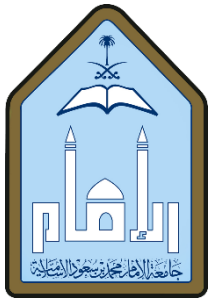
المملكة العربية السعودية

وزارة التعليم

جامعة الإمام محمد بن سعود الإسلامية

معهد تعليم اللغة العربية

قسم علم اللغة التطبيقي



التَّوْسِيمُ النَّحْوِيُّ لِلْمُدَوَّنَاتِ العَرَبِيَّةِ: نَمَاذِجُ تَوْسِيمِيَّةٌ مُقْتَرَحَةٌ

رسالة مقدمة لنيل درجة الدكتوراه في اللغويات التطبيقية

إعداد الطالبة:

أفراح بنت عبد العزيز التميمي

المشرف الأساسي:

أ. د. محمد يوسف حبلص

الأستاذ في علم اللغة

بجامعة الإمام محمد بن سعود الإسلامية

المشرف المساعد:

د. عبد المحسن الثبيتي

أستاذ البحث المشارك بالمركز الوطني لتقنية

الذكاء الاصطناعي والبيانات الضخمة

العام الجامعي:

١٤٣٩-١٤٤٠ هـ

٢٠١٨-٢٠١٩ م

مستخلص الدراسة بالعربية

عنوان الرسالة: التَّوْسِيمُ النَّحْوِيُّ لِلْمُدَوَّنَاتِ العَرَبِيَّةِ: نَمَاذِجُ تَوْسِيمِيَّةٌ مُقْتَرَحَةٌ

الباحث: أفراح بنت عبد العزيز بن حمد التميمي. **المشرفان:** أ. د. محمد حبلص ود. عبد المحسن الثبيتي

الدرجة العلمية: دكتوراه. **الجامعة والكلية:** جامعة الإمام محمد بن سعود الإسلامية–معهد تعليم العربية

القسم والتخصص: علم اللغة التطبيقي **العام الجامعي:** ١٤٣٩–١٤٤٠

يهدف هذا البحث إلى جمع مدونة لغوية شاملة ومتوازنة تتخذ من الإطار النموذجي للمدونة العربية (المدونة اللغوية العربية لمدينة الملك عبد العزيز للعلوم والتقنية) KACSTAC إطارا لمدونة من ١٠ آلاف كلمة تقطع وفق قائمة من متغيرات التقطيع التي يعتمد عليها التوسيم النحوي؛ لتقديم نماذج توسيمية نحوية، منطلقة من قواعد النحو العربي، صالحة للتوسيم الآلي، تدرب على جزء منها إحدى خوارزميات تعلم الآلة، ثم يختبر نظاما التقطيع والتوسيم النحوي على المتبقي منها.

وقد اتبعت الباحثة المنهج الوصفي التحليلي حيث وصفت أنظمة التوسيم بشكل عام، وما هو موجود في المدونات العربية الموسومة حاليا، مقترحة نماذج للوسوم النحوية (أقسام الكلام) تنطلق من نظرية تمام حسان في تقسيم الكلام. ثم طبقت هذه النماذج على المدونة التي جمعتها بتقطيع المدونة وتوسيمها يدويا. كما اتبعت الباحثة منهج تعليم الآلة الاحتمالي في تدريب خوارزمية الحقول العشوائية المشروطة Conditional Random Fields (CRF) على مدونة الدراسة (مدونة التدريب)، ثم اختبار نظامي التقطيع والتوسيم النحوي على المتبقي من المدونة (مدونة الاختبار).

وقد توصلت الباحثة إلى تقديم ثلاث مجموعات وسوم: مجموعة وسوم رئيسة، ومجموعة وسوم فرعية، ومجموعة وسوم موسعة. ثم زودت الباحثة نظامي التقطيع والتوسيم النحوي بها، وقاست أداءهما على مدونة الاختبار بعد تدريبه على التقطيع والتوسيم مع مجموعات الوسوم الثلاثة وظهرت نتائج الصحة في التقطيع ٠,٩٩ وفي مجموعة الوسوم الرئيسة ٠,٩٢ وفي مجموعة الوسوم الفرعية ٠,٨٢ وفي مجموعة الوسوم الموسعة ٠,٧٢، ثم زادت من حجم مدونة التدريب لتحسين أدائهما، وقد أظهرا تحسنا طفيفا.

مستخلص الدراسة بالإنجليزية

Title of Thesis: POS tagging for Arabic Corpus: Suggestion of Annotation Models
Name of Student: Afrah Abdulaziz Hamad Altamimi
Supervisors: Prof. Mohammed Hablas & Dr. Abdulmohsen Al-Thubaity
University & College: Al-Imam Mohammad Ibn Saud Islamic University- Arabic Language Teaching Institute
Department: Language of Preparation **Branch/Track**: Applied Linguistics
Degree: Decorate **Academic Year**: 2018/2019

The aim of this research is to collect a comprehensive and balanced corpus, based on the sample frame of the Arabic Corpus (KACSTAC). The study corpus consists of 10,000 words segmented according to a list of linguistic variables on which grammatical tagging based. This is in order to produce grammatical tagging models driven by Arabic grammar rules and valid for the automatic tagging. Part of the intended corpus is trained with machine learning algorithms, and the remaining of the .corpus is tested to evaluate the systems of grammatical segmentation and tagging

The researcher adopted a descriptive and analytical method. She reviewed tagging systems in general, especially those currently existing in the tagged Arabic corpora. Afterwards, she suggested models of parts-of-speech tags, which are based on Tammam Ḥassan's. These models are applied on the study corpus that was segmented and tagged manually. The researcher also used the probabilistic machine learning. This was made by training the algorithms of conditional random fields (CRF) in the training corpus, and by testing the system of grammatical segmentation and tagging in the reset of the .corpus (test corpus)

She made up three tagsets: main tagset, sub-main tagset and extended tagset. These three tagsets were provided by the system of grammatical segmentation and tagging. The performance of the systems was measured in the test corpus after the stage of training, and the accuracy of segmentation was 0.99. Regarding the main tagset, sub-main tagset and extended tagset, the accuracy was 0.92, 0.82 and 0.72 respectively. the training corpus was increased for improving the performance of systems. and this showed a slight improvement

فهرس المحتويات