مجلة قراءات للدراسات والبحوث النقدية والأدبية واللغوية والترجمة، باللغة العربية وباللغات الأجنبية
تصدر عن كلية الآداب واللغات، جامعة مصطفى اسطمبولي
مجلد: 12 عدد: 1 شهر جوان سنة 2024
ISSN:1112 -8658
E-ISSN: 260*2-7259*

العنوان: **المدونات اللغوية ونماذج من استخداماتها الحاسوبية**

المدونات اللغوية ونماذج من استخداماتها الحاسوبية

Language corpus and examples of their computer uses

أ د رضا بابا أحمد[1]

جامعة معسكر، r.babaahmed@univ-mascara.dz[1]

تاريخ الإرسال	تاريخ القبول
2023/10/19	2024/03/05

ملخص:

يتزايد وعي اللسانيين يوما بعد يوم بضرورة استثمار المدونات اللغوية في التحليل والدراسة، وقد حدد لها الباحثون مجموعة من المعايير والضوابط التي تميز هذه المدونات من غيرها حتى تكون النتائج المتحصل عليها أكثر دقة ونجاعة، طالما أنها تتكون من معطيات لغوية مزودة بمختلف المعلومات اللغوية والمرجعية والسياقية وقد شجع على ذلك تطور الأدوات الحاسوبية التي تساعد على المعالجة، مما سمح بتطور مجال لسانيات المدونة الذي لا يعتبره بعض الباحثين منهجا خاصا في دراسة المدونات اللغوية فحسب بل مجالا علميا يصل فيه الباحث إلى حقائق جديدة قد لا ينوي البحث عنها أيضا. ولأجل تنظيم طرق الاستفادة منها، تم تصنيف تلك المدونات اللغوية تبعا للعوامل المتحكمة فيها إلى عدة أصناف، واستخدامها لعدة أغراض لسانية وترجمية وتعليمية وحاسوبية، لذلك تم ابتكار مختلف البرامج ومنها برنامج 'غواص' الذي يساعد على معالجة المدونات اللغوية العربية وإقامة مختلف العمليات عليهاوالتي تتضمن إحصاء مكوناتها ومقارنتها وتوظيفها لخدمة الإغراض السابقة.

الكلمات المفتاحية: لسانيات المدونة؛ المدونات اللغوية العربية؛ برنامج 'غواص'.

Abstract: The awareness of linguists is increasing day after day of the necessity of investing linguistic corpus in analysis and study. Researchers have identified a set of standards and controls that distinguish these corpusfrom others so that the results obtained are more accurate and effective, as long as they consist of linguistic data provided with various linguistic,

reference, and contextual information. This was encouraged by the development of computer tools that aid in processing, which allowed for the development of the field of corpus linguistics, which some researchers consider not only a special approach to studying linguistic corpus, but also a scientific field in which the researcher gains new facts that he may not intend to search for as well.In order to organize the ways to benefit from them, these linguistic corpus were classified according to the factors controlling them into several categories, and used for several linguistic, translational, educational and computer purposes. Therefore, various programs were invented, including the "Ghawwas" software, which helps to process Arabic linguistic corpus and perform various operations on them, which include: Counting their components, comparing them, and using them to serve the previous purposes.

Keywords: corpus linguistics; Arabic language corpus; 'Ghawwas' software.

مقدمة:

لقد ساهم انتشار الحواسيب والاستعمال الواسع للشابكة في ظهور مدونات واسعة لنصوص إلكترونية موسومة تكون كل كلمة فيها مصحوبة بسمة صرفية تركيبية، وأحيانا مصحوبة بمشجرات تركيبية في المدونات المشجرة، وهي كثيرة الانتشار في اللغات الأجنبية خاصة الإنجليزية، وتشيع فيها أدوات الاستعلام التي تثري وسائل العنونة أو التحشية، والتي تتضمنها برمجيات الواسمات والمحللات التركيبية وغيرهما. فما هو الغرض من إنشاء هذه المدونات، وما هي الاستخدامات الحاسوبية الممكنة لها؟

سأحاول في هذه الأسطر أن أتعرض إلى تعريف المدونات اللغوية وبيان أنواعها والغرض من إنشائها، وسأتطرق أيضا إلى إبراز استخداماتها الحاسوبية من خلال برنامج نموذجي يسمى 'غواص' يتضمن تطبيقات ذات أدوات متعددة في معالجة المدونات اللغوية العربية.

