



علم الإحصاء في لغويات المدونة الحاسوبية

دليل عملي

تأليف

Vaclav Brezina

ترجمة

د. سلطان بن ناصر المحيول

أستاذ مشارك - اللغة العربية

كلية الآداب - جامعة الملك سعود

دار جامعة
الملك سعود للنشر
KING SAUD UNIVERSITY PRESS



ص.ب ٦٨٩٥٣ - الرياض ١١٥٣٧ المملكة العربية السعودية

فهرسة مكتبة الملك فهد الوطنية أثناء النشر

بريزينا، فاكلاف.

علم الإحصاء في لغويات المدونة الحاسوبية دليل عملي. / فاكلاف بريزينا؛
سلطان بن ناصر المحيول - الرياض، ١٤٤٢هـ.

٣٩٥ ص؛ ١٧ سم × ٢٤ سم

ردمك: ١ - ٩٤٧ - ٥٠٧ - ٦٠٣ - ٩٧٨

١- الاحصاء ٢ - الطرق الاحصائية أ. المحيول، سلطان بن ناصر. (مترجم)

ب. العنوان

١٤٤٢/٧٤٨٦

ديوي ٣١٠

رقم الإيداع: ١٤٤٢/٧٤٨٦

ردمك: ١ - ٩٤٧ - ٥٠٧ - ٦٠٣ - ٩٧٨

هذه ترجمة عربية محكمة صادرة عن مركز الترجمة بالجامعة لكتاب:

Statistics In Corpus Linguistics A Practical Guide

By: Vaclav Brezina

© Vaclav Brezina 2018

وقد وافق المجلس العلمي على نشرها في اجتماعه الثاني عشر للعام ١٤٤٢هـ المعقود

بتاريخ ١٩/٦/١٤٤٢هـ الموافق ١/٢/٢٠٢١م

جميع حقوق النشر محفوظة. لا يسمح بإعادة نشر أي جزء من الكتاب بأي شكل وبأي وسيلة سواء كانت إلكترونية أو آلية بما في ذلك التصوير والتسجيل أو الإدخال في أي نظام حفظ معلومات أو استعادتها بدون الحصول على موافقة كتابية من دار جامعة الملك سعود للنشر.

مقدمة المترجم

اعتمدتُ في أثناء ترجمة مصطلحات علم الإحصاء على الأكثر تداولاً من المقابلات العربية، وسيجد القارئ أيضاً أن الأرقام في معادلاتها ونتائجها الإحصائية قد وُضعت بالأرقام العربية بدلاً من الأرقام الهندية؛ ذلك أن كل البرامج الحاسوبية من محررات نصوص أو أوراق عمل أو برامج إحصائية أو لغات برمجة لا تتعامل مع الأرقام الهندية، ولتفادي الالتباس المحتمل الذي قد يقع فيه القارئ العربي عند التعامل مع القيم العددية وكسورها العشرية التي تظهر بالأرقام الهندية، والتي لا تكون ممثلة في البرامج الحاسوبية المتعلقة، اعتمدتُ الأرقام العربية، على سبيل المثال: قراءة هذا العدد العشري (0.001) ستكون أفضل على قارئ هذا الكتاب من قراءة ما يقابله بالأرقام الهندية (0.001)، كذلك الحال مع الأعداد الصحيحة التي تطول أجزاؤها العشرية كالعدد (12.8123) الذي سيُمثّل هكذا في البرامج الحاسوبية، وقد تُحدث كتابته بالأرقام الهندية عدم تساق أو ارتباكاً عند القارئ، ذلك أن الأرقام العربية هي المعمول بها في جداول البيانات.

هذا الكتاب سيكون دليلك العملي إلى تحقيق المهارات الإحصائية المتقدمة في تحليل البيانات اللغوية، سعى فيه المؤلف إلى تغطية الموضوعات اللغوية التحليلية بإضفاء أساليب وتقنيات التحليل الإحصائي في لغويات المدونة الحاسوبية، وفي كل فصل من فصوله موضوعات تحليلية لغوية محددة، شُرحت في إطار ما يناسبها من القياسات الإحصائية المتقدمة للجمع بين لغويات المدونة الحاسوبية وعلم الإحصاء.

وقامت هذه الإجراءات الإحصائية على أداة لغة البرمجة R، ووفّر المؤلف أدوات التحليلات للبيانات اللغوية، كما وُفّر أيضاً الأكواد الخاصة بتلك التحليلات، وذلك كله في موقع أدوات لانكستر الإحصائية.

تعتمد لغويات المدونة الحاسوبية على البيانات اللغوية، ومن المهم أن يُدرك القارئ أنَّ المدونة corpus في هذا الكتاب تشير إلى مفهوم البيانات اللغوية linguistic data. وقد مرَّ مصطلح corpus بصفته مدونةً حاسوبيةً بثلاث مراحل زمنية متتابة، كل مرحلة زمنية تشكّل فيها هذا المصطلح بمفهوم محدد، وهذه المراحل هي:

الأولى: المفهوم المتعلق بمجموعة من النصوص المكتوبة أو المنطوقة التي تُجمع وتُحفظ إلكترونياً. الثاني: المفهوم المتعلق بتصميم مدونات حاسوبية من النصوص المكتوبة والمنطوقة، وبنائها حسب أوعية النصوص أو مصادرها كالإنترنت والصحف والكتب ووكالات الأنباء والمجلات، إلخ، وحسب مجالات كل وعاء، وحسب موضوعات كل مجال، وحسب مناطقها الجغرافية، وحسب فترات الزمنية، وبلي هذا التصميم وذلك البناء دمج وظائف وأدوات فيها لمعالجتها، أو استعمال أدوات قائمة بذاتها، وتعمل هذه الوظائف والأدوات للبحث عن الكلمة، وعن عدد مرات تكرارها، وعن متصاحباتها اللفظية، وعن بعض القياسات الإحصائية التي تدل على معنوية التصاحب ودلالاتها الإحصائية من عدم الدلالة، وعن كيفية الاستفادة من هذا كله في اللغويات والدراسات الاجتماعية.

الثالث: المفهوم المتعلق بإعداد بيانات لغوية وتبويبها بدقة، والتحقق من صحتها، ومن ثم تحليلها تحليلًا إحصائيًا متقدماً، وهو ما سيجده القارئ في هذا الكتاب. وتكون البيانات اللغوية على صيغة مجموعة بيانات dataset، تتوزع في خلاياها المنفصلة قيمٌ ورموز تُمثل حقائق وأرقامًا تفصيلية عن حالة (لغوية) معينة تُدرس وتُعالج بصفته وحدة كاملة. وسيتدرب القارئ في هذا الكتاب على كيفية تطبيق القياسات الإحصائية والتمثيل البياني لها على النحو التسلسلي الآتي:

أولاً: الإحصاءات الوصفية للبيانات اللغوية في الفصل الأول.

ثانياً: التكرار والتشتت والتنوع في الفصل الثاني.

ثالثاً: التصاحبات اللفظية والكلمات المفتاحية والترميز اليدوي للخطاب في الفصل الثالث.

رابعاً: من الإحصاءات البسيطة إلى النمذجات المعقدة للنحو المعجمي في الفصل الرابع.

خامساً: معامل الارتباط والتجميع وتحليل العوامل لتنوعات السجل اللغوي في الفصل الخامس.

سادسًا: قياس التنوع الفردي والتنوع الاجتماعي في اللغويات الاجتماعية وعلم الأسلوب في الفصل السادس.

سابعًا: تحليل البيانات اللغوية التاريخية بأساليب التمهيد في الفصل السابع.

ثامنًا: التعرف والاطلاع على عشرة مبادئ مهمة للتفكير الإحصائي والتحليل البعدي وأحجام التأثير في الفصل الثامن.

أتمنى أن يكون هذا الكتاب أحد المقررات الجامعية ومصادرها المرجعية، كما أتمنى رحلة تطبيقية ممتعة لكل متدرب على هذا الكتاب.

المترجم

د. سلطان المجبول

محرم ١٤٤٣هـ

عرفان المؤلف

أتوجه بالشكر إلى توني مكناري Tony McEnery ودانا جابلاسوفا Dana Gablasova على دعمهما غير المنقطع وعلى تشجيعهما في أثناء فترة تأليف هذا الكتاب، وأشكرهما على ملحوظاتهما الدقيقة على كل فصل من فصول الكتاب. تلقى هذا الكتاب عدداً من الملحوظات المستنيرة من جابريلي باللوتي Gabriele Pallotii الذي قرأ جزءاً كبيراً منه. أتقدم بالعرفان أيضاً إلى الزملاء الذين أبدوا ملحوظاتهم على أجزاء من الكتاب، وهم: بيتر ديجل Peter Diggle، ومايكل غوتيه Michael Gauthier، وأندرو هاردي Andrew Hardie، وإيرينييه مارين سيرفانتس Irene Marin Cervantes. أشكر جيل سميث Gill Smith لمساعدتها على تنسيق النسخة الأولى من الكتاب، وأشكر إيرين مارتين سيرفانتس على تحضير مداخل الكشافات. كل الشكر حقيقة لمحكمي الكتاب على ملحوظاتهم المشجعة والمفيدة.

دُعم تأليف هذا الكتاب من قبل ESRC بعقد رقم EP/P001559/1 ورقم ES/K002155/1

عن هذا الكتاب

ما موضوع هذا الكتاب؟

يُمهّد هذا الكتاب الطريق إلى الإجراءات الإحصائية التطبيقية في لغويات المدونة الحاسوبية الذي يُعد حقلاً يعتمد على الحاسب الآلي في معالجة اللغة وتحليلها، وتغطي موضوعات الكتاب بنية الكلمة، والنحو، واللغويات الاجتماعية، وتحليل الخطاب، واللغويات التاريخية، ويعرض صورة عامة عن منهجيات تحليل اللغة الأكثر دقة باستخدام المدونات (الحاسوبية)، ويفيد بالأساليب الجديدة التي لم تُطبّق في حقل لغويات المدونة الحاسوبية، ولا يُفترض من القارئ أن تكون له خلفية معرفية بالإحصاء؛ لأن الكتاب يشرح المفاهيم والمناهج الضرورية بلغة أوضح من تعقيداتها. من الممكن تطبيق كل الإجراءات التحليلية المذكورة في هذا الكتاب باستخدام أدوات لانكستر الإحصائية Lancaster Stats Tools الشبكية (انظر إلى المبحث التالي: كيف تستخدم هذا الكتاب؟). زُوّد الكتاب بكثير من الأمثلة الإحصائية وبعده من حالات الاستخدام case studies لتلك الأمثلة، كما وُوقِف على الإحصاءات وقوفاً مقنناً، ويُركز الكتاب في الأوجه التطبيقية للتحليل الإحصائي من خلال طبيعة البحث المراد تصميمه، ومن خلال طبيعة ما تتضمنه أُطر (أو أشكال أو صيغ) البيانات، ولذلك وُضع موقعٌ شبكي يتضمن مجموعات البيانات الكاملة التي أُستخدمت في الكتاب من أجل تكرار التحليل من القارئ. يُعدّ حقل لغويات المدونات الحاسوبية علماً منهجياً متعدد الاستعمالات لتحليل اللغة، وتطبيقاً لحقول عدّة كاللغويات، والعلوم الاجتماعية، والإنسانيّات الرقمية، وغيرها، ويرمي إلى تسهيل الاستخدام الجادّ للبيانات اللغوية لأكبر قدر ممكن من المحللين.

لمن هذا الكتاب؟

لكل مهتم بلغويات المدونة الحاسوبية بالتحليل الكمي للغة، من الطلاب والباحثين في حقول اللغويات وعلم الاجتماع والتاريخ والتربية وغيرها. ويرمي الكتاب أساسًا إلى إعانة القراء على تطبيق طرائق التفكير الإحصائي الأساسية من أجل تحقيق أبحاث مستنيرة. وبناءً على ذلك، يتضمن هذا الكتاب أيضًا أسئلة للمناقشة (مباحث: فِكر Think about)، وتمارين يمكن للقراء التدرب عليها من أجل اختبار مدى استيعابهم لكل تحليل إحصائي. يُمكن الحصول على إجابات أسئلة المناقشة بالعودة إلى الموقع الآتي: <http://corpora.lancs.ac.uk/stats/material.php>.

كيف تستخدم هذا الكتاب؟

يلبي الكتاب احتياجات الطلاب والباحثين في مجال لغويات المدونة الحاسوبية، وقد رُوِّد بأدبيات الحقل وبأفضل الأساليب التحليلية، ويُمكن استخدام هذا الكتاب بوصفه مرجعًا جامعيًا، كما يُمكن استخدامه للدراسة الذاتية. يستطيع القراء تلبية ما يرغبون في البحث عنه في اللغة بعد دراسة الإحصاءات الأساسية في الفصل الأول، وقد رُبطت الأساليب الإحصائية مع بعضها ببعض بين الفصول بالإحالات، ووُضعت في كشاف في آخر الكتاب.

