المجلة الدولية للهندسة وتقنية المعلومات
journal homepage:www.ijeit.misuratau.edu.ly
IJEIT
جامعة مصراتة MISURATA UNIVERSITY

استخدام نحو اللغة العربية الوظيفي في تحسين استخراج المعلومات من النص العربي غير المهيكل

حسين خليل خليل
جامعة مصراته-كلية تقنية المعلومات، قسم علوم الحاسوب، مصراته، ليبيا
Hussein.khalil@it.misuratau.edu.ly

محمد الحسين مليطان
جامعة مصراته- كلية الآداب، قسم اللغة العربية، مصراته، ليبيا
mhmlitan@gmail.com

محمد علي بادي
جامعة مصراته-كلية تقنية المعلومات، قسم الشبكات والاتصالات، مصراته، ليبيا
Mak.badi@it.misuratau.edu.ly

***الملخص*— ينشر – يومياً - كم هائل من المعلومات على شبكة الإنترنت في أشكال متعددة، وثائق، مدونات...إلخ عبر شبكات التواصل الاجتماعي. وبما أن معظم هذه البيانات متاحة في شكل وثائق غير منظمة، فإنه من الصعب البحث في هذا الكم الهائل من البيانات والوصول إلى نتائج محددة ومفيدة فضلا عن إنفاق الوقت الطويل في ذلك. وبالتالي، فإن البحث والوصول إلى المعرفة المفيدة من هذه البيانات والنصوص غير المهيكلة تتزايد أهميته كل يوم.**

يركز هذا العمل على استخلاص المعلومات من النص العربي غير المهيكل الموجود على شبكة الانترنت وهي مهمة صعبة خاصة إذا كنا نروم الوصول إلى معلومات عن طريق التحليل للمستويات الصرفية-النحوية والدلالية والتداولية للغة العربية. ومما يزيد من صعوبة المهمة وتعقيد المشكلة عدم وجود بحوث متقدمة في هذا المجال بالنسبة للغة العربية بالمقارنة مع اللغات الأوروبية مثلا.

الهدف الرئيسي من هذا العمل البحثي هو تقديم منهجية شاملة لاستخراج المعلومات من النصوص العربية غير المهيكلة الموجودة على شبكة الانترنت باستخدام تقنيات معالجة اللغات الطبيعية التي ثبتت فعاليتها في حل معظم مشاكل التصنيف من أجل تحسين عملية استخراج المعلومات من النصوص والوثائق غير المهيكلة على الانترنت.

تكمن أهمية أدوات هذه البرمجة في كونها تلعب دورا رئيسيا في تيسير الوصول إلى تعيين ووسم المفردات بناء على دلالتها في النصوص غير المهيكلة، لنتمكن بعد ذلك من الوصول إلى الغاية الكبرى وهو الويب الدلالي. ويعرض هذا العمل نهجا قائما على قواعد جديدة – على مستوى معالجة اللغة العربية حاسوبيا - باستخدام تقنيات علم اللسانيات، وتحديدا نظرية النحو الوظيفي لسيمون ديك، متخذا من المستوى الدلالي للغة موضوعا لهذا البحث، كما يسهم بحثنا في تطبيق نهج جديد فيما يتعلق باستخراج العلاقات من النص العربي غير المنظم في المجال الاقتصادي التي تعتبر واحدة من أهم البيانات غير المهيكلة التي يتم نشرها يوميا على الانترنت، بالاعتماد على نظرية النحو الوظيفي.

الكلمات المفتاحية: **استخراج المعلومات، العلاقات بين الكائنات، النحو الوظيفي، البيانات الغير مهيكلة، معالجة اللغات الطبيعية.**

1. المقدمة

يزداد حجم المعلومات المنشورة على شبكة الإنترنت طرديا وبسرعة هائلة مع زيادة عدد مستخدمي الإنترنت، ووفقا للإحصاءات العالمية للإنترنت فقد تجاوز عدد مستخدمي الإنترنت 5.168780607 عند كتابة هذه الورقة. ومن المعلوم لدى مستعملي شبكة الانترنت أن معظم المعلومات المنشورة هي معلومات غير مهيكلة، مما يزيد أهمية الحاجة إلى النظم التي يمكن أن تكمل مهمة استخراج المعلومات المفيدة من الوثائق غير المنظمة بحيث أصبحت ضرورية أكثر من أي وقت مضى، وهذا ما ساهم في تطوير استخراج المعلومات (Information Extraction).

ومع ذلك، فإن التقدم في مجال البحث والتطوير في نظم اللغة العربية ليس ملحوظا مقارنة باللغات الأوروبية وهو ما ينعكس في ندرة الموارد العربية المتاحة لمعالجة اللغات الطبيعية والأدوات (Natural language processing). الشبكة العنكبوتية العالمية، أو ببساطة الويب، تمثل واحدة من أكبر مصادر المعلومات في العالم. المعلومات على شبكة الإنترنت تأتي على أشكال وأنواع مختلفة مثل الوثائق النصية والصور ومقاطع الفيديو. بالرغم من ذلك، فإن استخراج المعلومات ذات الفائدة دون مساعدة بعض أدوات الويب ليست عملية هيّنة، وهنا يأتي دور التنقيب في الويب والذي يحتاج بالضرورة إلى توفير أدوات تساعدنا على استخراج المعلومات المفيدة من البيانات، وأصبح من الصعب على نحو متزايد على الإنسان تحليل محتويات النص يدويا (manually).

البيانات الاقتصادية تعتبر واحدة من أهم البيانات التي يتم نشرها يوميا على الانترنت، وهذه البيانات غير المهيكلة تعتبر مصدرا هاما للمعلومات من أجل الاستفادة منها في العديد من التطبيقات الاقتصادية. استخراج البيانات من المعلومات الاقتصادية تعتبر عملية معقدة بسبب احتوائها على معلومات كثيرة. وبالتالي فإن استخدام الطرق التقليدية من أجل استخراج المعلومات من هذه البيانات والحصول على نتائج جيدة تعتبر عملية مكلفة للوقت والجهد عند التعامل مع ذلك الكم الهائل من البيانات؛ لهذا فإننا نحتاج إلى أدوات وتقنيات من أجل التعامل مع هذا الكم الهائل من البيانات على الانترنت لاستخراج المعلومات المفيدة.

هناك العديد من البحوث التي سعت لاستخراج المعلومات المهيكلة من النصوص التي تحتوي على اللغات البشرية عن طريق معالجة اللغة الطبيعية (Natural Language Processing) وكذلك هناك العديد من التطبيقات التي طُورت باستخدام تقنية معالجة اللغة الطبيعية ومنها على سبيل المثال: استرجاع المعلومات و الترجمة الآلية والمجيب الآلي والتعرف على مكونات العبارة اللغوية والعلاقات بين هذه المكونات وهذه جميعا تعد فرعا من فروع اهتمامات معالجة اللغات الطبيعية التي تدرج عادة تحت فئة الذكاء الصناعي يوجد العديد من الخوارزميات التي استخدمت في التعرف على مكونات العبارة اللغوية (Named Entity Recognition) والعلاقات بينها (Relation Extraction) على سبيل المثال خوارزميات التعلم الآلي (Machine Learning) والنهج القائم على قواعد اللغة في استخراج البيانات. وهناك من جمع بين النهجين في الطريقتين السابقتين . في هذه الورقة، نقدم مقاربة لاستخراج العلاقات بين مكونات العبارة اللغوية في النص غير المهيكل، وتحقيقا للهدف من هذه الدراسة في التعرف على الدلالة في اللغة العربية (المستوى الدلالي) وليس التعرف على تركيبها (المستوى التركيبي)، فإننا نقترح الاعتماد على منهج لساني جديد قائم على مبادئ تداولية في وصف وتفسير اللغات الطبيعية، وبشكل محدد فإننا نعتمد أطروحات نظرية النحو الوظيفي، الذي سنقدم له تعريفا موجزا في الفقرات التالية.