المدونة: تعريفها، أنواعها واستخداماتها

يعرف **سانكلير** Sinclair المدونة بأنها «تشكيلة أو تجميع من المعطيات اللغوية المنتقاة والمنظمة وفق معايير لسانية واضحة لتستغل كعينات ممثلة للغة. والمدونة الحاسوبية هي مدونة مشفرة بطريقة موحدة ومتجانسة لمهام الاسترجاع المحددة بعد أن تم توثيق أجزائها المكونة للغة من حيث أصولها ومصادرها.»(Sinclair J., 1996 : 4-5) وعلى هذا الاعتبار فإن كثيرا من الموارد النصية تفقد تسمية المدونة حينما يتعلق الأمر بتشكيلات لنصوص إلكترونية بسيطة لم يتم جمعها تبعا لمعايير محددة، كما أن الاستفادة من تلك المدونة لا تحصل إلا إذا كانت نصوصها موثقة ومتأصلة وممثلة للغة إذ يستحيل أن تكون المدونة نهائية وشاملة لكل نصوص اللغة، ثم يتم إدماجها في الحاسوب باستخدام طرق

التشفير. كما يبرز في هذا التعريف مفهومان أساسيان: الانتقاء والتمثيل. يشمل الأول وضع معايير خارجية تتعلق بالفئات والأعمار والأساليب وغيرها، ويرتبط الثاني بأنماط المدونات.

من جهته، يعرف **راستييه**Rastier المدونة بأنها: «تجميع منظم لنصوص ووثائق كاملة قد تكون معنونة، وقد روعي فيها -من الناحية النظرية- أنواع الخطاب وأجناس النصوص، وخصصت -من الناحية العملية- للقيام بتطبيقات عليها.»(Rastier F., 2004) فبين بذلك الشروط والضوابط التي تحدد المدونة وهي:

- كونها مجموعة من النصوص والوثائق الكاملة التي تم جمعها وتنظيمها بحسب وجهة نظر محددة؛
- الأخذ بعين الاعتبار اختلاف تلك النصوص المكتوبة والخطابات الشفهية من حيث الشكل والغرض؛
- إعداد المدونة وتجميع النصوص يكون بهدف إجراء مختلف التطبيقات الحاسوبية عليها.

وعليه، يعمل اللسانيون على إنشاء مدونات بانتقاء مجموعة من النصوص انطلاقا من معايير خاصة من أجل القيام بدراسة لسانية، ومع انتشار المعلوميات ارتبطت المدونات اللغوية بالحاسوب وصارت تعني ذلك المستند أو مجموعة المستندات الرقمية التي يمكن أن تجرى عليها مختلف العمليات باستخدام أدوات معالجة المدونات كاستخراج ورود عبارة ما في السياق أو تحديد المصطلح الأكثر استخداما في مدونة علمية. هذه المهام يصعب بل يستحيل القيام بها يدويا لأنها تتطلب استقراء كل النصوص وهذا غير ممكن التحقيق.

والمدونة المشكلة قد تحتوي على نصوص مكتوبة أو منطوقة، ولكن التسجيلات الصوتية يتم نسخها أيضا بحروف مكتوبة حتى يسهل البحث فيها. وتجمع المدونة بهدف وصف نظام اللغة بالكشف عن معاني بعض الكلمات أو العبارات مثلا أو بفهم شروط استعمال بعض التراكيب النحوية؛ أو بهدف وصف مميزات وخصائص خطاب أو نمط خاص من الخطابات كالخطاب السياسي الذي يتناول كثيرا في تحليل الخطاب. وعليه، تكون طرائق البحث مختلفة بحسب الأهداف المرجوة.

الجديد بالنسبة للمدونات في هذه السنوات الأخيرة يتمثل في تزايد حجمها وكثرة انتشارها وتوفر أدواتها؛ فلم تعد مجرد سلسلة من الحروف المكتوبة، بل صارت مجموعة من الكلمات المعنونة أي المزودة بمعلومات مختلفة صرفية وتركيبية ودلالية وتنغيمية وغيرها ولا تزال في تزايد وتطور. ففي سنوات 1980 ركزت الأبحاث على العنونة الصرفية التركيبية، وفي سنوات 1990 اتجهت نحو المدونات المشجرة التي تحتوي على عدد ضخم من الكلمات مثل المدونة الوطنية البريطانية بـ100 مليون كلمة معنونة (Habert B. et *al.*, 1997 : 7).