يأتي هذا الكتاب مع مصدر شبكي مصاحب يحتوي على أمثلة إضافية، ومجموعة بيانات، وشروحات مرئية، وشرائح عرض، وأدوات سهلة الاستخدام للقياسات الإحصائية التي أنتجت الرسومات البيانية في الكتاب (<http://corpora.lancs.ac.uk/stats>). يُمكن للقارئ في الواقع أن يستخدم أدوات لانكستر الإحصائية Lancaster Stats Tools ليطلع على الإجراءات الإحصائية المطبقة في الكتاب، وعليه فإن مستخدمي تلك الأدوات ممن ليسوا على مَلاك مؤسسات تدعم لهم الحُزم الإحصائية التجارية كحزمة IBM SPSS لا يحتاجون إلى دفع مبالغ كبيرة للاشتراك، كما أنهم ليسوا بحاجة إلى تعلم لغات البرمجة كلغة R بفضل ما تقدمه أدوات إحصائيات لانكستر الشبكية من ولوج مجاني إلى أدوات إحصائية متقدمة تعمل عمل لغة R، وبواجهة استخدام سهلة الوظائف تُمكن النسخ واللصق من برنامج إكسل وبرنامج Calc (برنامج الأوفيس المفتوح OpenOffice Calc). أؤمن بأن علم الإحصاء ينبغي ألا يكون عصيًا عن الاستيعاب، لكون الحاسب الآلي قادرًا على مضغ كل الأعمال الرقمية الصعبة، وقادرًا على أن يكون أداة تحليلية فعالة بعد استيعاب مبادئ الفكر الإحصائي وتطبيقاته. دعونا نكتشف ذلك معًا!

المحتويات

هـ	مقدمة المترجم
ط	عرفان المؤلف
ك	عن هذا الكتاب
ك	ما موضوع هذا الكتاب؟
ل	لمن هذا الكتاب؟
ل	كيف تستخدم هذا الكتاب؟
ص	قائمة الأشكال
غ	قائمة الجداول

الفصل الأول: لغويات المدونة الحاسوبية في ضوء علم الإحصاء

١	(١, ١) ما الذي يتناوله هذا الفصل؟
٢	(١, ٢) ما المقصود بعلم الإحصاء؟ العلوم ولغويات المدونة الحاسوبية وعلم الإحصاء
٦	(١, ٣) مصطلحات علم الإحصاء الأساسية
٢٠	(١, ٤) بناء المدونات وتصميم الأبحاث
٣٠	(١, ٥) استكشاف البيانات وتمثيلها
٣٩	(١, ٦) التطبيق وأمثلة إضافية: هل يستخدم كُتّاب الرواية الصفات أكثر من الأكاديميين؟
٤٢	(١, ٧) تمارين
٤٧	تذكّر الآتي

٤٧	قراءات إضافية.....
٤٨	أدوات لانكستر الإحصائية.....

الفصل الثاني: الكلمات: التكرار والتشتت والتنوع

٤٩	(١, ٢) ما الذي يتناوله هذا الفصل؟.....
٤٩	(٢, ٢) الكلمات الفعلية، والكلمات النوعية، والمواد المعجمية، واللكسيات.....
٥٤	(٢, ٣) كلمات في قائمة التكرار.....
٥٩	(٢, ٤) مشكلة حلزون البحر: التشتت.....
٦٩	(٢, ٥) ما الكلمات المهمة؟ تكرار المتوسط المقلص.....
٧٣	(٢, ٦) التنوع المعجمي: نسبة الكلمات النوعية إلى الكلمات الفعلية والنسبة الموحدة والنسبة المتوسطة المتحركة بينهما.....
٧٦	(٢, ٧) التطبيق وأمثلة أخرى: هل يتحدث البريطانيون عن الطقس طوال الوقت؟.....
٨١	(٢, ٨) تمارين.....
٨٥	قراءات إضافية.....
٨٥	أدوات لانكستر الإحصائية.....

الفصل الثالث: علم الدلالة والخطاب: التصاحبات

اللفظية، والكلمات المفتاحية، وموثوقية الترميز اليدوي

٨٧	(١, ٣) ما الذي يتناوله هذا الفصل؟.....
٨٨	(٢, ٣) التصاحبات اللفظية ومقاييس التصاحب.....
٩٩	(٣, ٣) مخططات التصاحب والشبكات: استكشاف المتجاورات المتقاطعة.....
١٠٣	(٣, ٤) الكلمات المفتاحية وكلمات الإيقاف.....
١١٤	(٣, ٥) مقاييس ملائمة المقدّر البيني.....
١٢٠	(٣, ٦) التطبيق وأمثلة أخرى: ما رأي قراء الصحف البريطانية في الهجرة؟.....
١٢٥	(٣, ٧) تمارين.....

١٣٢	قراءات إضافية
١٣٢	أدوات لانكستر الإحصائية

الفصل الرابع: النحو المعجمي: من الإحصاءات البسيطة إلى النماذج المعقدة

١٣٣	(١, ٤) ما الذي يتناوله هذا الفصل؟
١٣٤	(٢, ٤) تحليل خصائص البيانات المعجمية النحوية
١٤١	(٣, ٤) الجدولة المتقاطعة والنسب المئوية واختبار كاي تربيع
١٥٢	(٤, ٤) الانحدار اللوجستي
١٦٩	(٥, ٤) التطبيق: That أم Which ؟
١٧٥	(٦, ٤) تمارين
١٨٠	قراءات إضافية
١٨٠	أدوات لانكستر الإحصائية

الفصل الخامس: تنوع السجل اللغوي: الارتباط والتجميع والعوامل

١٨١	(١, ٥) ما الذي يتناوله هذا الفصل؟
١٨٢	(٢, ٥) العلاقات بين المتغيرات: الارتباطات
١٩٥	(٣, ٥) التصنيف: تحليل التجميع المتكامل الهرمي
٢٠٥	(٤, ٥) التحليل متعدد الأبعاد
٢٢٠	(٥, ٥) التطبيق: السجلات اللغوية في الإنجليزية النيوزيلندية
٢٢٩	(٦, ٥) تمارين
٢٣٣	قراءات إضافية
٢٣٤	أدوات لانكستر الإحصائية

الفصل السادس: اللغويات الاجتماعية وعلم الأسلوب: التنوع الاجتماعي والفردية

- (٦, ١) ما الذي يتناوله هذا الفصل؟ ٢٣٥
- (٦, ٢) الأسلوب الفردي والتنوع الاجتماعي: أين تبدأ المتغيرات اللغوية الاجتماعية؟ ٢٣٦
- (٦, ٣) مقارنة المجموعة: اختبار-ت، وتحليل التباين، واختبار مان-ويتني U، واختبار كروسكال-والاس ٢٤٠
- (٦, ٤) الأسلوب الفردي: تحليل التناظر ٢٥٧
- (٦, ٥) السياق اللغوي: نمذجة التأثير الممتزج ٢٦٥
- (٦, ٦) التطبيق: من هذا الشخص من البيت الأبيض؟ ٢٧١
- (٦, ٧) تمارين ٢٧٥
- قراءات إضافية ٢٧٩
- أدوات لانكستر الإحصائية ٢٧٩

الفصل السابع: التغير عبر الزمن: البيانات التعاقبية (التاريخية)

- (٧, ١) ما الذي يتناوله هذا الفصل؟ ٢٨١
- (٧, ٢) قياس الزمن وتمثيله بوصفه متغيرًا ٢٨٢
- (٧, ٣) إيجاد الاختلافات وتفسيرها: تغيّر النسبة المئوية واختبار التمهيد ٢٩٢
- (٧, ٤) تجميع الفترات الزمنية: تحليل التجميع المجاور ٢٩٩
- (٧, ٥) نمذجة التغيرات في الخطاب: القمم والقعور وتحليل التقلب الاستعمالي ٣٠٥
- (٧, ٦) التطبيق: الألوان في القرن السابع عشر ٣١٢
- (٧, ٧) تمارين ٣١٦
- قراءات إضافية ٣٢٠
- أدوات لانكستر الإحصائية ٣٢٠

الفصل الثامن: الاختتام: عشرة مبادئ للتفكير الإحصائي والتحليل البعدي وأحجام التأثير

٣٢١	(١, ٨) ما الذي يتناوله هذا الفصل؟.....
٣٢٢	(٢, ٨) عشرة مبادئ للتفكير الإحصائي.....
٣٣٣	(٣, ٨) التحليل البعدي: التوليف الإحصائي لنتائج الأبحاث.....
٣٤٣	(٤, ٨) أحجام التأثير: دليل للاستخدام الصحيح.....
٣٥٠	(٥, ٨) تمارين.....
٣٥٢	قراءات إضافية.....
٣٥٢	أدوات لانكستر الإحصائية.....
٣٥٣	ملحوظات ختامية.....
٣٥٥	قائمة المراجع.....
٣٦٣	ثبت المصطلحات: أولاً: (عربي-إنجليزي).....
٣٧٥	ثانياً: (إنجليزي-عربي).....
٣٨٧	كشاف الموضوعات.....

قائمة الأشكال

- شكل (١, ١). العلاقة بين التكرار النسبي للصفات والتكرار النسبي للأفعال ٦
- شكل (١, ٢). عملية التحليل الإحصائي ٨
- شكل (١, ٣). مثال على مجموعة البيانات ٩
- شكل (١, ٤). توزيع ضمير المتكلم (I) في مدونة تريتي لانكستر ١٢
- شكل (١, ٥). التوزيع الطبيعي القياسي ١٢
- شكل (١, ٦). توزيع تكرارات الصفات في ١١ ملفاً نصياً ١٤
- شكل (١, ٧). حالتان لمجالات الثقة ١٩
- شكل (١, ٨). التصميمات البحثية في لغويات المدونة الحاسوبية ٢٩
- شكل (١, ٩). خطاطة أعمدة: المتغير س في ثلاث مدونات ٣٢
- شكل (١, ١٠). مخطط رسم مربع: المتغير س في ثلاث مدونات ٣٣
- شكل (١, ١١). مخطط رسم مربع للمتغير س في ثلاث مدونات ٣٣
- شكل (١, ١٢). مخطط درجي: أداة التعريف في مدونة الإنجليزية البريطانية الفرعية BE06 ٣٤
- شكل (١, ١٣). مخطط درجي: كلمة fuck وتصريفاتها في المدونة الوطنية البريطانية الفرعية BNC64 ٣٥
- شكل (١, ١٤). مخطط تشتت لأداة التعريف the وضمير المتكلم I في مدونة BNC64 ٣٦
- شكل (١, ١٥). مخطط تشتت: I وyou في مدونة BNC64 ٣٧
- شكل (١, ١٦). أكثر الأماكن ارتباطاً باستخدام go وtravel في المدونة الوطنية البريطانية ٣٧
- شكل (١, ١٧). أنواع أخرى للتمثيل ٣٨

- شكل (١٨, ١). استخدام الصفات من قبل الكتّاب الأكاديميين والروائيين: مخطط رسم المربع ٤١.
- شكل (١٩, ١). استخدام الصفات من قبل الكتّاب الروائيين والأكاديميين: أعمدة الخطأ ٤١
- شكل (٢٠, ١). جزيرة بريطانيا العظمى ٤٣
- شكل (٢, ١). توزيع تكرارات الكلمة في المدونة الوطنية البريطانية ٥٨
- شكل (٢, ٢). حساب الانحراف المعياري لمدونة مكونة من ستة أجزاء ٦٣
- شكل (٢, ٣). توزيع الكلمة الأولى والكلمة الثانية ٧٠
- شكل (٣, ١). مدى التكرار والحصرية ٩٧
- شكل (٣, ٢). الشكل البياني للتصاحب love (الحب) في مدونة BE06 (قيمة لوج دايس الدنيا ٧، يم-3-يس3، تكرار المصاحبة ٥ وتكرار التصاحب ٥) ١٠٠
- شكل (٣, ٣). شبكات التصاحب: توضيح المفهوم ١٠١
- شكل (٣, ٤). مُصاحبات الدرجة الترتيبية الثالثة لكلمة time (الوقت) في مدونة LOB (قيمة المعلومات المتبادلة الدنيا ٥، يم-5-يس5: لم تُحدّد أي نقاط قيم فاصلة) ١٠٢
- شكل (٣, ٥). الشبكة التصاحبية لكلمة university (جامعة) استنادًا إلى مدونة BE06 (قيمة المعلومات المتبادلة الدنيا ٣، يم-5-يس5) ١٠٢
- شكل (٣, ٦). شبكة التصاحب لكلمة immigrants (المهاجرين) في الجارديان (قيمة المعلومات المتبادلة الدنيا ٦، يم-5-يس5، تكرار المتصاحب ١٠ وتكرار التصاحب ١٠. لم تُحدّد أي نقاط قيم فاصلة) ١٢٣
- شكل (٣, ٧). شبكة التصاحب لكلمة immigrants (المهاجرين) في ديلي ميل (قيمة المعلومات المتبادلة الدنيا ٦، يم-5-يس5، تكرار المتصاحب ٢٠ وتكرار التصاحب ٢٠. لم تُحدّد أي نقاط قيم فاصلة) ١٢٣
- شكل (٣, ٨). شبكات التصاحب المختارة ١٢٧
- شكل (٤, ١). أداة التعريف وأداة التنكير في المدونة الوطنية البريطانية الفرعية ١٣٥
- شكل (٤, ٢). مقتطف من مجموعة بيانات the مقابل a(n): تصميم الخاصية اللغوية ١٣٦
- شكل (٤, ٣). مخطط لنوع الأداة حسب التحديد السياقي ١٤٢