استلمت الورقة بالكامل في 18 يونيو 2021 وروجعت في 30 أغسطس 2021 وقبلت للنشر في 13 أكتوبر 2021

ونشرت ومتاحة على الشبكة العنكبوتية في 27 اكتوبر 2021

2. دراسات سابقة

على الرغم من محدودية الجهود البحثية في معالجة النص العربي غير المهيكل من أجل استخراج المعلومات المفيدة وبخاصة طبيعة العلاقات بين مكونات العبارة اللغوية في النص العربي غير المهيكل مقارنة باللغات الأخرى على سبيل المثال اللغة الانجليزية واللغات الأوربية الأخرى فإن هناك العديد من الجهود البحثية التي ركزت على استخراج العلاقات من النص العربي غير المهيكل ويمكن إرجاع هذه الجهود إلي ثلاث اتجاهات رئيسة هي: النهج القائم على القواعد التركيبية اللغة (Rule-based approach). والنهج القائم على التعليم الآلي (Machine Learning Approach) والنهج القائم على الجمع بين المنهجين السابقين (Hybridapproach).

أولا/ النهج القائم على القواعد التركيبية Rule-based approach

يعتمد هذا النهج على بناء أنماط تركيبية لاستخراج العلاقات بينها، باستخدام القواعد التركيبية للغة التي تصف هذه الأنماط. ويتم ترميز هذه الأنماط اللغوية اعتمادا على التنميط اللساني المسبق الذي أعده لسانيون كمجموعة من القواعد. وهذه النهج يستخدم المعاجم/القواميس لبناء القاعدة. وعادة ما تنفذ القواعد في شكل تعبيرات منتظمة وقواعد إجبارية لتحديد العلاقات.

كما استخدمت بعض الدراسات (حمدو وهيلا) تطبيق NOOJ للوصول إلى معرفة العلاقات بين مكونات العبارة اللغوية اعتمادا على القواعد التركيبية للغة، والتكوين المعجمي لنطاق المشكلة (specific domain) لاسيما العلاقة بين أسماء الأشخاص والذوات أو المفاهيم الواردة في العبارة اللغوية.

كما طورت مجموعة من الباحثين (فاطمة علي وآخرون 2012) نموذجا لاستخراج العلاقات المرتبطة بالظروف المكانية في النصوص العربية، ويمكن تلخيص هذه المقاربة بأنها مجموعة من القواعد التي استخدمت لاستخراج العلاقات التي ترتبط بالظروف المكانية داخل النص العربي، وتم تطبيق هذا النموذج على عبارات (جملية) خاصة.

وهناك فريق آخر من الباحثين (صادق و مزيان (2016)) اقترح نموذجا لاستخراج العلاقات السببية بين مكونات العبارة اللغوية التي تم التعبير عنها صراحة في النصوص العربية. وقد تضمن هذا النموذج حوالي 700 نمط لغوي يسمح بتمييز أجزاء الجمل التي تمثل السبب والمسبب.

ثانيا/ النهج القائم على التعليم الآلي (Machine learning approach)

استند هذا النهج على مجموعة من الخصائص أو العلامات التي يمكن أن تشمل خصائص أو علامات نحوية أو دلالية أو معجمية، وقد تم استخدام نهج التعلم الآلي بنجاح لاستخراج العلاقة من النص غير المنظم وهذه الطريقة تتطلب إعدادا كاملا لكل البيانات التي سيتم تدريب الحاسوب عليها (training corpus)

قام فريق بحث (بن يحيى وآخرون) باستخدام التعليم الآلي لاستخراج العلاقات بين مكونات العبارة اللغوية في النصوص العربية، واعتمد هذا المنهج على مجموعة من الكلمات الموجود في قاعدة البيانات (Ontology) من أجل توليد مجموعة من بيانات التدريب (training corpus). قدم باحثون (محمد وآخرون) رؤية جديدة لاستخراج العلاقة الدلالية باستخدام المراقبة البعيدة (distant supervision approach) التي تعتمد على استخدام الدبيديا (DBpidia) كمصدر البيانات لبناء بيانات التدريب، وتحصل الباحثون في هذا البحث على نتائج مرضية في اكتشاف 97 نوعا للعلاقة بين مكونات العبارات اللغوية التي تمت معالجتها.

ثالثا/النهج القائم على الجمع بين النهجين السابقين

قدم باحثان (أنس وسلمى) نهجا يستخدم نهجا هجينا يعتمد على النهج القائم على قواعد اللغة والنهج القائم على التعليم الآلي. يهدف إلى تحديد موقع الكلمة التي تعبر عن العلاقة الثنائية بين مكونات العبارة اللغوية. هذه النهج يهدف إلى استخراج العلاقات بين الأشخاص والأماكن والهيئات.

كما قام فريق بحث آخر (بن باشا وزويجنباوم، 2011) باستخراج العلاقات بين الأمراض وعلاجها، واعتمد هذا البحث على الجمع بين النهج القائم على القواعد والنهج القائم على التعليم الآلي، حيث تم بناء مجموعة من الأنماط يدويا من مجموعة البيانات التدريبية للتعبير عن النهج القائم على القواعد، أما فيما يخص النهج القائم على التعلم الآلي فقد تم استخدام خوارزمية شعاع الدعم الآلي (Support vector machine Algorithm) لتحليل البيانات من أجل تصنيفها بالاعتماد على المستوى المعجمي، والصرف-التركيبي، والسمات الدلالية، وتظهر النتائج التي تم الحصول عليها تحسنا في كفاءة النهجين معا.

من خلال المراجعة السابقة للعديد من الجهود التي بذلت لاستخراج العلاقات من النص العربي يمكن استنتاج أن منهج الجمع بين النهجين يعكس تحسنا كبيرا في أساليب استخراج البيانات بناء على المنهج القائم على قواعد اللغة وأساليب التعلم الآلي. في هذا الصدد، تحتاج الأساليب القائمة على القواعد إلى استغلال المناهج اللسانية المتطورة في البحث في اللغة العربية لحل المشاكل المتعلقة بتعقيدات اللغة العربية، وليس فقط في على المستوى التركيبي ولكن أيضا على المستويين الدلالي والتداولي. على سبيل المثال، تستخدم معظم الأنظمة المبنية على قواعد اللغة العربية القواعد الأساسية لتحديد أنماط العلاقة اللغوية، وذلك باستخدام السمات الأساسية الثلاثة للغة العلاقة: فعل (Verb) فاعل (Subject) مفعول (Object)

على الرغم من الفائدة التي تقدمها الدراسات البنيوية في استخدام القواعد التركيبية كمجموعة مجردة من القواعد النحوية المفصولة تماما عن سياق الاستخدام فإن العبارات اللغوية غالبا ما تحتوي على علاقات أخرى معقدة ولهذا فإن استخدام النحو الوظيفي إطارا نظريا يساعد في فهم أعمق وأوضح للخصائص الدلالية والتداولية للغة.