ولا يزال دور المدونات بارزا على الرغم من ثورة **تشومسكي**Chomsky في اللسانيات سنة 1957 التي أدت إلى استبدال حدس المتكلم بالمدونات واستبعاد الأعمال اللسانية الكمية والدراسات التجريبية للمعطيات اللغوية المشاهدة، بالمقابل ظلت هناك بعض الأعمال التي تستحضر المدونات الإلكترونية والمعاجم والأنحاء الوصفية، وتستغلها كذلك في فحص الفرضيات أو المقابلة بين النموذج النظري المقترح وبين التحققات اللغوية الفعلية.

ومما عضد الدراسات اللسانية المشتغلة على النصوص والمدونات الإلكترونية أيضا، دخول المعالجة الآلية للغة على الخط حيث أعطت مسارا جديدا لإنشاء المدونات المعنونة واستعمالها، وقد كان المنطلق في ذلك هو عدم ملائمة الأطر النظرية المستخدمة في المعالجة الآلية للغة آنئذ. هناك تفسيران ممكنان لهذه الحالة: أولهما احتياج نظام المعالجة الآلية للغة إلى موارد (معاجم وأنحاء) ضخمة من حيث عدد المداخل المعجمية والقواعد، ومفصلة حيث تتعلق بشروط استخدام الكلمات مثلا؛ ثانيهما تحسين تلك الأنظمة يتم بملاحظة مجموع المعطيات النصية المتوفرة.

حتى تلك اللحظة التي تبنت فيها المعالجة الآلية للغة مقاربة المدونة كانت تعتمد على قواعد لنمذجةوصورنة المعرفة الإنسانية واستخراج القواعد الضمنية وفق نماذج رمزية، وتستبعد النماذج الحسابية والإحصائية، لكن مشاهدة المعطيات اللغوية بكمية كبيرة ومعالجة تدفق المعلومات على الشبكة يقود لا محالة الى الاعتماد على مقاربات كميات ورمزية.

وبالنتيجة نعيش في الوقت الحاضر تغييرا كبيرا؛ فالرهانات الصناعية معتبرة جدا وإنشاء المدونات الضخمة يلاقي دعما كبيرا من قبل القطاع الخاص والقوة العمومية خاصة في العالم

الأنجلوسكسوني، وصار التركيز في مجال المعالجة الآلية للغة على المعطيات النصية لتؤدي معالجة متجذرة بقوة في المعطيات الشاهدة(Habert B. et *al.*, 1997 : 7).

بالنسبة للدراسات التي ترتبط بالمدونات هناك:

- دراسات **مبنية على المدونات**corpus-based حيث تستخدم عادة معطيات المدونة من أجل التحقق من صحة نظرية أو فرضية متداولة في الأدبيات الحالية أو دحضها أو تحسينها. وصف لسانيات المدونة بأنها منهج يدعم هذه المقاربة في استخدام معطيات المدونة في اللسانيات(McEnery T. &Hardie A., 2011 : 6). وهي تفيد من المدونات اللغوية في البحث اللساني وفق إطارمحدد مسبقا كالتعرف على شيوع كلمات أو تراكيب معينة أو التثبت من ظواهر لغوية معينة؛
- ودراسات **موجهة بالمدونات**corpus-directed حيث تعتبر المدونة المصدر الوحيد لفرضياتنا حول اللغة، وترفض بالتالي كون لسانيات المدونة منهجا. وهذه الدراسات تنتج من النظر في المدونات اللغوية حيث يصل الباحث إلى حقائق جديدة لم يكن يبحث عنها بالضرورة، لذلك لا يفرق البحث الموجه بالمدونات اللغوية بين المعجم والنحو والتداولية وعلم الدلالة والخطاب، بل تتبع منهجا كليا في البحث اللساني(العصيمي ص. ف. وآخرون، 2015 : 22).

من جهة أخرى، يمكن تصنيف المدونات من عدة زوايا بحسب طبيعة النصوص وأنماطها التي تحتوي عليها، والأهداف التي أنشئت من أجلها؛ أو بحسب الزمن فقد تحتوي على نصوص أنتجت في فترة زمنية محددة، أو في فترات تاريخية متعاقبة؛ أو بحسب كون معطياتها اللغوية خالية من أي وسم، أو محشاة بالمعلومات الصرفية والنحوية والمعجمية وغيرها(العصيمي ص. ف. وآخرون، 2015 : 27).