- شكل (٤, ٤). المخطط الأساسي للانحدار اللوجستي ١٥٥
- شكل (٤, ٥). عينة من استخدام الأداة في اللغة الإنجليزية من مجموعة بيانات ١٥٩
- شكل (٤, ٦). صُححت جملة من هذا الكتاب نحوياً ١٦٩
- شكل (٤, ٧). تمثيل العلاقة بين which و that والفاصلة ١٧٢
- شكل (٤, ٨). need to و have to و must في الإنجليزية البريطانية (BE06) ١٧٦
- شكل (٥, ١). الأسماء والصفات في مدونة BE06 ١٨٢
- شكل (٥, ٢). الأفعال والصفات في مدونة BE06 ١٨٣
- شكل (٥, ٣). الضمائر وأدوات العطف في مدونة BE06 ١٨٣
- شكل (٥, ٤). خمس نقاط بيانات للارتباط ١٨٦
- شكل (٥, ٥). التغيرات للارتباط ١٨٦
- شكل (٥, ٦). الدلالة الإحصائية (<0.05) لارتباطات بيرسون وعلاقتها بعدد المشاهدات (الملحوظات) ١٨٨
- شكل (٥, ٧). مخطط التبعر متعدد ١٩٣
- شكل (٥, ٨). مصفوفة الارتباط: الأسماء والصفات والأفعال والضمائر وأدوات العطف ١٩٤
- شكل (٥, ٩). مصطلحات الألوان في المدونة الوطنية البريطانية ١٩٦
- شكل (٥, ١٠). الخطوات الأربع لإنشاء التجميعات ٢٠١
- شكل (٥, ١١). النتيجة النهائية لإنشاء التجميعات ٢٠١
- شكل (٥, ١٢). مصطلحات الألوان: مخطط تفرعي بقيمة-2 المعيرة والمسافة الإقليدية وطريقة الربط المنفرد ٢٠٢
- شكل (٥, ١٣). مخطط تفرعي لطريقة الربط المنفرد ٢٠٢
- شكل (٥, ١٤). مخطط تفرعي لطريقة الربط التام ٢٠٣
- شكل (٥, ١٥). مخطط تفرعي لطريقة الربط المتوسط ٢٠٣
- شكل (٥, ١٦). مخطط تفرعي وفق منهج Ward ٢٠٤
- شكل (٥, ١٧). مجموعة بيانات للتحليل متعدد الأبعاد (مقتطف صغير) ٢١١

- شكل (١٨, ٥). اختزال البيانات من ١٠ متغيرات إلى عاملين ٢١٣
- شكل (١٩, ٥). تدوير بروماكس للعامل ٢١٤
- شكل (٢٠, ٥). استخلاص العامل: مخطط ركامي ٢١٥
- شكل (٢١, ٥). متوسط قيم السجلات اللغوية الموضوعة على البُعد 1: التواصلية مقابل المعلوماتية ٢١٨
- شكل (٢٢, ٥). مصفوفة الارتباط: ٤٤ متغيرًا ٢٢٣
- شكل (٢٣, ٥). الارتباط بين متوسط طول الكلمة والاختصارات: تجميعات السجل اللغوي ٢٢٤
- شكل (٢٤, ٥). مخطط التجميع: سجلات لغوية في الإنجليزية النيوزيلندية ٢٢٥
- شكل (٢٥, ٥). البعد الأول: الإنجليزية ٢٢٨
- شكل (٢٦, ٥). البعد الثاني: الإنجليزية ٢٢٨
- شكل (٢٧, ٥). العلاقة بين متوسط طول الكلمة (عدد الأحرف) ومتوسط طول الجملة (عدد الكلمات) في المدونة الوطنية البريطانية ٢٢٩
- شكل (٢٨, ٥). العلاقة بين استخدام الماضي والمضارع في مدونة BE06 ٢٣٠
- شكل (٢٩, ٥). العلاقة بين استخدام الصفات ومصطلحات اللون في مدونة BE06 ٢٣٠
- شكل (٣٠, ٥). العلاقة بين طول النص (الكلمات الفعلية) ونسبة الكلمات النوعية/ الفعلية (TTR) في المدونة الوطنية البريطانية ٢٣١
- شكل (٣١, ٥). البعد الثالث ٢٣٢
- شكل (٣٢, ٥). البعد الرابع ٢٣٢
- شكل (١, ٦). توزيع الضمائر الشخصية في مدونة BNC64 للمتحدثات ٢٤٢
- شكل (٢, ٦). حساب تحليل التباين: التباين بين المجموعة (العلوي)، التباين داخل المجموعة (السفلي) ٢٤٩
- شكل (٣, ٦). مجموعة بيانات من مدونة BNC64 والتكرارات النسبية والرتب لاستخدام الضمائر الشخصية ٢٥١
- شكل (٤, ٦). توزيع ain't في مدونة BNC64 للمتحدثين بتأثير الطبقة الاجتماعية ٢٥٥

- شكل (٦, ٥). مجال الثقة لـ ain't في مدونة BNC64 ٢٥٦
- شكل (٦, ٦). مخطط التناظر لفئات الكلمات في خطاب المتحدثين الفرديين ٢٥٩
- شكل (٦, ٧). ملفات تعريف المتحدث (الصف): المسافة الإقليدية ٢٦٢
- شكل (٦, ٨). مسافة كاي تربيع لملف تعريف المتحدث (الصف) ٢٦٣
- شكل (٦, ٩). نموذج من مجموعة البيانات اللغوية الاجتماعية: العوامل الداخلية والخارجية ٢٦٧
- شكل (٦, ١٠). المخرجات لنموذج التأثير الممتزج ٢٦٨
- شكل (٦, ١١). تحليل التناظر لاستخدام أمناء الصحافة بالبيت الأبيض فئات الكلمات ٢٧٤
- شكل (٦, ١٢). تحليل التناظر لاستخدام العلامات المعرفية في BNC64 ٢٧٧
- شكل (٧, ١). الأفعال الناقصة في مدونات عائلة براون ٢٨٢
- شكل (٧, ٢). تفسير بديل لنماذج في مدونات عائلة براون ٢٨٦
- شكل (٧, ٣). مشاهد نغرامية جوجل لكلمة (الرجل) وكلمة (المراة) ٢٨٧
- شكل (٧, ٤). الأفعال الناقصة في مدونات عائلة براون: (القمة) الأصلية و(القاع) معاد قياسه ٢٨٨
- شكل (٧, ٥). الأفعال الناقصة في اللغة الإنجليزية البريطانية: (أ) مخطط رسم المربع، (ب) أعمد الخطأ بمجال الثقة ٩٥٪ ٢٩٠
- شكل (٧, ٦). مخطط الشموع: تطور الأفعال الناقصة الفردية بين ١٩٣١-٢٠٠٦ ٢٩١
- شكل (٧, ٧). التمهيد: وصف المفهوم ٢٩٤
- شكل (٧, ٨). مثال على مجموعة بيانات اختبار التمهيد لـ: its في مدونة EEBO ٢٩٦
- شكل (٧, ٩). نقاط البيانات مع مرور الوقت: مثال مبتكر ٢٩٩
- شكل (٧, ١٠). مبدأ التجميع: (أ) تجميع التكتل الهرمي. (ب) تجميع الجوار القائم على المتغيرية ٣٠٠
- شكل (٧, ١١). المخططات التفرعية: (أ) تجميع التكتل الهرمي، (ب) تجميع الجوار القائم على المتغيرية ٣٠٢
- شكل (٧, ١٢). مخطط تفرعي لاستخدام ضمير الملكية its في القرن السابع عشر ٣٠٣

- شكل (١٣, ٧). مخطط تفرّعي لاستخدام ضمير الملكية its في القرن السابع عشر ٣٠٤
- شكل (١٤, ٧). الشكل البياني للقمم والقعور الناتجة ٣٠٩
- شكل (١٥, ٧). نتائج تحليل التقلّب الاستعمالي لكلمة war (حرب) بين عامي 1940 و2009 (قيمة المعلومات المتبادلة الدنيا ٦، ٥-٥، تكرار المتصاحب النسبي ١٠ وتكرار المتصاحب النسبي (١٠) ٣١١
- شكل (١٦, ٧). تكرار مصطلحات الألوان في القرن السابع عشر ٣١٣
- شكل (١٧, ٧). مخطط الشموع: الألوان في القرن السابع عشر ٣١٤
- شكل (١٨, ٧). تحليل التقلّب الاستعمالي لكلمة red (أحمر) بين عامي 1600 و1699 (قيمة المعلومات المتبادلة الدنيا ٣، ٥-٥، تكرار المتصاحب النسبي ١٠ وتكرار المتصاحب النسبي (١٠) ٣١٥
- شكل (١٩, ٧). تجميع الجوار القائم على المتغيرة VNC: كلمة red (أحمر) في القرن السابع عشر ٣١٦
- شكل (٢٠, ٧). عدد التغيرات المرتبطة ببرنامج X Factor البريطاني (١٦ نوفمبر ٢٠١٤ م ٧-١١ مساءً) ٣١٦
- شكل (٢١, ٧). تطور تكرارات handsome (وسيم)، وpretty (مليح)، وbeautiful (جميل) متبوعة بالذكر M: أو بالأنثى F في القرن السابع عشر ٣١٧
- شكل (٢٢, ٧). تطور تكرارات ضمير الملكية its في القرن السابع عشر ٣١٧
- شكل (٢٣, ٧). أربعة سيناريوهات لتغير التكرار ٣١٨
- شكل (٢٤, ٧). كلمة handsome (وسيم) في القرن السابع عشر ٣١٩
- شكل (٢٥, ٧). كلمة pretty في القرن السابع عشر ٣١٩
- شكل (١, ٨). نظرة عامة على الأنواع في BE06 (Baker 2009) ٣٢٥
- شكل (٢, ٨). الزمن الماضي في أنواع الكتابة المختلفة من مدونة BE06 ٣٢٥
- شكل (٣, ٨). مخطط رسم المربع في أنواع مكتوبة مختلفة BE06: (أ) الزمن الماضي، (ب) الزمن المضارع ٣٣٢

قائمة الأشكال

ذ

- شكل (٤, ٨). البحث عن مسرح غلوب ٣٣٥
- شكل (٥, ٨). مخطط غابي للتحليل البعدي لأربع دراسات..... ٣٤٢
- شكل (٦, ٨). مقارنة بين مدونتين فرعيتين ٣٤٨
- شكل (٧, ٨). مخطط غابي للمثال الأول ٣٥١
- شكل (٨, ٨). مخطط غابي للمثال الثاني ٣٥١

قائمة الجداول

جدول (١, ١). حجم التأثير (r) وتفسيره القياسي.....	١٩
جدول (١, ٢). إطار عينة عائلة براون.....	٢١
جدول (١, ٣). تكرار الكلمات والتعبيرات المختارة في ثلاث مدونات باللغة الإنجليزية.....	٢٦
جدول (١, ٤). مستويات مختلفة من التحليل في لغويات المدونة الحاسوبية.....	٢٧
جدول (١, ٥). المدونتان الفرعيتان في بحث مصغر.....	٣٩
جدول (٢, ١). الكلمة النوعية والمادة المعجمية واللكسيم: المزايا والعيوب.....	٥٣
جدول (٢, ٢). أكثر ١٠ كلمات تكررًا في المدونة الوطنية البريطانية.....	٥٤
جدول (٢, ٣). مدونة مكونة من مليون كلمة فعلية.....	٦١
جدول (٢, ٤). حساب توزيع النسب في المدونة المكونة من ستة أجزاء.....	٦٨
جدول (٢, ٥). المدونة BE06.....	٧٧
جدول (٢, ٦). المواد المعجمية المرتبطة بالطقس في مدونة BE06.....	٧٨
جدول (٢, ٧). رُتب المواد المعجمية المتعلقة بالطقس في مدونة BE06.....	٨٠
جدول (٢, ٨). المدونة الوطنية البريطانية: توزيع أربع كلمات مختارة.....	٨٣
جدول (٣, ١). التكرارات الملحوظة.....	٩٢
جدول (٣, ٢). التكرارات المتوقعة: منطلق الورود العشوائي.....	٩٤
جدول (٣, ٣). مقاييس التصاحب: نظرة عامة.....	٩٥
جدول (٣, ٤). ترتيب مُصاحبات (جديد: new) في BE06 (يس-3-يم3).....	٩٦
جدول (٣, ٥). مجموعة بارامترات المصاحبة collocation parameters notation.....	٩٨

- جدول (٦, ٣). الكلمات المفتاحية الإنجليزية الأمريكية من مدونة Ame06 ١٠٣
- جدول (٧, ٣). قرارات إجراء الكلمة المفتاحية: خيارات أساسية..... ١٠٥
- جدول (٨, ٣). مقارنة عناصر معجمية مختارة في مدونتي BE06 و Ame06..... ١٠٩
- جدول (٩, ٣). الكلمات المفتاحية في اللغة الإنجليزية الأمريكية: تطبيقات مختلفة لتحديد الكلمة المفتاحية..... ١١٢
- جدول (١٠, ٣). مدونة BE06: كشافات سياقية مختارة للعقدة (الدين: religion): أي الكشافات السياقية إيجابية وأيها سلبية؟ ١١٥
- جدول (١١, ٣). الترميز المزدوج للكشافات السياقية ١١٦
- جدول (١٢, ٣). نظرة عامة على مقاييس ملاءمة المقدّر البيني ١١٩
- جدول (١٣, ٣). الكلمات المفتاحية..... ١٢١
- جدول (١٤, ٣). تقييمات immigrant(s) في مدونتي الجارديان وديلي ميل ١٢٥
- جدول (١٥, ٣). مُتصاحبات كلمة issue في مدونة BE06 ١٢٦
- جدول (١٦, ٣). الكلمات المفتاحية..... ١٢٨
- جدول (١٧, ٣). أمثلة للتصنيف ١٣٠
- جدول (١, ٤). أمثلة على المتغيرات المعجمية النحوية ذات الإطار النحوي..... ١٣٩
- جدول (٢, ٤). الجدولة المتقاطعة لنوع الأداة حسب التحديد السياقي ١٤٢
- جدول (٣, ٤). النسبة المئوية للخيارات في الجدولة المتقاطعة ١٤٤
- جدول (٤, ٤). الأفعال (شبه) الناقصة الإلزامية الجازمة في مختلف الأوعية في الإنجليزية البريطانية والأمريكية ١٤٥
- جدول (٥, ٤). التكرارات المتوقعة لنوع الأداة حسب التحديد السياقي ١٤٨
- جدول (٦, ٤). تفسير معامل كيرمر ف ١٤٩
- جدول (٧, ٤). الاحتمالات لنوع الأداة حسب التحديد السياقي ١٥١
- جدول (٨, ٤). الاحتمالات لنوع الأداة حسب نوع الاسم ١٥٧
- جدول (٩, ٤). نظرة عامة عن النماذج ١٦١