ومن هنا، نعتقد أن نظرية النحو الوظيفي تقدم مقترحات أكثر مرونة لنمذجة العلاقات الدلالية المعقدة في اللغة العربية، ولهذا تم اعتماد نظرية النحو الوظيفي إطارا منهجيا لهذه الدراسة لمحاولة استخراج العلاقة الدلالية في النص العربي. تركز هذه الدراسة بشكل خاص على المقالات المتعلقة بالأخبار في سوق الأوراق المالية، والتي تحتوي على عبارات تتضمن علاقات دلالية معقدة من شأنها أن تجعل استخراج العلاقات عملية ليست سهلة.

3. نظرية النحو الوظيفي

عرفت نظرية النحو الوظيفي منذ ظهورها على يد اللساني الهولندي سيمون دك (Simon Dik) سنة 1978 عددا من النماذج المتعاقبة قامت على ما يشبه مبدأ "النشوء والارتقاء"، حيث شهد كل جيل من هذه النماذج تطورا على النموذج السابق مستفيدا من القصور في الوصف والتفسير الكافيين في سلفه. قصور كان في أغلبه متمثلا في بنية النحو الإجرائية دون أن يطال المبادئ الرئيسة التي قام عليها. فالجهاز الواصف في نظرية النحو الوظيفي مرّ، من حيث بناؤه وصياغته، بمراحل أفرزت عددا من النماذج من أهمها النموذج النواة، والنموذج المعيار، ونموذج محو الطبقات القالبي، ونموذج نحو الخطاب الوظيفي، ونموذج نحو الخطاب الوظيفي الموسع، وقد تناولتها كتب الأستاذ أحمد المتوكل بالشرح والتوضيح وما أضافه كل نموذج على سابقه ما يستوجب الاكتفاء هنا بالعرض الموجز لأهم ثلاثة نماذج وهي النموذج النواة والنموذج المعيار ونموذج نحو الخطاب الوظيفي.

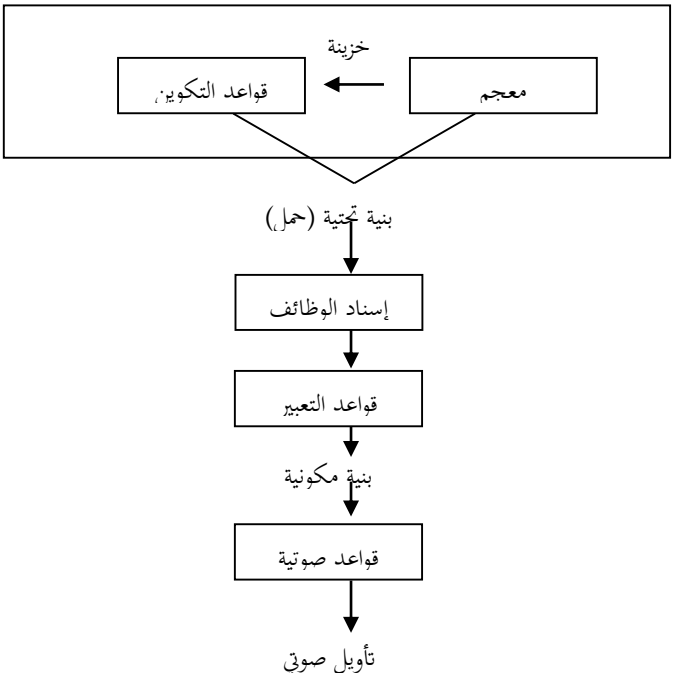
أ. النموذج النواة، سيمون دك (Simon Dik) 1978

عرض اللساني الهولندي سيمون دِكْ (Simon Dik) أوّل نماذج نظرية النحو الوظيفي في كتابه ((Functional Grammar (1978)). وكان لبنة أولى في صياغة النماذج التي تلته على أساس ما توصلت إليه نظرية النحو الوظيفي في سعيها نحو إحراز الكفايات التداولية والنمطية والنفسية.

يتكون هذا النموذج من أربعة مكونات مرتبة حسب آلية اشتغالها: خزينة ثم قواعد إسناد الوظائف ثم قواعد التعبير ثم القواعد الصوتية. تتكون الخزينة من شقين اثنين: معجم يضم المفردات الأصول وقواعد تكوين مسؤولة عن اشتقاق المفردات الفروع (كأفعال الانعكاس وأفعال المطاوعة وأسماء الفاعلين وغيرها) من المفردات الأصول.

تشكل البنية التحتية حملا تحدّد فيه كل الخصائص الدلالية المؤشر لها في شكل مخصّصات وسمات (جهية، زمنية...) ووظائف دلالية. ويُنقل الحمل إلى بنية وظيفية تامة التحديد عن طريق إسناد وظيفتي الفاعل والمفعول ثم إسناد الوظيفتين التداوليتين المحور والبؤرة.

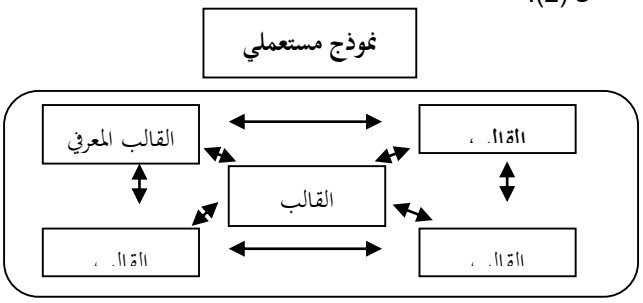
تتخذ البنية الوظيفية دخلاً لقواعد التعبير وهي مجموعة القواعد المسؤولة عن تحديد الخصائص الصرفية والتركيبية (الرّتبيّة) والتطريزية (النبرية والتنغيمية) على أساس ما يُورَد في البنية الوظيفية. يأخذ خرج قواعد التعبير شكل بنية مكوّنية تنقل بواسطة القواعد الصوتية إلى تأويل