إن عملية الوسم تمكننا من إجراء أفضل بحث على النصوص من خلال خفض نسبة المعلومات غير المميزة (الضجيج) التي يمكن أن تنتج عند البحث: مثل الحصول على الأفعال التي على وزن 'فَعَل' إذا بحثنا عن الأفعال فقط دون الأسماء، وخفض نسبة الإجابات المميزة التي لم يمكن استخراجها عند البحث أيضا (الصمت). سأركز فيما يأتي على المدونات الموسومة وأبين أنواعها واستخداماتها.

ففي مجالالمدونات الموسومة، المدونة المعنونةهي المدونة التي يتم وسم كلماتها بمقولاتها النحوية عادة كوسم أقسام الكلام الذي يحدد الصنف الذي تنتمي إليه الكلمة (فعل، اسم، ضمير، حرف...)،

والذي يسهم في إزالة اللبس بين الكلمات المشتركة اللفظ، وكوسم تجريد الكلمة الذي يعمل على رد الكلمات إلى أصولها المعجمية بحذف الزوائد الصرفية والاشتقاقية، بالإضافة إلى التحشية الدلالية التي تعكس الخصائص والحقول الدلالية والتحشية المرجعية التي تعين الاتساق اللفظي بين الكلمات، وغيرها من العناوين التي تضاف إلى المعطيات اللغوية داخل المدونة(الدكروري أ.، 2018 : 74).

والمدونة المشجرةوتسمى البنك الشجري Treebank وهو يتضمن بعض التحليلات النحوية التي تتجاوز بيان الأجزاء الكلامية إلى الكشف عن العلاقات التركيبية بين هذه الأجزاء. وهو أساس هام لبناء التطبيقات الإحصائية لمعالجة اللغات الطبيعية مثل المحللات النحوية وتطبيقات الترجمة الآلية ورفع اللبس الدلالي عن الكلمات؛ كما تستخدم في الكشف عن الخصائص الأسلوبية للنصوص، وإحصاء ترددات بعض الظواهر المتصلة بالجوانب النفسية(روبي أ.، 2017 : 11).

أما فما يخص استخداماتها،فإن المدونات اللغوية تستخدم لعدة أغراض، وبالتالي تطبق لسانيات المدونات على عدة ميادين. من أبرز تلك الميادين المعجمية والمصطلحية حيث لا يمكن أبدا الاهتمام بدراسة المعجم اللغوي عموما والمعجم العلمي المتخصص دون الاعتماد على مدونات ضخمة الحجم حتى يمكنها أن تكون ممثلة للغة المدروسة وللغة التخصص المقصودة. إن تلك الدراسات المعجمية والمصطلحية تمكننا من تحديد المفردات والمصطلحات الجديدة والمهملة وتحديد المعاني والمفاهيم التي تعبر عنها حتى يتم تجديد المعاجم وتطوير اللغة المدرسة.

كما تشكل المدونات اللغوية موارد مهمة للترجمة لأنها توفر الأمثلة العملية المتداولة بين المترجمين والمختصين، وتدلنا على الأساليب المستخدمة في المقابلة بين المفردات والعبارات ضمن مختلف الموضوعات والمجالات. وتعاني اللغة العربية نقصا في هذا الصدد حيث تحتاج إلى جمع مدونات عربية ومقارنتها بنظيرتها في اللغات الأجنبية(روبي أ.، 2017 : 46).

من جهتها، تظهر الصلات بين المدونات اللغوية وبين المعالجة الآلية للغة في تطبيق تقنيات هذه الأخيرة في تحشية المدونات كنسخ العبارات المنطوقة وعنونة الكلمات بطريقة آلية. كما أن أدوات المعالجة تتطور بفضل المدونات اللغوية من خلال التقييمات الدورية والتعديلات التي يجريها مستعملو تلك الأدوات في تحليل المدونات اللغوية.

وبالعروج على مجال تعليم اللغات، نجد المدونات اللغوية قد حظيت باهتمام الباحثين بحيث ابتكرت مناهج تعليم جديدة معتمدة على المدونات؛ إذ يتم فيها الاعتماد على نصوص حقيقية لتعلم اللغة عوض تبني مقاربة ترتكز على القواعد باتباع ثلاث مراحل لتحسين كفاءة المتعلم اللغوية، وهي: التعيين من خلال إعداد معطيات الكشف السياقي التي تبين سياق استعمال المفردة أو العبارة، ينتقل بعدها المتعلم إلى تصنيف العناصر اللغوية أو الخطابية بهدف استنباط عموميات لها في اللغة (مثلا كالتمييز بين شهيد وقتيل في العربية).