- جدول (١٠، ٤). أخطاء المعيار الكبيرة لجزء من ناتج الانحدار اللوجستي ١٦١
- جدول (١١، ٤). الجدولة المتقاطعة لاستخدام الفاصلة مع ضميري الوصل which و that ١٧٠
- جدول (١٢، ٤). الجدولة المتقاطعة لـ which و that في مواقف سياقية مختلفة ١٧٣
- جدول (١٣، ٤). تقديرات الانحدار اللوجستي لكل من which و that ١٧٤
- جدول (١٤، ٤). تصميم البحث المناسب ١٧٥
- جدول (١، ٥). رُتب الأسماء والصفات في خمسة نصوص من مدونة BE06 ١٩٠
- جدول (٢، ٥). جدول الارتباط (ارتباطات بيرسون) للأسماء والصفات والأفعال والضمائر وأدوات العطف ١٩٢
- جدول (٣، ٥). المجموعة الكاملة لخصائص باير (١٩٨٨) القائمة على كونراد وباير (٢٠٠١): ١٨-١٩ ٢٠٨
- جدول (٤، ٥). العامل 1: تحميلات المتغيرات الفردية ٢١٧
- جدول (٥، ٥). السجلات اللغوية في مدونة ICE-NZ ٢٢٠
- جدول (٦، ٥). نتائج تحليل العامل للغة الإنجليزية النيوزيلندية: تحميلات العامل ٢٢٦
- جدول (٧، ٥). نتائج تحليل عوامل الإنجليزية النيوزيلندية: تحميلات العامل (3) و (4) ٢٣٢
- جدول (١، ٦). الجدولة المتقاطعة لفئات الكلمات في خطاب المتحدثين الفرديين ٢٥٩
- جدول (٢، ٦). الجدولة المتقاطعة للأفعال والأدوات في خطاب متحدثين ٢٥٩
- جدول (٣، ٦). ملفات التعريف للمتحدثين ٢٦١
- جدول (٤، ٦). مدونة البيت الأبيض ٢٧٣
- جدول (٥، ٦). القَسَم والجنس: مدونة BNC64 ٢٧٨
- جدول (١، ٧). مقارنة فترتين في مدونة EEBO: الكومنولث والمحمية & Commonwealth Protectorate والاسترداد Restoration ٢٩٣
- جدول (٢، ٧). نتائج التمهيد لمقارنة فترتين في مدونة EEBO ٢٩٧
- جدول (٣، ٧). التقسيم النهائي للنتائج: must و its و pestilence ٢٩٧
- جدول (٤، ٧). التكرار النسبي (لكل مليون كلمة) لضمير الملكية its في القرن السابع عشر ... ٣٠٢

جدول (٥, ٧). التكرار النسبية (لكل مليون كلمة) لكلمة war	٣٠٧
جدول (٦, ٧). الفروق بين التكرارات النسبية لكلمة war	٣٠٧
جدول (٧, ٧). التكرار النسبي ذو اللوغاريتم المحول (لكل مليون) لكلمة war	٣٠٧
جدول (٨, ٧). ملف تعريف مصاحبات war	٣١٠
جدول (١, ٨). استخدام الزمن الماضي والحاضر في سجلات لغوية مختلفة (مجموعة البيانات الأصلية)	٣٢٢
جدول (٢, ٨). استخدام الزمن الماضي والمضارع في سجلات لغوية مختلفة (تقرير بحثي)	٣٢٢
جدول (٣, ٨). أمثلة على استخدام الزمن الماضي في الكتابة الأكاديمية وأدب الغموض	٣٢٦
جدول (٤, ٨). أمثلة على نصوص من الكتابة الأكاديمية وأدب الغموض	٣٢٩
جدول (٥, ٨). موجز الأسئلة	٣٣٤
جدول (٦, ٨). دراسات رُوِجعت للتحليل البعدي	٣٤٠
جدول (٧, ٨). بيانات الإدخال للتحليل البعدي البسيط	٣٤١
جدول (٨, ٨). قياسات حجم التأثير الواردة في هذا الكتاب	٣٤٤
جدول (٩, ٨). تحويل حجم التأثير والاستخلاص الرياضي	٣٤٦
جدول (١٠, ٨). حجم التأثير: التفسير المعياري	٣٤٧
جدول (١١, ٨). قياسات حجم التأثير: التحقق من صلاحية المدونة الوطنية البريطانية	٣٤٩
جدول (١٢, ٨). تحويلات حجم التأثير	٣٥١

الفصل الأول

لغويات المدونة الحاسوبية في ضوء علم الإحصاء

(١, ١) ما الذي يتناوله هذا الفصل؟

يتناول هذا الفصل مبادئ التفكير الإحصائي الضرورية لتطبيق العمليات الإحصائية على بيانات المدونة الحاسوبية تطبيقاً دقيقاً. ويبدأ الفصل بشرح دور الإحصاء في البحث العلمي عامةً وفي لغويات المدونة الحاسوبية corpus linguistics خاصةً، ثم يُعرِّج على موضوعات أخصّ وأقرب صلةً، وهي: إنشاء المدونات، وأنواع تصميم البحث، والمصطلحات الإحصائية الأساسية، واستكشاف البيانات وتمثيلها، لينتهي بعدئذٍ بدراسة حالة معينة توضح طبيعة استخدام علم الإحصاء statistics في أبحاث المدونات.

سنُجيب في هذا الفصل عن الأسئلة الخمسة الآتية:

- ١- ما دور علم الإحصاء في الأبحاث العلمية وأبحاث المدونات؟ (المبحث ٢, ١).
- ٢- ما المصطلحات الأساسية في إحصائيات المدونات؟ (المبحث ٣, ١).
- ٣- كيف يمكننا بناء المدونات وإنشاءها؟ (المبحث ٤, ١).
- ٤- كيف يمكننا استكشاف البيانات وتمثيلها؟ (المبحث ٥, ١).
- ٥- كيف يمكن استخدام علم الإحصاء في أبحاث المدونات؟ (المبحث ٦, ١).

(٢, ١) ما المقصود بعلم الإحصاء؟ العلوم ولغويات المدونة الحاسوبية وعلم الإحصاء

فكّر...

قبل قراءة هذا المبحث فكر في الأسئلة التالية:

١- ما المقصود بالعلم؟ ما السمات الأساسية للمبحث العلمي؟

٢- ما البيان الذي يُعدّ علمياً لتفسير اللغة؟

(أ) يغلبُ على خطاب المرأة بوجه عام أن يتضمن تعبيرات لفظية من قبيل:

(حسناً) و(كما تعلم) و(إلى حدّ ما) وما إلى ذلك ...

(ب) الكلام سهل كالهواء.

(ج) البناء للمجهول أكثر شيوعاً بكثير في النشر الأكاديمي [عما في السجلات

اللغوية الأخرى]، إذ يتكرر قرابة ٥٠٠, ١٨ مرة في كل مليون كلمة.

(د) من المنطقي اعتبار القدرة اللغوية (عضوًا لغويًا).

(هـ) تُظهر نتائجنا أنه حدثت تغيرات كبيرة في نطاق رنيني^(١) واحد على الأقل لـ

١٠ من ١١ صائتًا، وفي كلا النطاقين الرنينيين لـ ٥ من ١١ صائتًا من خمسينيات

وحتى ثمانينيات القرن العشرين ... نخلص إلى أنّ الملكة لم تعد تتحدث إنجليزية

الملكات في الخمسينيات ...

خلافاً لمصادر المعلومات الأخرى مثل الأساطير أو الفلسفة أو الفنون، تعتمد العلوم على الجمع المنهجي الذي يحقق ما يفيد تجريبياً ويختبر النظريات والفرضيات. وقد عرّف كارل بوبر، أحد أبرز المنظرين في حقل العلوم، الاتجاه العلمي أو النظرية العلمية بأنها شيء يمكن دحضه من حيث المبدأ (Popper, 2005 [1935])، وعليه فإنه لا يمكننا أن نصف نظرية بأنها علمية وعملية إلا إذا كان من الممكن اختبارها تجريبياً، وهذا يحتاج منا إلى جمع البيانات، وتقييم ما إذا كانت البيانات متوافقة مع نظريتنا، وإن لم تكن متوافقة فإن الأدلة المتاحة من البيانات حيثتد تعارض مع النظرية. وإذا نظرنا إلى البيانات الواردة في السؤال الثاني من مهمة (فكّر)، نلاحظ أنها تختلف وتتنوع إلى حد كبير من حيث

(١) النطاق الرنيني مكوّن من صائت في البحث الصوتي.

قابلية إخضاعها للاختبار من خلال جمع الأدلة التجريبية، فيمكن اعتبار ما قيل في (ج) و(هـ) علميين^(٢)، ليس لأنهما قابلان للاختبار تجريبيًا وحسب، بل لأنهما مشفوعان بأدلة تجريبية. ومن ناحية أخرى، فإن البيان الشعري (ب) عبارة عن استعارة، والتي على الرغم من قوتها سيكون من الصعب اختبارها من خلال جمع البيانات. يمكن اختبار البيان (أ) المأخوذ من كتاب لاكوف Lakoff (١٩٧٥): (اللغة ومكانة المرأة) Language and Woman's Place، تجريبيًا، والذي اختبره كثير من الباحثين بالفعل، ومع ذلك لا تقدم المؤلفة نفسها أي أدلة تجريبية في كتابها غير الحكايات. ويقترح البيان (د) لتشومسكي Chomsky (٢٠٠٠) وجهة نظر لغوية تعتمد بنحو أكبر على الفهم الفلسفي (ما قبل التجريبي) للقدرة اللغوية البشرية ولا تسعى بالضرورة إلى تأكيد تجريبي. وبناء على كل تلك البيانات عن اللغة، فإن البيان العلمي هو ذلك الذي يُمهد مباشرة إلى أدلة تجريبية تُجمع وتُختبر بأسلوب منظم^(٣).

يشير مصطلح: لغويات المدونة الحاسوبية إلى الطريقة العلمية التي بها تُحلّل اللغة، إذ تتطلب من المحلّل تقديم أدلة تجريبية في صورة بيانات مستمدة من مدونات لغوية وتدعم أي بيان يتعلق باللغة، وتتطلب أمرًا آخر يتبعه لغويو المدونة من حيث المبدأ، ألا وهو قابلية تكرار النتائج، أي: القدرة على تأكيد نتائج دراسة واحدة في دراسات متتابعة من الباحثين (المبحث ٣، ٨). ومن أجل أن تكون النتائج قابلة للتكرار. يتعين على لغويي المدونة أن تكون اختياراتهم للمدونات والتقنيات التحليلية محددة وصریحة. ومن الممارسات البحثية الدقيقة أيضًا في لغويات المدونة الحاسوبية إتاحة الفرصة للباحثين الآخرين الذين يمكنهم استكشاف مجموعة البيانات نفسها بدقة، الأمر الذي يساعد على تعزيز المعرفة في المجال الخاضع للتحليل^(٤).

(٢) انظر إلى المصادر الآتية: لاكوف Lakoff (١٩٧٥: ٥٣)، وشكسبير Shakespeare (١٩٩٢ [١٥٩٩]: ٢٦٩)، وبايبر وآخرين Biber et al. (١٩٩٩: ٤٧٦)، وتشومسكي Chomsky (٢٠٠٠: ٤)، وهارينغتون وآخرين Harrington et al. (٢٠٠٠: ٩٢٧).

(٣) يمكن أن يكون البحث التجريبي نوعيًا ووصفيًا وتفسيريًا من جانب، أو أن يكون كميًا باستخدام الأرقام من جانب آخر، ويعد هذان الجانبان مكملين لبعضهما بعضًا.

(٤) لا تُتاح المدونات الرسمية في بعض الأحيان للأسف، ولا يُفصح عن طريقة بنائها، ويدعو ذلك إلى غض الطرف عن استخدامها في الأبحاث العلمية الجادة، ويُجعل من قراء البحث المعتمد على هذا النوع من المدونات في شك من الأمر. وعليه فإن حقل لغويات المدونة الحاسوبية يدعو الباحثين إلى ألا يتضمن بحثهم عبارات من قبيل: هذا المثال اللغوي مذكور في مدونة ضخمة ليست متوافرة.

يقوم جوهر لغويات المدونة الحاسوبية على المنهج الكمي، وهذا يعني انطواءها عادةً على أرقام تعكس تكرارات الكلمات والمركبات في المدونات (McEnery & Hardie 2011). ولهذا السبب يُعدّ علم الإحصاء ضروريًا بالنسبة إلى لغويي المدونة لأنه يساعد على التعامل بفعالية مع المعلومات الكمية. ولإطار علم الإحصاء أطر عديدة، والإطار المطبق في هذا الكتاب هو الجانب الذي يُعِين على فهم البيانات اللغوية الكمية بعد جمعها وتفسيرها (Diggle & Chetwynd 2011: vii)، ويُلمّز عند التفسير قياس هذه البيانات أو تقدير كميتها بطريقة محددة. ومن الأدوات المهمة في علم الإحصاء استخدام التعبيرات الرياضية، ولهذا السبب فإننا سنقف على معادلات مختلفة في هذا الكتاب، إذ تُعِين تلك التعبيرات على فهم الحقائق المعقدة والغامضة من خلال الكشف عن سمات مهمة للبيانات من طريق الأرقام والرموز، الأمر الذي يجعل من التعامل مع تحليل البيانات يسيرًا وواضحًا.