الشكل (1) يوضح اشتقاق العبارة اللغوية في النموذج النواة

ب. النموذج المعيار، سيمون دك (Simon Dik) 1989

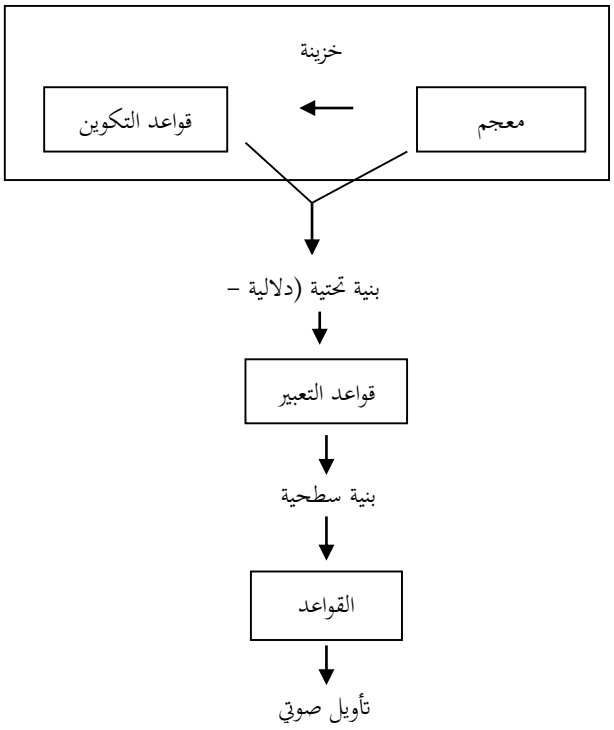
يتصدر "نموذج مستعملي اللغة الطبيعية" أهم ما أتى به النموذج المعيار، ويقصد به المنتج للخطاب والمتلقي للخطاب والمؤول له، هذا النموذج وضع لرصد الملكات التي تتوافر لدى المتكلم/المخاطب، حصر منها سيمون دِكْ (Simon Dik) خمس ملكات على أساس أنه يمكن إضافة ملكات أخرى إليها، وتقوم خمسة قوالب برصد ملكات القدرة التواصلية الخمس وتتفاعل فيما بينها على أساس أن كل قالب يتمتع باستقلال مبادئه وإولياته لكنه يشكّل دخلا – خرجا لباقي القوالب حسب الشكل (2):



الشكل (2) يوضح نموذج مستعملي اللغة

كانت الخصائص الممثل لها في النموذج النواة محصورة في الوظائف التداولية الخارجية والداخلية، وفي النموذج المعيار أصبحت تشمل السمات الإنجازية والوَجهية، كما أضيفت إلى وظيفتي المبتدأ والذيل وظائف خارجية أخرى صنفت إلى فواتح ونواقل وحوافظ وخواتم باعتبار دورها في تنظيم بنية الخطاب كما فُرّعت وظيفة المحور إلى محور معطى ومحور معاد ومحور جديد ووظيفة البؤرة إلى بؤرة جديد وبؤرة تعويض وبؤرة قصر وبؤرة انتقاء.

هذه الإضافات التداولية لا تغير، رغم أهميتها، من قانون الاشتقاق المعتمدة في النموذج الأول إذا استثني أن البنية التحتية لم تعد مجرد حمل بل بنية تامّة التحديد من حيث الدلالة والتداول معاً كما يتبين من الشكل (3):



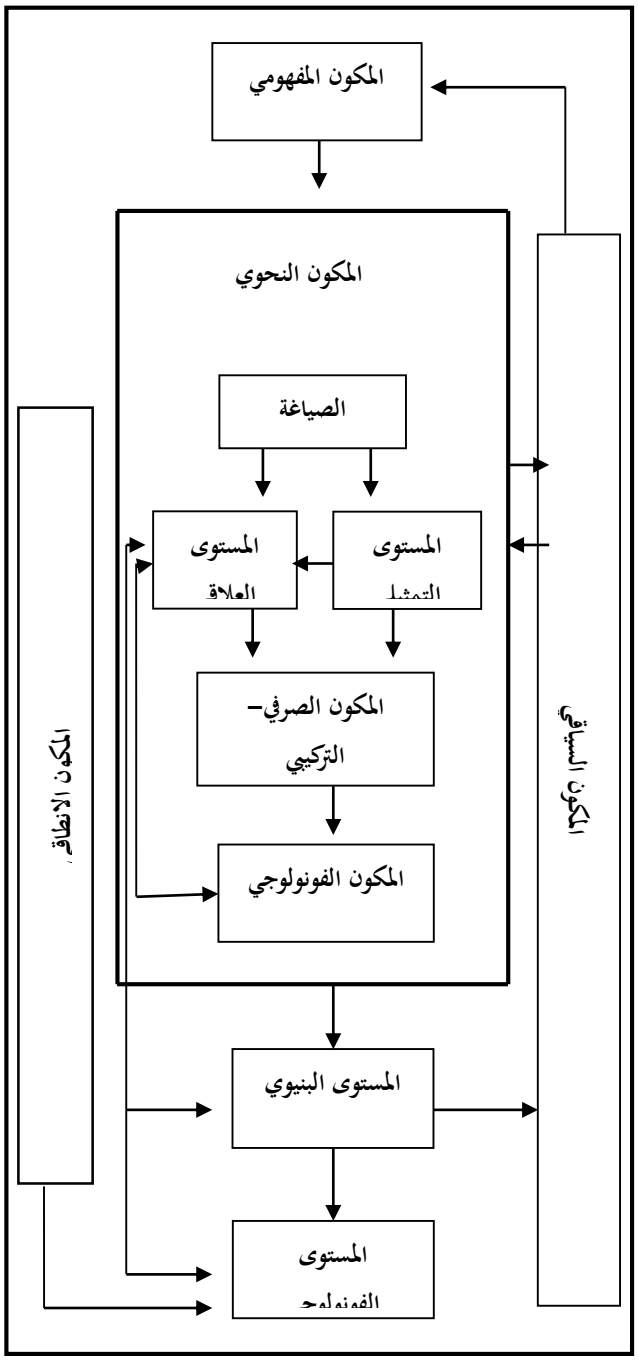
الشكل 3 يوضح النموذج المعيار

ت. نموذج نحو الخطاب الوظيفي، هنخفلد (Hengeveld) وماكنزي (Mackenzie) 2008

سعت نظرية النحو الوظيفي إلى تحصيل الكفاية النفسية منذ نشأتها إلى جانب إحراز الكفايتين التداولية والنمطية، هذا السعي تمثل في إخضاع قواعد النحو للواقعية النفسية حيث أقصيت القواعد التحويلية باعتبارها لا تطابق إواليات إنتاج الخطاب ولا إواليات تأويله. الجهاز الواصف في نحو الخطاب الوظيفي اشتمل على أربعة مكونات هي: "المكون المفهومي" (أو المعرفي) و"المكون النحوي" و"المكون الإصاتي" و"المكون السياقي".

- يرصد المكون المفهومي المعارف اللغوية وغير اللغوية كما يرصد قصد المتكلم من الخطاب المزمَع إنتاجه. ويعدّ هذا المكون "القوة الدّافعة" بالنظر إلى المكونات الأخرى؛
- تُحدَّد خصائص الخطاب في المكون النحوي في ثلاثة مستويات: مستوى علاقي (تداولي) ومستوى تمثيلي (دلالي) ومستوى بنيوي. المستويان العلاقي والتمثيلي خرجان لإوالية "الصياغة" التي تمثل للخطاب في المستوى الأول في شكل فعل خطابي يتضمن فحوى قضويا قوامه فعل إحالي وفعل

- حملي ولخصائص الخطاب الدلالية في المستوى الثاني . وتتكفل إوالية قواعد التعبير بنقل المستويين العلاقي والتمثيلي إلى مستوى بنيوي تحدّد فيه الخصائص الصرفية – التركيبية والخصائص الصوتية.
- خرج قواعد التعبير بنية صرفية-تركيبية – صوتية مجرّدة يضطلع المكون الإصاتي بإنطاقها في شكل عبارة لغوية محققة.
- المكون السياقي محط رصد العناصر المقامية والمقالية التي تواكب إنتاج الخطاب ويقوم بدور الربط بين المكونات الثلاثة الأخرى. الشكل (4) يوضح مكونات جهاز نحو الخطاب الوظيفي وطريقة اشتغاله:



الشكل (4) يوضح مكونات جهاز نحو الخطاب الوظيفي وطريقة اشتغاله

نحو الخطاب الوظيفي يختلف عن النحوين الآخرين (النموذج المعيار ونحو الطبقات القالبي) من حيث القيام بعمليتين اثنتين: عملية اختزال وعملية إضافة، حيث اختزل فحوى القالبين الاجتماعي والإدراكي في مكون واحد هو المكون السياقي الذي يرصد السياق العام والسياق المقامي (سياق الموقف التواصلي الآني) والسياق المقالي (الخطاب السابق والخطاب اللاحق اللذين يشكّلان بُعد "التّناص" للخطاب المنتج أو المروم فهمه)، واختزل فحوى القالبين المعرفي والمنطقي في مكون واحد يرصد فيه كل ما يفترض أن يشكّل "القوة الدافعة" في إنتاج الخطاب أو فهمه.