برنامج معالجة المدونات اللغوية العربية 'غواص'

عدة برامج حاسوبية عربية طورت لتحليل المدونات اللغوية منها برنامج 'غواص'، وهو برنامج مجاني مفتوح المصدر يتسم بإمكاناته الكبيرة في معالجة النصوص العربية. وهو ذو واجهتي استخدام عربية وإنجليزية. يستطيع عرض تكرار تردد هياكل الكلمات tokens والكلمات الفريدة types، والوثائق. ويدعم البرنامج أشكالا مختلفة من الملفات: html, doc-docx, txt. كما يدعم نظامي التشفير: UTF-8, ANSI. ويتيح البرنامج للباحثين والمستفيدين بتحميل ومعالجة ملف يحتوي على أكثر من 50 مليون كلمة، من خلال العمل في بيئة جافا.

ويحتوي البرنامج على ثلاث واجهات استخدام:

يمكن من خلال واجهة الاستخدام الأولى أن يقوم الباحث بتحميل ملفات النصوص، سواء الخاصة بالمدونة اللغوية الرئيسة، أو المدونة اللغوية المرجعية

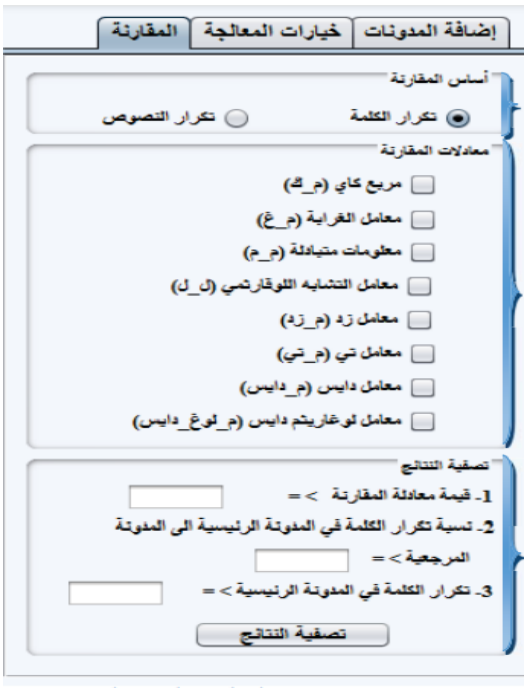
الشكل 1: واجهة إضافة المدونات في برنامج 'غواص

بينما توفر واجهة الاستخدام الثانية مجموعة من الإمكانات هي: تحليل المتتابعات اللفظية n-grams،وهي عملية من التتابعات الإحصائية لوحدات النص أو الجملة في ضوء خوارزميات رياضية، وتتوقف N أو عدد المتتابعات على ما تصبو إليه التطبيقات المنشودة. والخوارزميات مجموعة من الخطوات المتسلسلة الرياضية والمنطقية لحل مشكلة

ما (روبي أ.، 2017 : 25)، كما توفر هذه الواجهة محرك بحث المدونة اللغوية، وحذف أو إبقاء علامات التشكيل، أو تعديل بعض التمثيلات، وتحديد الملفات التي يرغب المستفيدون في البحث فيها، ورفع قوائم الاستثناء أو قوائم الاعتبار.

الشكل2: واجهة خيارات المعالجة في برنامج 'غواص'

فيما تعرض واجهة الاستخدام الثالثة مجموعة من الإحصاءات والتحليلات. حيث يمكن حساب قيمة مربع كاي للدلالة الإحصائية Chi-square، وطريقة احتمالات سجل الأداء log-likelihood، ومعامل ارتباط الغرابة Weirdness Coefficient، ومعامل المعلومات المتبادلة Mutual Information، ومعامل ارتباط دايسDice Coefficient.



الشكل3: واجهة المقارنة في برنامج 'غواص'

يتم تقييم ملفات النصوص بحيث تمثل الدلالة الإحصائية لتوزيع كلمة معينة. وتستخدم طريقة احتمالات سجل الأداء في المقارنات التي تقدر قيم الاحتمالات، ومن ثم يمكن تقدير قيم معاملات الارتباط. حيث يشار إلى ذلك بنتائج عالية ومنخفضة ضمن اختبار معياري.