لِنُلْقِ نظرة على مثالين يوضحان هذه النقطة. تَحْيَلُ أننا مهتمون بعدد الصفات التي يستخدمها كُتّاب الرواية البريطانيون في نصوصهم. قد نفترض بأن استخدام المزيد من الصفات يؤدي إلى المزيد من الأوصاف المفصلة واقعيًا في الروايات. اخترنا أحد عشر نصًا روائيًا عشوائيًا لمؤلفين مختلفين من المدونة الوطنية البريطانية (BNC)، وحسبنا عدد الصفات في كل نص (انظر إلى عدد التكرار المطلق absolute frequency في المبحث ٣، ٢). أجرينا بعد ذلك معايرة التكرار المطلق من أجل القابلية للمقارنة^(٥)، وتُسمّى تلك النصوص الأحد عشر عينة sample. وفيما يلي التكرارات النسبية للصفات في كل 10,000 كلمة:

508, 542, 552, 553, 565, 567, 570, 599, 656, 695, 699

لا يمكن عرض قائمة طويلة من النتائج على أي حال، وليس ذلك فعلاً عند التعامل مع البيانات الكمية، فلك أن تتخيل حجم النتائج إذا كان علينا إدراج 100 أو 1000 نتيجة مختلفة. وبدلاً عن ذلك، يمكننا استخدام مقياس إحصائي بسيط للغاية لتلخيص نتائجنا يُعرف بـ: المتوسط

(٥) تختلف النصوص من ناحية الطول، وعليه فقد سجلنا التكرارات النسبية لكل 10,000 كلمة لإظهار متوسط عدد الصفات التي يستخدمها كل مؤلف في ذلك العدد (انظر المبحث ٣، ٢ للاطلاع على التكرار النسبي)، وقد قُرِبَت التكرارات النسبية إلى أقرب عدد صحيح.

mean، ويعطينا قيمة متوسطة تمثل مجموعة كاملة من القيم values، ووسيط الأرقام المذكورة سابقاً هو (591.45) كما في المعادلة (١، ١) الآتية:

$$\text{مُحَسَّب المتوسط بالطريقة التالية: المتوسط} = \frac{\text{مجموع القيم}}{\text{عدد الحالات}}$$

وتطبيقاً على مجموعة البيانات الواردة أعلاه فإن:

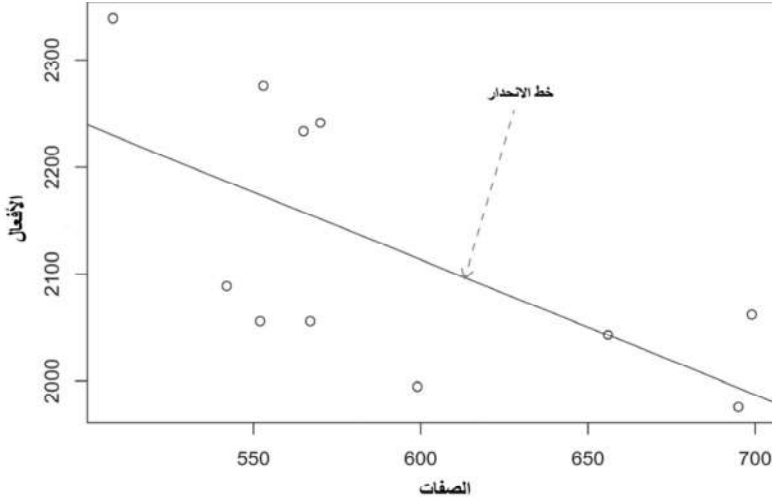
$$\text{المتوسط} = \frac{508+542+552+553+565+567+570+599+656+695+699}{11} = 591.45$$

ونظراً إلى ما يصفه الوسيط من العينة فإنه يُعدّ جزءاً مما نسميه علم الإحصاء الوصفي descriptive statistics. ومن الأمثلة الأخرى على التمثيل الرياضي للواقع اللغوي المعقد خطياً، الذي يسمى في علم الإحصاء بـ: خط الانحدار regression line أو خط التوافق الأفضل line of the best fit (انظر الفصل الرابع لمطالعة شرح نماذج الانحدار)، تصوّر بأننا مهتمون بما إذا كان المؤلفون الذين يستخدمون المزيد من الصفات يستخدمون أيضاً المزيد من الأفعال. يمكننا سرد تكرارات الأفعال^(٦) أسفل تكرارات الصفات مباشرة لمعرفة ما إذا كانت هناك أي علاقة بين هاتين الخاصيتين اللغويتين:

التكرار											
٥٠٨	٥٤٢	٥٥٢	٥٥٣	٥٦٥	٥٦٧	٥٧٠	٥٩٩	٦٥٦	٦٩٥	٦٩٩	الصفة
٢٣٣٩	٢٠٨٩	٢٠٥٦	٢٢٧٦	٢٢٣٣	٢٠٥٦	٢٢٤١	١٩٩٥	٢٠٤٣	١٩٧٦	٢٠٦٢	الفعل

ومع ذلك، فمن أفضل الطرق لمعرفة ما إذا كان بين استخدام الصفات واستخدام الأفعال علاقة هي تمثيل هذه الأرقام في رسم بياني (يُبين المبحث ١، ٥ كيفية إنشاء الرسومات البيانية). ويوضح الشكل البياني في الشكل (١، ١) نزعة واضحة يُفسرها خط الانحدار، ويشير خط الانحدار إلى أن عدد الأفعال والصفات في العينة في علاقة تناسبية عكسية، إذ كلما زاد استخدام المؤلفين للصفات قلَّ عدد الأفعال التي يستخدمونها والعكس صحيح.

(٦) نُدَّكر القارئ هنا بأن هذه التكرارات النسبية هي لكل 10,000 كلمة وقد عُرِضَتْ وفقاً لأقرب عدد صحيح.



شكل (١, ١). العلاقة بين التكرار النسبي للصفات والتكرار النسبي للأفعال.

ويُمثّل بالشكل البياني كما هو ظاهر بهذه الطريقة للعثور على التوافق الأفضل لجميع نقاط البيانات (أي: أرقام التكرارات). ومن قبيل الصدفة البحتة وقوع إحدى النقاط على الخط بالتحديد؛ لأنه في الغالب لا يمر الخط بأيّ من نقاط البيانات لكونه تجريداً رياضياً يمثل مجموعة البيانات عموماً. ويفيدنا هذا النموذج الرياضي بشيء مهم عن مجموع البيانات التي لن نلاحظ تفاصيلها بالضرورة إن نظرنا إلى تكرارات الصفات والأفعال بالعين المجردة، وعليه فإن هذا النموذج لأمثلة الصفات والأفعال يوضح لنا الهدف من وراء العمليات الإحصائية التي سيقف هذا الكتاب على أشكال مختلفة منها. إذن: يدور علم الإحصاء في لغويات المدونة الحاسوبية حول النمذجة الرياضية لواقع لغوي معقد، ويمكن لهذه النمذجة أن تساعدنا على اكتشاف الأنماط والاتجاهات في البيانات التي تظل مخفية.

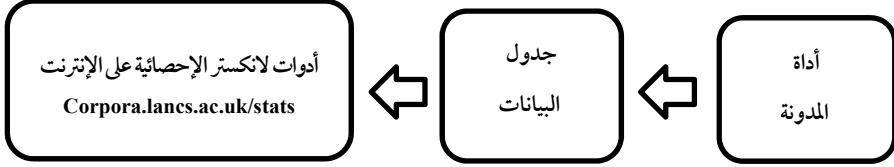
(١, ٣) مصطلحات علم الإحصاء الأساسية

فكّر...

قبل مطالعة هذا المبحث، فكّر في معاني المصطلحات التالية. هل سمعت بها من قبل؟ وإن كنت قد سمعتها من قبل، ففي أي سياق؟ هل تستطيع أن تُعرّفها؟

- الافتراض assumption
- الحالة case
- مجال الثقة confidence interval
- مجموعة البيانات dataset
- التوزيع distribution
- التشتت dispersion
- حجم التأثير effect size
- التوزيع الطبيعي normal distribution
- الفرضية الصفريّة null hypothesis
- القيمة المتطرفة outlier
- قيمة الاحتمال p-value
- (إحصاء) متين robust
- القيمة الشاذة rogue value
- القياس الإحصائي statistical measure
- الاختبار الإحصائي statistical test
- الانحراف المعياري standard deviation
- متغير variable

نلقي فيما يلي نظرة عامة على المصطلحات الإحصائية الأساسية المستخدمة في هذا الكتاب، مع ذكر أمثلة لكل مصطلح منها من أبحاث المدونة، وهي مرتبة بدءًا بالمفاهيم الأساسية ثم وصولاً إلى المفاهيم الأكثر تعقيداً التي تعتمد على فهم المصطلحات السابقة، وستعين معرفة هذه المصطلحات على فهم كثير من تطبيقات لغويات المدونات الحاسوبية بيسر وسهولة.



شكل (١, ٢). عملية التحليل الإحصائي.

المدونة corpus

(جمعها مدونات corpora): شكل محدد من أشكال البيانات اللغوية، وهي عبارة من النصوص المكتوبة أو النصوص المنطوقة التي يمكن البحث فيها بالحاسوب باستخدام برامج معينة. تُمثّل المدونة عادة عينة لغوية، أي مجموعة فرعية من إنتاج لغوي معين، ويمكن في بعض الحالات المحدودة للمدونات المتخصصة أن تشمل المدونة **المجموع اللغوي population** بأكمله، أي اللغة التي تهتم بها الباحث (المبحث ٤, ١). ينفذ البرنامج الذي يُستخدم للبحث في مدونة ما أنواعاً أساسية من التحليل الإحصائي، مثل التعريف الإحصائي للمتصاحبات اللفظية والكلمات الرئيسية (انظر: الفصل الثالث). بالنسبة للتحليلات الإحصائية الأكثر تعقيداً، نحتاج عادةً إلى استخدام حزم إحصائية مناسبة، وعليه فإن هذا الكتاب سيستخدم أدوات لانكستر الإحصائية على الإنترنت، وأدوات إحصائية مجانية متاحة في أدوات لانكستر (يوضح الشكل ١, ٢ عملية التحليل باستخدام أدوات لانكستر الإحصائية على الإنترنت).

لاحظ أن إعداد جدول البيانات بالتنسيق المناسب لا يقل أهمية عن التحليل الإحصائي، وبذلك يقدم المرجع الذي بين يديك كثيراً من أمثلة مجموعات البيانات استناداً إلى مدونات مختلفة من أجل أن يتناسب ذلك مع أنواع التحليلات المختلفة. ومن المفيد دائماً مقارنة بياناتك بالأمثلة النموذجية المقدمة من البيانات في أدوات لانكستر من أجل معرفة ما إذا كانت بياناتك قد أعدت بالتنسيق المناسب أم لا.

مجموعة البيانات dataset

هي سلسلة من النتائج القائمة على المدونة يمكن تحليلها إحصائياً. إنها مجموعة منتظمة من النتائج التي تُخزّن في جدول وفي أحد برامج جدول البيانات مثل إكسيل Excel أو كالك Calc وغيرهما، وكل سطر يمثل نقطة أو حالة بيانات **data point or case** فردية، وكل عمود يمثل متغيراً

منفصلاً. يقدم الشكل ٣، ١ مثلاً لمجموعة بيانات بها خمسة متغيرات وحالات متعددة، تمثل كل حالة متكلمًا واحدًا. من المهم بمكان أن يُلاحظ القارئ أمثلةً لمجموعات البيانات في موقع أدوات لانكستر والتآلف مع هيكل البيانات المعين الذي يساعد على تصميم أنواع أخرى من الأشكال للتحليلات الإحصائية.

المتغير variable

يُوحى الاسم بما يمكن أن يتغير ليأخذ قيمًا مختلفة. على سبيل المثال، يُعد عُمر المتحدث متغيرًا يمكن أن يأخذ قيمًا مختلفة من عام واحد تقريبًا (عندما يتعلم الأطفال عادة كلماتهم الأولى) إلى أكثر من 100 عام. يمكن وصف كثيرٍ من أبحاث المدونة بأنها تبحث عن متغيرات في المدونات وتحلل العلاقة بينها. وتدعو الحاجة إلى التمييز بين المتغيرات اللغوية والمتغيرات التأويلية، فـ: المتغيرات اللغوية linguistic variables تصف تكرارات السمات اللغوية الرئيسية بالمدونة. والمتغيرات التأويلية explanatory variables (تسمى أحيانًا المتغيرات المستقلة independent variables) تصف السياقات التي تظهر فيها السمات اللغوية. على سبيل المثال، يمكن أن يكون المتغير التأويلي هو الوعاء/المجال اللغوي، أو السجل اللغوي، أو تاريخ نشر النص، أو عمر المتحدث وجنسه ولغته.

المتغيرات التأويلية				المتغيرات اللغوية		
1	speaker_id	gender	proficiency	I	you	pers_pronouns_all
2	6_SP_51	0	1	38.75969	9.302326	80.62015504
3	6_SL_7	0	1	33.46856	19.26978	100.4056795
4	8_ME_24	1	2	39.10112	38.65169	129.8876404
5	8_IT_28	1	2	51.98181	11.04613	122.8070175
6	8_IT_14	0	2	33.41584	8.663366	108.9108911
7	IT_65	1	3	37.127	19.43095	100.9715475
8	7_CH_17	0	2	58.64198	23.91975	100.308642
9	7_ME_6	1	2	42.48573	10.14585	119.2136969
10	6_CH_15	0	1	56.12245	22.95918	145.4081633
11	IT_54	1	3	25.81369	19.01969	101.010101
12	6_ME_2	1	1	34.90401	33.15881	108.2024433
13	6_CH_25	1	1	47.82147	11.68969	145.5897981
14	CH_6	1	3	52.44601	25.1212	121.6394888
15	6_IN_3	1	1	20.83539	26.74897	131.6872428

gender: 0... male, 1... female; English proficiency: 1...pre-intermediate, 2...intermediate,

شكل (٣، ١). مثال على مجموعة البيانات (هوية المتكلم speaker_id والجنس gender والكفاية proficiency وكل

الضائرات الشخصية pers_pronouns_all)

تحتوي مجموعة البيانات في الشكل (٣، ١)، المأخوذة من مدونة تريتي لانكستر لإنتاج اللغة الثانية المنطوق (Gablasova et al. 2017) على متغيرين تأويليين (الجنس والكفاءة اللغوية) وثلاثة متغيرات لغوية (التكرارات النسبية للضمير I وyou وجميع الضمائر الشخصية معاً).