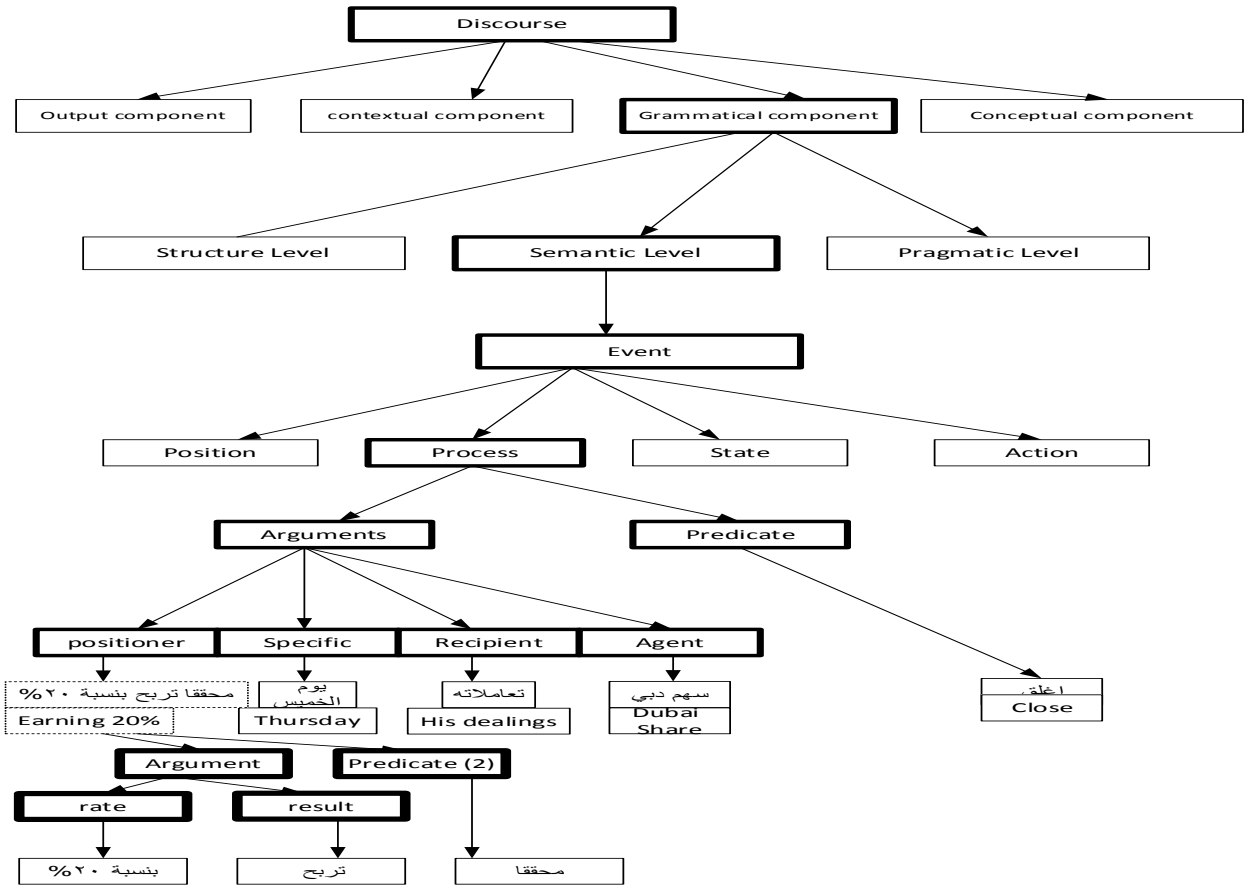
أمّا الإضافة فتخص "المكون الإنطاقي" الذي يقوم بنقل المستوى الفونولوجي المجرّد إلى تحقّق فعلي قد يكون صوتا أو خطّاً (في الخطاب المكتوب) كما يمكن أن يكون إشارة.

4. النظام المقترح

من الأهداف التي تسعى النظريات اللسانية إلى الوصول إليها ما يعرف بـ"الكفاية الحاسوبية"Computational adequacy وبلوغ هذه الكفاية يمكّن من التأكد من مدى دقة وصحة صورنة النظرية اللسانية، ومن خلال ما تم عرضه سابقا نلحظ أن معظم الجهود التي بذلت سابقا اهتمت بالحوسبة البنيوية للغة، وتحديدا المستوى التركيبي للجملة (فعل، فاعل، مفعول) دون الاهتمام بشكل كبير بالمستويات الأخرى للغة حسب التقسيم اللساني المعروف.

ومن المعلوم لدى اللسانيين أن مستويات التحليل اللغوي هي الصوتي (Phonetic)والإصاتي (Phonological) والصرفي (Morphological) والتركيبي (syntactic) والدلالي (Semantic) والتداولي (Pragmatic) وتنقسم النظريات اللسانية بشأنها إلى فرعين رئيسيين، قسم يرى أن الأهمية للمستويات البنيوية (التركيب) وهي التي تحكم ما عداها، وفريق يرى أن التداول والدلالة يحكمان البنية وهم أصحاب التيار المعروف بالتداوليين أو الوظيفيين، وهو الاتجاه الذي تتبناه هذه الورقة ليكون إطارا نظريا لها، ولهذا فإن هذه الورقة تقترح مقاربة تحاول فيها تقديم تجربة لحوسبة المستوى الدلالي – وليس التركيبي – للعبارة اللغوية في الخطاب الاقتصادي، بناء على ما تقترحه النظرية اللسانية المعروفة بنظرية النحو الوظيفي لسيمون دك والتي نقلها للعربية وساهم فيها وقام بتحليل الخطاب في اللغة العربية وفق هذه النظرية اللساني المغربي أحمد المتوكل.

ومن المعلوم أيضا أن المستوى الدلالي (Semantic level) داخل نظرية النحو الوظيفي يقوم بتحديد الخصائص الدلالية للخطاب. ويتضمن طبقتين اثنتين: طبقة عُليا هي "القضية Proposition " وطبقة سفلى هي "الواقعة State of affairs ". ما يهمنا في هذه المقاربة هي طبقة الواقعة فهي مكان التمثيل للواقعة وللذوات المشاركة فيها مشاركة رئيسية أو مشاركة فرعية، مشاركة وجوب أو مشاركة جواز. ويتم التمثيل للواقعة في شكل بنية حملية Predicate frame تتضمن محمولا Predicate (فعلا، أو اسما، أو صفة) وفئتين من الحدود Terms: حدود موضوعات Argument term رئيسية ضرورية وحدود لواحق Satellite term فرعية اختيارية تحمل جميعها وظائف دلالية داخل العبارة اللغوية. الشكل رقم 5يوضح معمارية نظرية نحو الوظيفي .