فيما يستخدم معامل ارتباط الغرابة في استخراج الكلمات المفتاحية والمتلازمات اللغوية من النصوص؛ من خلال المضاهاة بين مدونة لغوية رئيسة، ومدونة لغوية مرجعية. ويشار إلى ناتج معامل ارتباط الغرابة بأربع قيم: القيمة الأولى للمستوى 0، والقيمة الثانية أكبر من 1 حينما تكون الكلمات أكثر ترددا في المدونة اللغوية الرئيسة عنها في نظيراتها المرجعية. والقيمة الثالثة أقل من 1 عندما تكون الكلمات أكثر ترددا في المدونة اللغوية المرجعية عنها في نظيراتها الرئيسة. بينما القيمة الرابعة إلى 'ما لا نهاية' حينما ترد الكلمات في المدونة اللغوية الرئيسة فقط.

ويفيد معامل المعلومات المترابطة في التحقق من قوة الارتباط بين المتلازمات اللغوية. فكلما ازدادت القيمة ازدادت قوة الارتباط بين المتلازمات اللغوية. وتشير عادة القيمة الأقل من 3 إلى عدم الارتباط بينها.بينما يدل معامل ارتباط دايس على قوة الارتباط أو ضعفه بين الكلمات والوثائق. وتتراوح قيم هذا المعامل بين 0 و14. وتظهر هذه القيم في حالة وجود ارتباط ين المتلازمات اللغوية. وكلما اقتربت القيمة من 14 ازدادت قوة الارتباط(الدكروري أ.، 2018 : 123-125).

الخاتمة:

أصبحت المدونات اللغوية من الأهمية بمكان بحيث صارت تعتمد عليها الدراسات اللسانية والإنسانية عموما خاصة مع تطور الوسائل الحاسوبية التي تسهل عملية البحث في تلك المدونات والقيام بمختلف العمليات الإحصائية بغرض الاستفادة من نتائجها في عدة مجالات كتعليم اللغات وتحليل الخطابات وغيرهما. في الأخير، يمكن الخروج بمجموعة من النتائج أوردها فيما يلي:

- انتقاء النصوص والمعطيات اللغوية وجمعها في مدونات يتمان وفق معايير لسانية محددة لتحقيق أهداف واضحة؛
- التوسل بالأدوات الحاسوبية في تدوين تلك النصوص والمعطيات اللغوية؛
- وسم المدونات بعنونة كلماتها وعباراتها يساهم بقدر كبير في تيسير العمل على المدونات وجعله أكثر فائدة بالنسبة للباحثين؛
- برنامج أدوات معالجة المدونات اللغوية العربية 'غواص' نموذج للبرامج التي توفر مختلف الإمكانات للعمل على المدونات من خلال الوسائل الإحصائية التي يوفرها والتي تعين على تحديد تكرار الكلمات والعبارات وتكشف عن ورودها في مختلف السياقات concordance.

قائمة المراجع:

.الدكروري، أ. (2018) **المدونات اللغوية العربية ودورها في معالجة النصوص العربية**. الرياض: مركز الملك عبد الله لخدمة اللغة العربية.

.روبي، أ. (2017) **البنك الشجري النحوي: بناؤه وتوظيفه في إطار تقنيات الذكاء الاصطناعي**. الرياض: مركز الملك عبد اللهلخدمة اللغة العربية.

.العصيمي، ص. ف. وآخرون. (2015) **المدونات اللغوية العربية: بناؤها وطرائق الإفادة منها**. الرياض: مركز الملك عبد الله لخدمة اللغة العربية.

Habert, B., Nazarenko, A., & Salem, A. (1997). *Les linguistiques de corpus*. Paris: Colin..

McEnery, T., &Hardie, A. (2011). *Corpus linguistics: Method, theory and practice*. Cambridge : Cambridge UniversityPress..

RASTIER, F. Enjeux épistémologiques de la linguistique de corpus. *Texto !* [en ligne], juin 2004. Rubrique Dits et inédits. . Disponible sur : <http://www.revue-texto.net/Inedits/Rastier/Rastier_Enjeux.html>. (Consultée le 04/11/2022).

Sinclair J. (1996). Preliminary recommendations on Corpus Typology, Technical report, EAGLES (Expert Advisory Group . on Language Engineering Standards).