يمكن أن تكون المتغيرات (اللغوية والتأويلية على حد سواء) متغيرات اسمية أو ترتيبية أو مقياسية. يشتمل المتغير الاسمي *nominal variable* على قيم تمثل فئات مختلفة يمكن تجميع الحالات غير التسلسلية أو الهرمية فيها. ويُعدّ جنس المتحدث مثالاً للمتغير الاسمي، وعليه يُعيّن ذلك الجنس حسب المتحدثين إلى مجموعتين من دون أن تتضمن أي تسلسل هرمي تصنيفي آخر، وهاتان المجموعتان هما: المتحدثون والمتحدثات. ومن المهم أن نستخدم الأرقام للإشارة إلى هاتين المجموعتين، ففي مجموعة البيانات في الشكل (٣، ١) يشير الرقم 0 إلى (المتحدث) والرقم 1 إلى (المتحدثة)، ولكن هذه الأرقام ليس لها قيمة لكونها مجرد وسيلة رقمية للتسمية، وعليه يمكن وضع أي رقمين آخرين للإشارة إلى كل جنس من الجنسين. يشبه المتغير الترتيبي *ordinal variable* المتغير الاسمي؛ لأنه يجمع الحالات في فئات مختلفة؛ غير أن الفئات يمكن ترتيبها وفقاً لتسلسل هرمي. على سبيل المثال، تُعدّ كفاءة المتحدث في لغة أجنبية متغيراً ترتيبياً لإمكان ترتيب المتحدثين وفقاً لإجادتهم بدءاً من المبتدئين إلى المتقدمين. يُشير رقم 1 في مجموعة البيانات في الشكل (٣، ١) في عمود (الكافية proficiency) إلى كفاءة لغوية أقل من رقم 2، ويشير رقم 2 إلى كفاءة لغوية أقل من رقم 3. أما متغير المقياس *scale* فهو متغير كمي؛ لأنه يمكن أن يأخذ أي قيمة على مقياس يتضمن أرقاماً كمية لسمات محددة. يعني ذلك أيضاً أنه يمكن إضافة القيم المأخوذة إلى مثل هذه المقاييس وطرحها وضربها وقسمتها؛ لأنها تمثل كميات قابلة للقياس، وليست مجرد ترتيبات^(٧). في حالة المتغيرات اللغوية، يُظهر المقياس التكرارات النسبية لسمات لغوية في نصوص مختلفة أو من متحدثين مختلفين أو مدونات فرعية في مدونة ما، فالأرقام التي تشير إلى التكرارات النسبية لكل 1000 كلمة من ضمير المتكلم I (أنا) في الشكل (٣، ١) ما هي إلا قيم للمتغير المقياسي، وجميع المتغيرات اللغوية الثلاثة في مجموعة البيانات في هذا الشكل ما هي إلا من هذا النوع.

(٧) يُصنّف (المتغير المقياسي scale variable) الفترة المرحلية، إذ لا يكون للصفر فيه أي معنى، أما متغير النسبة ratio الذي يكون الصفر فيه ذا معنى فيُعدّ مميّزاً لبعض الأغراض التحليلية، غير أن التمييز بينها ليس جوهرياً في تحليل المدونة.

يمكن استخدام البيانات في الشكل (١, ٣) لاستكشاف أنواع مختلفة من أسئلة البحث. على سبيل المثال:

- هل هناك علاقة بين جنسي المتحدثين (متغير تأويلي اسمي) واستخدام الضمائر الشخصية (متغير لغوي مقياسي)؟
- هل لكفاءة المتحدث اللغوية (متغير تأويلي ترتيبي) تأثير في استخدام ضمير المتكلم (متغير لغوي مقياسي)؟
- هل بين استخدام ضمير المتكلم واستخدام ضمير المخاطب (وكلاهما متغيران لغويان مقياسيان) علاقة؟

توزيع التكرار frequency distribution

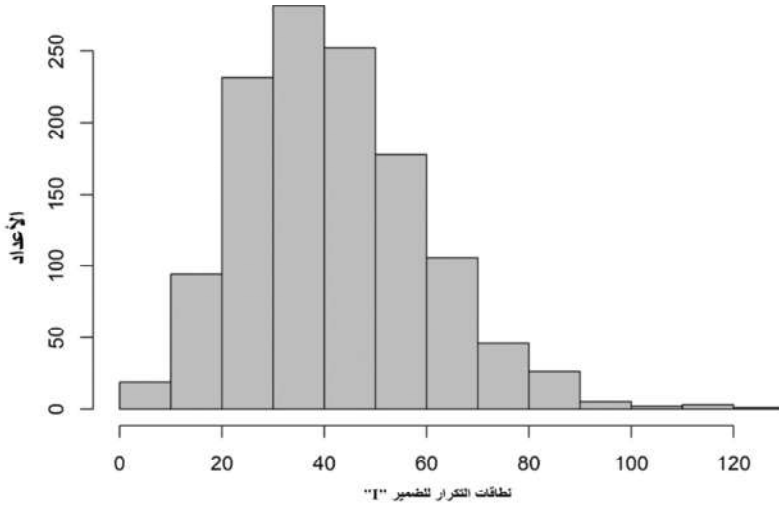
يتضمن توزيع تكرار المتغير معلومات عن القيم التي يأخذها المتغير وتكراراته. يمكن إظهار توزيعات المتغيرات المقياسية scale في مخطط دَرَجِي (المبحث ٥, ١). يعرض الشكل (١, ٤) توزيع ضمير المتكلم من بيانات الشكل (١, ٣)، ويسرد المحور السيني (X) نطاقات تكرار مختلفة لهذا المتغير اللغوي (ضمير المتكلم) لكل ألف كلمة، في حين يُظهر المحور الصادي (Y) عدد الحالات في البيانات لكل ألف كلمة. وعلى هذا النحو، يوضح الشكل البياني أن في المدونة تسعة عشر نصًّا (متحدثًا) استُخدم فيها ضمير المتكلم عشر مرات أو أقل في كل ألف كلمة (الشريط الأول من اليسار في الشكل)، و٨٨ نصًّا ظهر فيها هذا الضمير من ١١ إلى ٢٠ مرة (الشريط الثاني من اليسار)، و٢١٤ نصًّا ظهر فيها الضمير من ٢١ إلى ٣٠ مرة (الشريط الثالث من اليسار) وهكذا.

تكون التوزيعات في الغالب طبيعية -التوزيع الطبيعي^(٨) normal distribution- بوصفها المعيار الرئيس في علم الإحصاء، الذي يأخذ في شكله النهائي شكل الجرس كما هو موضح في الشكل (١, ٥). وعلى الرغم من أن كثيرًا من البيانات في العلوم الطبيعية والاجتماعية تتبع هذا النوع التوزيعي الطبيعي فإن معظم البيانات اللغوية تنحرف انحرافًا إيجابيًا (كما في الشكل أعلاه)، فيكون التوزيع في اليسار أكثر من التوزيع في اليمين كما في الشكل (١, ٤)، وتعتبر التوزيعات في علم الإحصاء

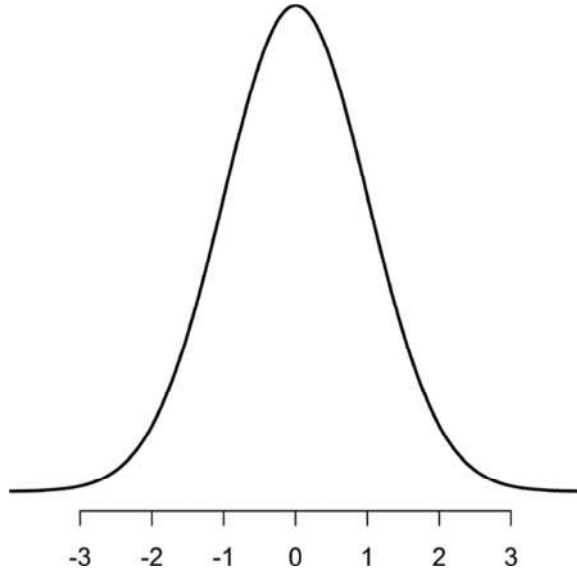
(٨) دلالة (الطبيعي) هنا مستخدمة بالمعنى التقني التخصصي، وهي تسمية أطلقها بيرسون Pearson (١٩٢٠: ٢٥)

للتوزيع الذي يُبنى عن دلالة إحصائية محددة، وعليه فإنه لا شيء غير طبيعي في كل أنواع التوزيعات التكرارية.

حاسمة؛ لأنها توفر نظرة عامة على البيانات، وتُعدّ أكثر الأساليب الإحصائية المناسبة استخدامًا. وعليه فإن لشكل التوزيع دورًا مهمًا في حسم الافتراضات الإحصائية (انظر: الافتراضات assumptions أدناه).



شكل (٤، ١). توزيع ضمير المتكلم (I) في مدونة ترينتي لانكستر.



شكل (٥، ١). التوزيع الطبيعي القياسي standard normal distribution.

القيمة المتطرفة outlier value أو القيمة الشاذة rogue value

يُمكن في كثير من الأحيان التحقق من القيم المتطرفة عندما ننظر إلى التوزيعات، وهي قيم تكون بعيدة جدًا عن القيم الأخرى. سوف يقدم المبحث (٥, ١) مخططات الصندوق boxplots، وهي وسيلة مفيدة لتحديد القيم المتطرفة. عندما نعثّر على قيمة متطرفة فإننا نحتاج إلى التحقق مما إن كانت القيمة المتطرفة هي قيمة حقيقية أو نتيجة خطأ في القياس - وهو ما يسمى بـ: القيمة الشاذة. يمكن أن تنجم القيمة الشاذة على سبيل المثال من طريق الخطأ في كتابة البيانات في جدول بيانات أو الخطأ في الرسم في المدونة. تُعدّ القيمة الشاذة نقطة بيانات مفيدة قد تتميز عن غيرها لسبب ما، ومع أن القيم المتطرفة ليست في حد ذاتها أخطاء فإنها تشكل مشكلات للنماذج الإحصائية تؤدي إلى حجب النزعة العامة في البيانات (انظر قياس النزعة المركزية أدناه)، ونبغي للباحث أن يقرر كيفية إجراء تحليل للبيانات التي تشمل قيمًا متطرفة. ويمكن استبعاد القيم المتطرفة من جزء من التحليل الذي يركز على النزعة المركزية في البيانات إذا اقتضى الأمر ذلك.

قياس النزعة المركزية measure of central tendency

المتوسط أو المعدّل average، وهو قيمة تلخيصية واحدة لسلسلة من قيم المتغير المقياسي، ويُعدّ نموذجًا إحصائيًا بسيطًا مقترنا افتراضيًا مفيدًا مع التشتت في إظهار وصف ملخص للبيانات. يمكن استخدام أنواع مختلفة من المتوسط، ومن أهمها في لغويات المدونة الحاسوبية: المتوسط mean، والوسيط median، والمتوسط المرتب العشرين بالمائة (20%). يُعرّف المتوسط (M أو $\bar{x}$) - كما هو مذكور في المبحث (٢, ١) - على أنه مجموع كل القيم مقسومًا على عدد الحالات، وهو مقياس مفيد في التوزيعات التي ليس لها قيم متطرفة (outliers) تجعل المتوسط يميل تجاهها. وفي التوزيعات ذات القيم المتطرفة قد يمثل المتوسط القيمة المتطرفة أكثر من بقية القيم مما يؤدي إلى فشل المتوسط في أن يكون نموذجًا مفيدًا. خذ على سبيل المثال تكرار الصفات في ١١ نصًا روائيًا مقتطعا من المدونة البريطانية الوطنية (BNC) والمستخدم كمثال لحساب المتوسط في المبحث (٢, ١).

508, 542, 552, 553, 565, 567, 570, 599, 656, 695, 699

وكما قلنا سابقا فمتوسط هذه القيم الإحدى عشرة هو 591.45، ومع ذلك لك أن تتصور ماذا سيحدث لو كانت القيمة الأخيرة في السلسلة هي 6,990 بدلاً من 699. في هذه الحالة، سيزحف المتوسط باتجاه القيمة القصوى، وسوف نحصل على القيمة (1,163.36)، وهي قيمة رديئة بالنسبة للبيانات؛ لأن قيمة واحدة فقط من أصل 11 قيمة تزيد على 1000. الطريقة الوحيدة لحل مشكلة حساسية المتوسط للقيم المتطرفة هي استخدام الوسيط بدلاً من المتوسط. الوسيط (median) هو القيمة المتوسطة في سلسلة من القيم المرتبة من الأصغر إلى الأكبر. بالنسبة لقيمنا الإحدى عشرة، يبلغ الوسيط 567 كما يتبين من التوضيح أدناه.