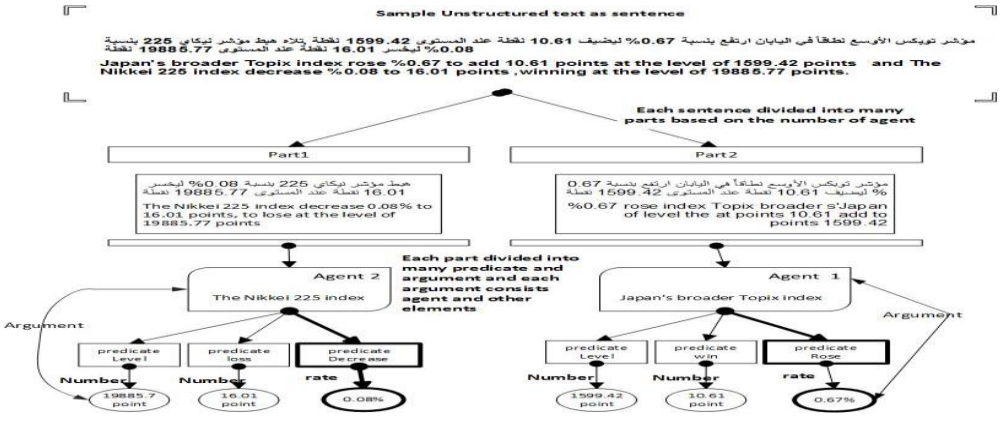
شكل (5) معمارية نظرية النحو الوظيفي

هذا الإطار النظري هو الذي تم اعتماده في هذه المقاربة التي حاولنا من خلالها الوصول إلى استخراج العلاقات الدلالية المعقدة في العبارة اللغوية العربية وتحديدا في الخطاب الصحفي الاقتصادي. تتوزع الوظائف الدلالية (Semantic functions) في نظرية النحو الوظيفي إلى (منفذ Agent ، متقبل Patient \ Goal ، مستفيد Beneficiary ، مستقبل Recipient،...) بناء على علاقتها بالمحمول الذي يمثل الواقعة.

وفي هذه المقاربة تم التعامل مع العبارات اللغوية بناء على هذا التصور، وتم الاهتمام باستخراج شبكة العلاقات الدلالية داخل العبارة اللغوية كما هو مبين في المثال التالي:

- ارتفع سهم الخليج بقيمة 10% في تداولات اليوم

حيث يمثل يقوم المحمول (Predicate) بتمثيل الواقعة ويتحقق هذا في المكون (ارتفع) ويمثل المنفذ (Agent) لهذه الواقعة العنصر الأساسي (سهم الخليج) بينما أسندت الوظيفة الدلالية النسبة (percentage) للمكون (بقيمة 10%) والوظيفة الدلالية الزمان (Tense) للمكون (في تداولات اليوم) الشكل رقم 6 يوضح آلية الخوارزمية المعتمدة



شكل 6: آلية الخوارزمية استخراج العلاقات باستخدام نظرية النحو الوظيفي

في المثال السابق ((مؤشر توبكس الأوسع نطاقاً في اليابان ارتفع بنسبة 0.67% ليضيف 10.61 نقطة عند المستوى 1599.42 نقطة، تلاه هبط مؤشر نيكاي 225 بنسبة 0.08% ليخسر 16.01 نقطة عند المستوى 19885.77 نقطة)

خوارزمية الحل المقترحة تتعامل مع العبارة في النص العربي التي تحتوي على علاقات معقدة تتمثل في عدد غير محدد من المشاركين في الواقعة الواحدة، فقد يكون للواقعة1 مثلا، منفذ1 ومستقبل1 ومتقبل1 ومنفذ2 ومستقبل2 وزمان2 ومنفذ ن ومستقبل ن وهكذا. وعلى هذا فإن خوارزمية الحل المقترحة تتعامل مع الجملة بتقسيمها إلى مجموعة من الوقائع التي تنتمي إلى قضية واحدة، كما هو مبين في الترسيمة التالية:

خطاب اقتصادي= (قوة إنجازية (قضية (واقعة (محمول، منفذ....))))

وعلى سبيل المثال فإن الخطاب الاقتصادي لسوق الأوراق المالية تضمن عبارات لغوية من قبيل:
(قوة إنجازية1 خبر (القضية1 (الواقعة 1 (محمول [ارتفع] منفذ/متموضع [مؤشر توبكس الأوسع نطاقاً] نسبة [بنسبة 0.67%] نتيجة1 [ليضيف 10.61 نقطة] نتيجة2 [المستوى 1599.42 نقطة])))
(قوة إنجازية2 خبر (القضية2 = الواقعة 2 (محمول [هبط] منفذ/متموضع [مؤشر نيكاي] نسبة [بنسبة 0.08%] نتيجة1 [ليخسر 16.01 نقطة] نتيجة2 [المستوى 19885.77 نقطة])

الاستراتيجية الرئيسية للخوارزمية تقوم على أن المحمول الرئيسي في الواقعة هو الذي يتحكم في عدد ونوع المشاركين في الواقعة والأدوار الدلالية لكل منها داخل هذه الواقعة.

وعلى سبيل المثال فإن المنفذ1 (Agent1) في العبارة هو "مؤشر توبكس الأوسع نطاقا"، الذي تم تحديده باعتباره المشارك الأول في الواقعة التي يمثلها المحمول1 (Predicate1) "هبط" والتي قد تكون فعلا / اسما / صفة. ويحدد المحمول أي موضوع (argument) محتمل في هذه العلاقة مثل ("10.61 نقطة") الذي يصف نتيجة العملية في العبارة لاستكمال تحديد العناصر الثلاثة في العلاقة. وعلاوة على ذلك، تم استخدام الخوارزمية لاستخراج العلاقات في الجملة من خلال مواءمتها مع العلاقات في قاعدة البيانات المعرفة.

المثال التالي يوضح طريقة استخراج العلاقات بين المكونات بناء على الوظائف الدلالية للعبارة اللغوية. الخطوات التالية توضح طريقة عمل النهج المقترح