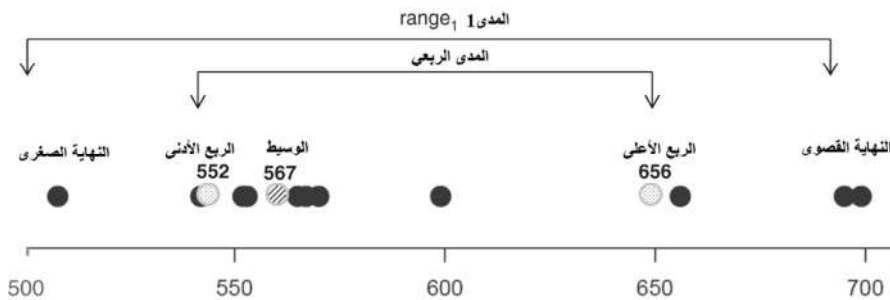
508, 542, 552, 553, 565, **567**, 570, 599, 656, 695, 699

سيبقى الوسيط دائماً في منتصف التوزيع بغض النظر عما يحدث في الأطراف، أي سواء كانت القيمة القصوى 699 أو 6,990 أو حتى 69,900.

ولو كان لدينا 10 قيم بدلاً من 11 قيمة، وهو رقم زوجي، لكان الوسيط في المنتصف بين القيمتين المركزيتين 565 و 567، كما هو موضح أدناه.

508, 542, 552, 553, **566**, **567**, 570, 599, 656, 695

والقاعدة العامة للوسيط هي أن الوسيط يكون القيمة المتوسطة في حال وجود عدد فردي من القيم؛ أما في حالة وجود عدد زوجي من القيم فتكون القيمة المتوسطة بين القيمتين المركزيتين، وذلك بجمع القيمتين المركزيتين ثم قسمة ناتج الجمع على اثنين.



شكل (٦، ١). توزيع تكرارات الصفات في أحد عشر ملفاً نصياً.

التشتت dispersion

هو انتشار قيم المتغير في مجموعة البيانات. لنعد مرة أخرى إلى عدد الصفات المحللة في المبحث (٢, ١) والمرتبة من أقل قيمة (508) إلى أعلى قيمة (699):

508, 542, 552, 553, 565, 567, 570, 599, 656, 695, 699

عندما نُمثل هذه القيم على خط يمكننا أن نرى تشتتها، وإلى أي مدى تتباعد القيم الفردية. تُسمى المسافة بين أصغر قيمة وأكبر قيمة بـ المدى $range_1$. يشير الرمز السفلي إلى وجود مفهوم آخر لـ (المدى) في لغويات المدونة الحاسوبية يسمى المدى₂، وهو موضح بالتفصيل في الفصل الثاني. يعطينا المدى₁ إشارة إلى تشتت المتغير المقياسي، ويتأثر المدى بالقيم المتطرفة، ولهذا يُفضل عادةً المدى الربعي $interquartile\ range$. ويُعدّ المدى الربعي الفاصل بين الربع الأدنى والربع الأعلى، أي: الحد الأدنى والأعلى لـ (الكتلة المتوسطة middle bulk) للقيم المرتبة من الدنيا إلى العليا، ويمثل 50% من قيم التوزيع ما عدا الوسيط وحدود القيم الربعية، وقد وُضّحت مفاهيم المدى₁ والمدى الربعي في الشكل (٦, ١). وقد قسمنا التوزيع بالكامل إلى أربعة أرباع، وحددنا القيم التي تشكل حدود الأرباع (الربع الأدنى والأعلى)؛ وعليه فالمدى الربعي يكون الفاصل الذي يحتوي على جميع القيم بين الربع الأدنى والأعلى.

ومقياس التشتت البديل للمدى₁ والمدى الربعي هو الانحراف المعياري $standard\ deviation\ (SD)$ ، وهو الجذر التربيعي لمجموع مربعات تباعدات القيم الفردية عن المتوسط. يعطينا ذلك إشارة إلى التباعد الإجمالي للقيم الفردية عن المتوسط (انظر: الفصل الثاني لمزيد من التفاصيل).

القياس الإحصائي statistical measure

هو مصطلح عام لأي عملية إحصائية، ويمكن أن يكون بسيطاً مثل المتوسط أو ينطوي على نمذجة إحصائية معقدة مثل نماذج التأثيرات الممتزجة mixed-effects models (الفصل السادس). من الأمثلة الأخرى للمقاييس الإحصائية التي نوقشت في هذا الكتاب: الانحراف المعياري SD، ومؤشر كوهن-د Cohen's d وقيمة المعلومة المتبادلة mutual information (MI)، ودلتا ب Delta P، وقيمة-ت t-score، وقيمة-ف F-score.

الاختبار الإحصائي statistical test

هو إجراء في فرع من فروع علم الإحصاء الاستدلالي inferential statistics، وهو إحصائيات تتجاوز العينة (المدونة) للاستدلال على شيء يتعلق بالمجموع اللغوي (الاستخدام اللغوي ذاته). يعتمد المنطق الأساسي له على (اختبار دلالة الفرضية الصفرية null-hypothesis significance testing NHST)، وتُعدّ قيمة الاحتمال p-value أشيع رموز الاختبار الإحصائي للفرضية الصفرية. نصادف كثيرا في تقارير الأبحاث -على سبيل المثال- رموزًا مثل $p < 0.05$ أو $p < 0.01$ ، وبناءً على قيمة الاحتمال على شاكلة هذه الرموز يُمكننا أن نقرر ما إذا كانت لدينا أدلة كافية في المدونة (العينة) لرفض الفرضية الصفرية أم لا، وذلك باتباع الخطوات الآتية:

١- نبدأ بالفرضية التي نريد اختبارها والتي تسمى الفرضية البديلة أو H_1 . على سبيل المثال، يمكن أن تفترض الفرضية اللغوية الاجتماعية H_1 أن الرجال والنساء يختلفون في استخدام الكلمات النابية.

٢- نصوغ الفرضية الصفرية (H_0) التي هي عكس H_1 . وببساطة شديدة، تنص الفرضية الصفرية null hypothesis على أنه لا يوجد فرق بين البيانات اللغوية (المدونة) التي نحللها أو بين مدونتين (فرعيتين). وعليه سنفترض بأن H_0 تنص على أنه لا فرق بين الرجال والنساء في استخدام الكلمات النابية.

٣- نختبر الفرضية الصفرية باستخدام اختبار إحصائي مثل اختبار t-test للعينات المستقلة. لكن قبل ذلك، نحتاج إلى التحقق من أن بياناتنا تنطبق عليها افتراضات الاختبار المحدد (انظر: الافتراضات assumptions أدناه).

٤- نحصل عادةً على قيمتين مهمتين من الاختبار الإحصائي؛ الأولى: إحصائية الاختبار، والثانية: قيمة الاحتمال p-value. استنادًا إلى قيمة الاحتمال (أي: القيمة الاحتمالية المشاهدة في المدونة عن طريق المصادفة وحدها) نقرر ما إذا كنا سنرفض الفرضية الصفرية أم لا. إذا كانت قيمة الاحتمال صغيرة بما يكفي (عادةً أصغر من 0.05، أي 5%)، فإننا نرفض الفرضية الصفرية، ونخلص إلى أنه من غير المرجح أن يكون الاختلاف الملحوظ في استخدام الكلمات النابية نتيجة صدفة، وبهذا تكون النتيجة ذات دلالة إحصائية (أي: إن الفرق بين الرجال والنساء في استخدام

الكلمات النابية صحيح). هذا يعني أن الاختلاف الذي لوحظ في المدونة (العينة) من المحتمل أن يكون فرقاً حقيقياً في المجموع اللغوي (أي: في الاستخدام اللغوي بأكمله). إذا كانت قيمة الاحتمال مساوية أو أكبر من 0.05 فإننا نستنتج بأنه لا يوجد دليل كافٍ في المدونة لرفض الفرضية الصفرية. نحتاج هنا إلى توخي الحذر عند تفسير نتيجة مثل هذه، وينبغي ألا نفهم على أن الفرضية البديلة (H_1) خاطئة أو أن الفرضية الصفرية (H_0) صحيحة، لأنه لا يوجد دليل كافٍ على رفض H_0 ، وقد يصبح الاختبار الإحصائي ذا دلالة إحصائية إذا جمعنا المزيد من البيانات. لاحظ أن 0.05 أو 5% هي نقطة الفصل الشائعة (أو نقطة مستوى المعنوية) التي يمكن تصوّر مخاطرها التي نود تحملها عند الاستدلال من العينة على المجموع اللغوي (انظر: قيمة الاحتمال أدناه). إذا كنا على استعداد لتحمل مخاطرة أقل من 5% فقط، فيمكننا اتخاذ قرار بشأن نقطة الفصل لقيمة الاحتمال لتكون 0.01 (1%) أو حتى 0.001 (0.01%).

غالباً ما تكون قيمة الاحتمال p -value هي العلامة الأكثر ظهوراً في الاختبار الإحصائي، وسيكون من المضلل اختزال جميع الإحصاءات إلى قيم الاحتمال. القيمة p هي قيمة الاحتمال (p) تعني الاحتمال) وهي إحدى نتائج الاختبار الإحصائي. يمكن تعريف قيمة الاحتمال على أنها احتمالية أن تكون البيانات غير دقيقة إذا كانت الفرضية الصفرية صحيحة. في حال كانت الفرضية الصفرية صحيحة، فإن المجموع اللغوي إذا لم يكن فيه اختلاف حقيقي، وكان الفرق الذي لوحظ في المدونة (العينة) عائداً فقط إلى الصدفة نتيجة لخطأ في أخذ العينات فإن بحثاً لغوياً اجتماعياً يبحث في استخدام الرجال والنساء للكلمات النابية سيفيدنا بقيمة احتمال من شأنها أن تعطينا احتمال وجود فرق ملحوظ بين المجموعتين (الرجال والنساء عند استخدام الكلمات النابية).

افتراضات assumptions

تُشير افتراضات الاختبار الإحصائي حسب المفهوم الشائع إلى الشروط التي ينبغي توافرها ليحقق الاختبار الإحصائي نفسه نتائج صالحة. من الافتراضات النموذجية لعدد من الاختبارات الإحصائية التي تُعرف بالاختبارات البارامترية parametric tests (مثل اختبار -ت t-test أو تحليل التباين ANOVA) ما يُعرف بـ: الافتراض الاعتيادي normality assumption. يرى هذا الافتراض

أن توزيع تكرار المتغير اللغوي لا ينحرف كثيراً عن التوزيع الطبيعي؛ لأن الاختبارات البارامترية مثل اختبار-ت تقارن بالعادة المتوسطات للتوزيعات المنحرفة أكثر من المعقول (قياسات النزعة المركزية). في مثل هذه الحالات يمكن استخدام الاختبارات غير البارامترية مثل اختبار مان-ويتني يو Mann-Whitney U (نسخة غير بارامترية من اختبار-ت)، وتقارن هذه الاختبارات غير البارامترية مجموع رُتب القيم بدلاً من متوسطات القيم الفعلية (المبحث ٣, ٦). تشير الأبحاث الإحصائية (على سبيل المثال Boneau 1960، و Lumley et al. 2002، و Schmider et al. 2010) إلى أن كثيراً من الاختبارات البارامترية مثل اختبار-ت أو تحليل التباين ما هي إلا إحصاءات متينة للافتراض غير الاعتدالي. هذا يعني أنها تُظهر نتائج صالحة حتى في حالة نقض الافتراض الاعتدالي وانحراف البيانات. من الأمثلة الأخرى على الافتراضات الشائعة للاختبارات الإحصائية الآتي:

- تجانس التباينات: للمدونات (الفرعية) التي نقارنها تباين متماثل، أي: قيم فردية متشعبة.
- الاستقلالية: نفترض معظم الاختبارات الإحصائية المستخدمة في هذا الكتاب (مثل اختبار-ت للعينات المستقلة أو اختبار كاي-تربيع) استقلالية الملاحظات في مدونات (فرعية) مختلفة. هذا يعني، على سبيل المثال، أن السمات اللغوية في نص ما لا يُفترض أن تتأثر بسمات لغوية في نص آخر.
- الخطية: يمكن نمذجة العلاقة بين بعض المتغيرات بخط مستقيم (الشكل ١, ١)، وفي حالات أخرى، قد يكون هناك نموذج مختلف مثل المنحنى أكثر ملاءمة يؤدي إلى نقض الافتراض الخطي. وستناقش افتراضات هذه الاختبارات بمزيد من التفصيل في الكتاب.

مجال الثقة confidence interval

في الإحصاءات الاستدلالية، مجال الثقة هو محاولة للابتعاد عن التفكير الثنائي الذي يرتبط غالباً باختبار دلالة الفرضية الصفرية NHST والاختبارات الإحصائية وقيم الاحتمال. وبدلاً من أن يُقدم مجال الثقة قراراً تقريرياً (أي بـ نعم / لا) عن الدلالة الإحصائية، يُقدم تقديراً للقيمة الحقيقية للقياس الإحصائي (مثل المتوسط) أو الفرق بين قياسين إحصائيين (مثل الفرق بين متوسطين) في المجموع اللغوي.



شكل (٧, ١). حالتان لمجال الثقة.

جدول (١, ١). حجم التأثير (r) وتفسيره القياسي.