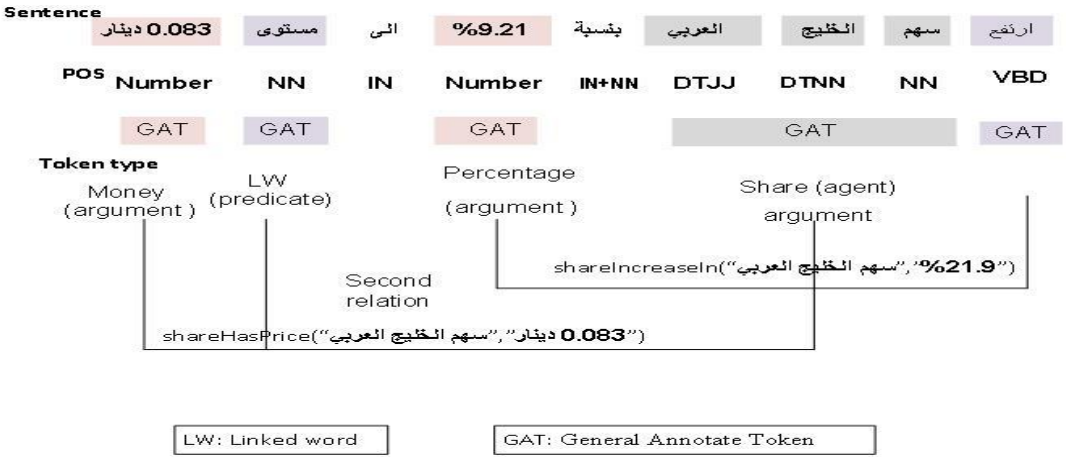
1- تحديد أسماء الكيانات المسماة وإعطاء اسم موحد لهذه الكائنات ((General Annotation Token (GAT وإضافة بعض الخصائص لكل كائن، على سبيل المثال: نوع الكائن و التصنيف والجدر وتخزين جميع الكيانات في مصفوفة تحت اسم NE_Array حسب موقع كل كائن في الجملة
2- قراءة المصفوفة NE_Array وتحديد أول منفذ في الجملة المراد معالجتها وعادة ما يكون اسما على سبيل المثال " المؤشر العام لسوق أبوظبي " وتحديد خصائصه.
3- البدء في قراءة البيانات الموجودة في المصفوفة من أول موقع في المصفوفة بالتتابع ومقارنة كل عنصر في المصفوفة مع وظيفة المنفذ في الخطوة 2 حسب الآتي:
   I. إذا كان أول عنصر في المصفوفة والذي نوعه GAT يحمل نفس خصائص وظيفة المنفذ الذي تم تعيينه في الخطوة 2 فحينئذٍ يتخطى الى العنصر الذي يليه
   II. إذا كان أول عنصر في المصفوفة والذي نوعه GAT يحمل وظيفة محمول (predicate) يتم تعينه كمحمول وقراءة خاصية الجذر.
   III. إذا كان في الجملة محمولين متتاليين يتم استبعاد الأول واعتماد الثاني
   IV. إذا تم تحديد المنفذ والمحمول سوف تعمل الخوارزمية على تحديد الجزء الثالث من العلاقة حسب نوع جذر المحمول على سبيل المثال:
      أ. إذا كان جذر المحمول (رفع، خفض) فإن العنصر الثالث من العلاقة سوف يكون نسبة مئوية
      ب. إذا كان جذر المحمول (فتح، أغلق) فإن العنصر الثالث من العلاقة سوف يكون قيمة
   V. إذا تم تصنيف مكونات العلاقة (المنفذ، المحمول، النتيجة) سيتم مطابقة هذه المكونات مع العلاقات الثلاثية التي تم استردادها من قاعدة المعرفة.
   VI. إذا كانت العلاقة الثلاثية المستخرجة من نص غير منظم تتطابق مع العلاقة في قاعدة المعرفة، فسيتم إضافة الميزات العديدة إلى العلاقة مثل (subject، object ، predicate)
4- اختيار المنفذ التالي في الجملة (المصفوفة)، إذا وجدت الخوارزمية أن المنفذ الجديد ويكون اسم علم، وليس له علاقة بالمنفذ الأول، فسيتم تعيين المنفذ الجديد كعامل جديد والعودة إلى الخطوة 2.

الشكلان 7 و 8 يوضحان آلية عمل خوارزمية الحل لاستخراج العلاقات:

أنهى المؤشر العام لسوق أبوظبي تعاملاته في جلسة يوم الخميس على هبوط واضح نسبته 1.24% بخسائر بلغت 62.43 نقطة، ليصل إلى مستوى 4.992.52 نقطة

	1	2	3	4	5	6	7	8	9
Agent	المؤشر العام لسوق أبوظبي								
Predicate		أنهى		هبوط		بخسائر		ليصل	
Argument			الخميس		1.24		62.43		4.992.52
Relation 1									
Relation 2									
Relation 3									
Relation 4									

يوضح الشكل 7 عملية استخلاص العلاقة من عبارة تحتوي على واقعة واحدة شارك في تنفيذها عدد من المشاركين



الشكل 8: استخلاص العلاقة من عبارة تحتوي على واقعة واحدة شارك في تنفيذها عدد من المشاركين

5. اختبار النظام

حتى وقت كتابة هذه الورقة، لم نتمكن من العثور على أي بحث منشور قام بتقييم جهود معالجة اللغة الطبيعية في استخراج العلاقات بين الكيانات المسماة العربية باستخدام النحو الوظيفي، وبالتالي لا يمكن مقارنة التقييم بالجهود المنشورة في هذا المجال.

في هذا البحث تم استخدام مجموعة الوثائق التي جمعها الباحثون من مواقع عربية مختلفة والتي سميت ARB_GEO Corpus لتنفيذ التجارب التي تهدف الى تقييم أداء المنهجية المقترحة باستخدام مجموعة ARB_GEO ، ولقد تم تقييم النظام باستخدام مقاييس التقييم الشائعة في هذا المجال والمتمثلة الدقة (precision) ، الاستدعاء (recall)، و مقياس F (F-measure) ، وقد حقق النظام نتائج جيدة عند الاختبار.

أجريت مجموعة من التجارب لتقييم أداء الخوارزمية المقترحة باستخدام مجموعة ARB_ECON التي تم جمعها من مواقع إخبارية اقتصادية عربية مختلفة.

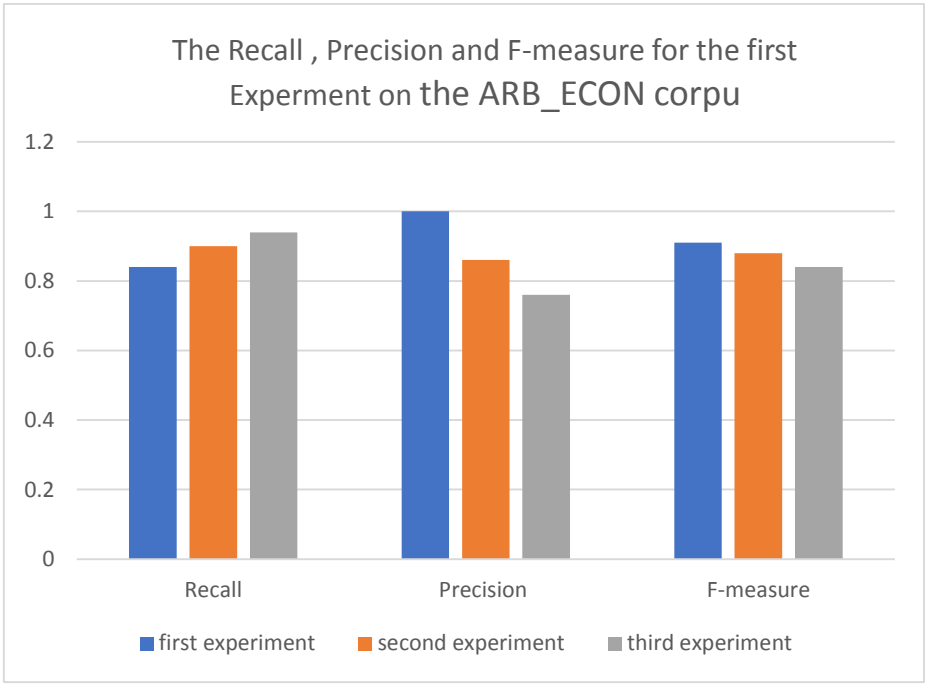
قسمنا اختبار الخوارزمية إلى ثلاث فئات: الأولى هي استخراج العلاقة من الجمل البسيطة، والثاني هو استخراج العلاقات من الجمل المعقدة، والثالث هو استخراج العلاقة من الجمل الأكثر تعقيدا.

التجربة الأولى هي استخراج العلاقة في العبارات البسيطة التي تحتوي على واقعة واحدة لها محمول واحد.

التجربة الثانية استخراج العلاقة من العبارات المعقدة التي تحتوي على أكثر من واقعة واحدة تضمنت محمولا واحدا رئيسيا وعدد من المحمولات الفرعية.

التجربة الثالثة استخراج العلاقة من عبارات أكثر تعقيدا تتضمن أكثر من واقعة وكل واقعة تتضمن أكثر من محمول رئيسي، وأكثر من محمول فرعي.

ويبين الشكل (9) التالي نتائج الخوارزمية استنادا إلى مجموعة التجربة



الشكل 9 :يوضح مقارنة بين نتائج التجارب الثلاثة

6. النتائج والتوصيات

- أظهرت الدراسة أن استخراج المعلومات (الدلالية تحديدا) حاسوبيا من النصوص العربية باستخدام تقنيات معالجة اللغات الطبيعية عملية معقدة جدا خاصة إذا ما نظرنا إلى خصوصية اللغة العربية وتراثها العريق والثرّي من جهة في مقابل أهمية تحديثها من جهة أخرى. ولكن على الرغم من التعقيدات التي واجهتنا فإنه بالإمكان حوسبة اللغة العربية دلاليا خاصة بعد أن عالجت اللسانيات الحديثة مشكلة الصورنة للنظام اللغوي العربي الذي صار من الممكن حوسبته.
- نظرية النحو الوظيفي الذي اختبرناها في هذه الدراسة أظهرت مرونة عالية وإمكانية لحوسبة اللغة العربية بشكل دقيق على المستوى الدلالي

ومن المفيد هنا أن نورد بعض التوصيات:

- العناية بالفروق الدلالية التي تسعف هندسة اللغة وإثراء حوسبتها بمستويات الدلالة وسياقات تعبيرها المجازية وسواها.
- العناية بتطوير الصورنة الدلالية للعربية باستخدام نظرية النحو الوظيفي وتوجيه الباحثين إلى المساهمة فيها.
- تفعيل ودعم العمل العلمي المشترك بين الجهود العلمية في مجالي اللغة العربية وعلوم الحاسب الآلي.

المراجع

[1] internetworldstats.com, "*Internet world stats*" [Online] Available:http://www.internetworldstats.com/stats.htm. [Accessed: 14-Dec-2017]

[2]-Moreno, Antonio, and Teófilo Redondo. "*Text Analytics: the convergence of Big Data and Artificial Intelligence*." IJIMAI 3.6 (2016): 57-64

[3]. Boujelben, S. Jamoussi and A. B. Hamadou, "*A hybrid method for extracting relations between Arabic named entities,*" Journal of King Saud University-Computer and Information Sciences, vol. 26, (4), pp. 425-440, 2014.

[4] Shaalan, K. (2010). Rule-Based Approach in Arabic Natural Language Processing. The International Journal on Information and Communication Technologies (IJICT), 3(3), 11-19

[5] A. B. Hamadou, O. Piton and H. Fehri, "*Multilingual Extraction of functional relations between Arabic Named Entities using NooJ platform*," 2010

[6] Fatma Ali Alnairia, Nazlia Omar and Mohammed Albared, (2012) "*Extraction of Spatial Relation in Arabic Text Using Rule-Based Approach*". In proceeding International Journal of Advancements in Computing Technology (IJACT) Volume4, Number15, September

[7] J. Sadek, F. MezianeExtracting Arabic causal relations using linguistic patterns ACM Trans. Asian Low-Resour. Lang. Inf. Process., 15 (3)(2016), pp. 14:1-14:20

[8]. Sarhan, Y. El-Sonbaty and M. A. El-Nasr, "*Arabic Relation Extraction: A Survey*," International Journal of Computer and Information Technology, vol. 5, (5), 2016

[9] M. Al-Yahya, L. Aldhubayi and S. Al-Malak, "*A pattern-based approach to semantic relation extraction using a seed ontology,*" in Semantic Computing (ICSC), 2014 IEEE International Conference On, 2014, pp. 96-99

[10] Mohamed, Reham, Nagwa M. El-Makky, and Khaled Nagi. "*ArabRelat: Arabic Relation Extraction using Distant Supervision*".

[11] I. Boujelben, S. Jamoussi and A. B. Hamadou, *"A hybrid method for extracting relations between Arabic named entities*," Journal of King Saud University-Computer and Information Sciences, vol. 26, (4), pp. 425-440, 2014

[12] Ben Abacha, A., Zweigenbaum, P., 2011. A hybrid approach for the extraction of semantic relations from MEDLINE abstracts. In: 12th International Conference on Intelligent Text Processing and Computational Linguistics CICLING2011, Tokyo, Japan, pp. 139–150.

[13] ينظر على الخصوص: أحمد المتوكل (2006) المنحى الوظيفي في الفكر اللغوي العربي الأصول والامتداد، و(2010) الخطاب وخصائص اللغة العربية

[14] ينظر مثلا: أحمد المتوكل (1986) دراسات في نحو اللغة العربية الوظيفي، و(1989) اللسانيات الوظيفية: مدخل نظري

[15] ينظر مثلا: أحمد المتوكل، (2006)، المنحى الوظيفي في الفكر اللغوي العربي: الأصول والامتداد

[16] ينظر: أحمد المتوكل، (2006) المنحى الوظيفي في الفكر اللغوي العربي: الأصول والامتداد، ونظرية النحو الوظيفي.. الأسس والنماذج والمفاهيم، محمد مليطان، منشورات ضفاف بيروت، دار الأمان الرباط المغرب، منشورات الاختلاف الجزائر، الطبعة الأولى، (2014)

[17] ينظر: الخطاب وخصائص اللغة العربية ص 35. وينظر أيضا: الخطاب وخصائص اللغة العربية، أحمد المتوكل، الدار العربية للعلوم بيروت، ومنشورات الاختلاف الجزائر، ودار الأمان الرباط. (2010)، وقضايا اللغة العربية في اللسانيات الوظيفية: البنية التحتية أو التمثيل الدلالي-التداولي. أحمد المتوكل، الرباط: دار الأمان. (1995)، وقضايا اللغة العربية في اللسانيات الوظيفية: بنية المكونات أو التمثيل الصرفي-التركيبي. أحمد المتوكل، الرباط: دار الأمان. (1996)، ومسائل النحو العربي في قضايا النحو الوظيفي، أحمد المتوكل، دار الكتاب الجديد، بيروت لبنان. (2009)، ونظرية النحو الوظيفي.. الأسس والنماذج والمفاهيم، محمد مليطان، منشورات ضفاف بيروت، دار الأمان الرباط المغرب، منشورات الاختلاف الجزائر، الطبعة الأولى، (2014)، والوظيفية بين الكلية والنمطية. أحمد المتوكل، الرباط: دار الأمان. (2003)