حجم التأثير (r)	التفسير
1.0	تأثير صغير
3.0	تأثير متوسط
5.0	تأثير كبير

لا يُمثّل مجال الثقة، كما يوحي الاسم، قيمة واحدة، بل مجموعة من القيم يمكن تمثيلها على هيئة أعمدة خطأ كما في الشكل (٧, ١). في لغويات المدونة الحاسوبية، ننشئ غالبًا فئات ثقة بنسبة ٩٥ بالمئة. تُبنى هذه النسبة الفتوية من الثقة حول قياس إحصائي قائم على العينة التي نحللها، وتكمن القيمة الحقيقية لهذا التدبير في هذه الفئة بالنسبة ٩٥ بالمئة من العينات المأخوذة من المجموع اللغوي نفسه. تُستخدم مجالات الثقة غالبًا عند مقارنة مدونتين (فرعيتين) أو أكثر، وحرى بأن نميز في هذا السياق بين حالتين نموذجيتين؛ الأولى: التداخل الكبير، والثانية: فئات ثقة غير متداخلة كما في الشكل (٧, ١). إذا كانت مجالات الثقة متداخلة إلى حد كبير فمن المحتمل جدًا أن تمثل المدونات (الفرعية) المعنية المجموع اللغوي نفسه، وبناءً عليه لا يكون هناك فرق بين المجموعتين. أما مجالات الثقة غير المتداخلة فتشير إلى احتمال أن تكون المدونات (الفرعية) قد جُمعت من مصادر لغوية مختلفة (للاطلاع على مزيد من المناقشة التفصيلية انظر Cumming et al. 2007).

حجم التأثير effect size

يُعدّ حجم التأثير في علم الإحصاء الوصفي مقياسًا موحدًا، أي مقياس يمكن مقارنته عبر دراسات مختلفة (انظر: المبحث ٣, ٨ لمناقشة التحليل البعدي) ويعبر عن الأهمية العملية للتأثير الملحوظ في المدونة أو المدونات. على سبيل المثال، إذا أثبتنا من خلال اختبار إحصائي أن مجموعتين

من المتحدثين (من الرجال والنساء مثلاً) تختلفان في استخدام متغير لغوي معين، أي بين هاتين المجموعتين فروق ذات دلالة إحصائية، فسنظل بحاجة إلى معرفة حجم هذا الاختلاف ومدى أهميته من الناحية العملية، ويمكن استخدام قياسات حجم التأثير مثل r أو نسبة الأرجحية odds ratio أو مؤشر كوهن-Cohen's d للوصول إلى حكم جيد. يقدم الجدول (١، ١) تفسيرات قياسية لثلاث نقاط فاصلة لقياس حجم التأثير r (Cohen 1988). تجدر الإشارة إلى أن هذا التفسير يقدم فقط إرشادات عامة، ولا ينبغي عدّه أمراً مسلماً به (المبحث ٤، ٨).

(٤، ١) بناء المدونات وتصميم الأبحاث

فكّر في ...

قبل قراءة هذا المبحث، فكّر في الأسئلة التالية:

١- ما عدد النصوص التي نحتاج إلى تجميعها لإنشاء مجموعة بيانات لغوية؟

٢- ما معنى كون المدونة مُثَلَّة representative؟

٣- هل المدونات الكبيرة أفضل دائماً من المدونات الصغيرة؟

المدونة هي مجموعة من النصوص^(٩) (أو النصوص المنسوخة من تسجيلات منطوقة) التي يمكن تحليلها باستخدام الحاسوب. يمكن أن تكون المدونات عامة أو تكون خاصة بتركز في وعاء معين أو مؤلف معين أو مجال معين. المدونة اللغوية العامة هي عينة من اللغة مأخوذة من مجموعة كبيرة جداً من المجاميع اللغوية، وفي حالة المدونة العامة يتكون المجموع اللغوي من كامل اللغة التي ينتجها الأشخاص خلال فترة زمنية معينة. تختبر المدونة التزامنية (الآنية) اللغة حول نقطة واحدة في فترة واحدة (تُنتج مثلاً في عام واحد)، في حين أن المدونة التعاقبية (التاريخية) تتضمن لغة عبر فترات مختلفة. من المقدّر أنه في المتوسط ينطق الشخص 16000 كلمة في اليوم (Mehl et al. 2007). ونظراً لوجود نحو ٤٠٠ مليون شخص يتحدثون الإنجليزية بصفتها لغتهم الأولى ومئات الملايين غيرهم

(٩) تُستخدم كلمة (نص) في اللغويات العامة في الغالب للدلالة على النصوص المكتوبة والمنطوقة. ينبغي أن تكون جميع

النصوص منسقة عند تحليل المدونة كي يمكن قراءتها آلياً.

يتحدثون الإنجليزية بصفتها لغتهم الثانية (Crystal 2003)، يمكن تقدير إنتاج المنطوق اليومي للغة الإنجليزية وحدها بتريليونات الكلمات. وزيادة على ذلك، فإن الأشخاص أيضا يكتبون رسائل نصية ورسائل بريد إلكتروني ومنشورات مدونة وقوائم تسوق وتقارير أعمال ومقالات وقصائد وما إلى ذلك، وكل ذلك يضاف إلى الإنتاج اللغوي اليومي. وفي المقابل، فإن المدونات (حتى أكبرها حجمًا) صغيرة نسبيًا عند مقارنتها بإجمالي الإنتاج اللغوي. وعليه، إذا كانت بنية المدونات جيدة، فيمكنها أن تفيدنا بشأن المجموع اللغوي، أي اللغة نفسها. ولكن ماذا تعني (البنية الجيدة)؟

يستخدم مصطلح (مُثَلَّة) لوصف المدونة في لغويات المدونة الحاسوبية، وهو واصف للعينة عندما يكون لها خصائص مماثلة للمجموع اللغوي الذي أُخذت منه. يتيح لنا هذا استخلاص استنتاجات عن المجموع اللغوي من العينة. يمكن تحقيق عنصر (المثلية) بأخذ عينات عشوائية^(١٠) random sampling من كل نص أو كلام شرط تحقيق الاتزان. جمع العينات عشوائيًا غيرٌ منهجي؛ لأنه لا يوجد تمثيل منضبط لكل إنتاج لغوي يمكن أن نشير إليه عند أخذ العينات. علاوة على ذلك، كل المُنتجات اللغوية لا يمكن أن تدوّن. لجعل عملية أخذ العينات ممكنة، يبدأ مصمم المدونة غالبًا بمجموعة من الفئات التي يرومون بها إلى جمع عينة غير متحيزة. تسمى هذه الفئات إطار العينة sampling frame. يقدم الجدول (١، ٢) مثالًا عليه من عائلة براون، ويُعدّ هذا الإطار من الأكثر شهرةً في لغويات المدونة الحاسوبية (Francis & Kucera 1979).

جدول (١، ٢). إطار عينة عائلة براون.

فئة النص	عدد النصوص
أ صحفي: تحقيق صحفي	44
ب صحفي: افتتاحية	27
ج صحفي: تقييمات نقدية	17
د ديني	17
ه مهارات ومهن وهوايات	36

(١٠) تختلف غالبًا العينات الاحتمالية عن العينات غير الاحتمالية (المناسبة) في علم الإحصاء، وتُعدّ العينات العشوائية نوعًا من العينات الاحتمالية.

تابع جدول (٢, ١).

فئة النص	عدد النصوص
و الثقافة الشعبية	48
ز الأدب البحث، والتراجم، والمقالات	75
ح متنوعات (وثائق حكومية، تقارير مؤسسية، تقارير صناعية، جدول جامعي، مجلة دار الصناعة)	30
ط كتابات مستنيرة وعلمية	80
ي الرواية	29
ك قصص الغموض والقصص البوليسية الروائية	24
ل الخيال العلمي	6
م المغامرات والرواية الغربية	29
ن القصص الرومانسية والغرامية	29
س الفكاهة	9
المجموع	500

تتكون عائلة براون من خمس عشرة فئة أختير لها ٥٠٠ عينة نصية. تتضمن كل عينة ٢٠٠٠ كلمة. إذن: تتكون كل مدونة من مدونات عائلة براون نحو مليون كلمة من اللغة الإنجليزية المكتوبة (٢٠٠٠ × ٥٠٠). في أعراف تصميم هذه المدونة، يتمثل هدف مصممي المدونة في تحقيق عينة من النصوص غير المتحيزة في فئات إطار العينة. لاستخدام طريقة العينات العشوائية داخل الفئات أساس منهجي، وتجنب التحيز في عملية الاختيار أساس تطبيقي. نعي بالتحيز **bias** الانحراف المنهجي الخفي للعينة المأخوذة من المجموع اللغوي.

وفما يلي قائمة بأنواع التحيز الأكثر شيوعاً ومبادئ اختيار النص التي تساعد على تجنب هذا التحيز.

- **تحيز عينة النص text sample bias**: للأقسام المختلفة من النصوص (مثل البداية والوسط والنهاية) خصائص لغوية مختلفة، فعل سبيل المثال، تميل نهايات النصوص إلى تضمّن لغة تلخص النقطة الرئيسية للنص كما في الآتي: in sum, to conclude, they lived happily ever after [الترجمة: باختصار، وفي الختام، عاشوا في سعادة دائمة بعد ذلك]. إذا كانت عينات المدونة تمثل أجزاء معينة

فقط، كأن تكون ممثلة لأول أو آخر ٢٠٠٠ كلمة من كل نص، فستحوز هذه الأجزاء نسباً تمثيلية تميل إلى المبالغة. من المهم تحقيق التوازن من جهة أنواع النصوص المختلفة الممثلة في المدونات التي لا تتضمن نصوصاً كاملة، ولكن تتضمن مجرد عينات نصية كإطار عينة عائلة براون.

• **التحيز الموضوعي topic bias**: ينشأ التحيز الموضوعي عندما يُضمّن كثير من النصوص المتعلقة بالموضوع نفسه في المدونة (ما لم يُتعمد بناء المدونة بوصفها مدونة محددة حول الموضوع). تحتوي هذه النصوص عادةً على عدد من عناصر المفردات المتعلقة بالموضوع التي تتكرر عدة مرات في كل نص. يرتبط هذا بما يسمى بمشكلة حلزون البحر whelk problem (المبحث ٤, ٢) عندما تُصبح بعض العناصر المعجمية النادرة ممثلة تمثيلاً زائداً جداً في المدونة. يمثل التحيز الموضوعي مشكلة، خاصة في المدونات الصغيرة، عندما يشكل كل نص فردي نسبة كبيرة من المدونة. في معظم الحالات، ينبغي لمصممي المدونة اختيار النصوص بتنّبه بين مجموعة من الموضوعات.

• **التحيز بدون تغطية non-coverage bias**: بعض النصوص أكثر ظهوراً من غيرها؛ لأن مصممي المدونة يعدّونها نماذج أولية لأسباب مختلفة. على سبيل المثال، قد تُفضّل النصوص المنشورة على الخطابات أو الرسائل الإلكترونية الخاصة. على مصممي المدونة أن يسعوا سعياً حثيثاً لتناول أكبر عدد ممكن من النصوص.

• **تحيز نوع النص التقليدي traditional text type bias**: هذا النوع من التحيز هو حالة خاصة من التحيز بدون تغطية. عند اختيار إطار العينة، يكون لدينا في الغالب استعداد لأن نرى أنواعاً من النصوص التي أدرجت تقليدياً في المدونات اللغوية بارزة وواضحة مثل تلك الموجودة في إطار عينة عائلة براون (انظر: الجدول ٢, ١). ومع التقنيات وأساليب الاتصال الجديدة تظهر أنواع جديدة مثل منشورات المدونات والتغريدات. لذلك يحتاج مصممو المدونة إلى التفكير ملياً في أنواع النصوص التي ينبغي أخذ عينات منها، وفي الآثار المترتبة على ذلك بالنسبة لتمثيل المدونة.

• **التحيز لاعتبارات قانونية legal considerations bias**: تُواجه مشكلة حقوق النشر من يصمّمون المدونات، وتظهر المشكلة عندما يرغب مصممو المدونة في مشاركة مدوناتهم مع باحثين آخرين. قد تؤدي الاعتبارات القانونية إلى اختيار مجموعة من النصوص التي لا تسري عليها حقوق الطبع والنشر (النصوص القديمة خارج نطاق حقوق النشر والنصوص الصادرة بموجب تراخيص المشاع

الإبداعي Creative Common License وما إلى ذلك) مما يخلق مشكلة التحيز في أخذ العينة. في الوقت الحالي، ليس لهذه المشكلة حل واضح، فالنُّهج قد تختلف وفقاً للمتطلبات القانونية في كل دولة على حدة، إلا أن على مصممي المدونة أخذ المسائل القانونية في الاعتبار.

- **التحيز التطبيقي العملي practicality bias:** يسهل الحصول على بعض النصوص أكثر من غيرها مثل صفحات الويب. قد يؤدي هذا الاعتبار التطبيقي العملي إلى إنشاء مدونة تبالغ في تمثيل النصوص التي يسهل الحصول عليها. يحتاج مصممو المدونة في هذه الحالة إلى مقاومة هذا الإغراء والسعي إلى الحصول على مدونة من النصوص بغض النظر عما إذا كان يسهل الحصول عليها أم لا.
- **انحياز الاختيار الذاتي self-selection bias:** ينشأ هذا الانحياز عندما يُطلب من مؤلفي النصوص تقديم نصوصهم تطوعاً. فعلى سبيل المثال، إذا كنا نرغب في إنشاء مدونة من الكتابات الصفية ونطلب من الطلاب أن يتطوعوا ويسهموا بنصوصهم، فقد ينتهي بنا المطاف إلى نصوص من طلاب محفّزين لا تعكس حقيقة الإنتاج اللغوي الذي يعكس مستوى الطلاب الصفّي. لذا فمصمّمو المدونة بحاجة إلى الوصول إلى مجموعات مختلفة من الإنتاج اللغوي وتشجيع الطلاب جميعهم على الكتابة العفوية من أجل الحصول على عينة مُثّلة.

كل ما ذُكر عن المدونة يعد من الطرائق التقليدية. في المقابل، برزت طريقة مختلفة لتمثيل المدونة وأخذ العينات من الويب كمدونة **web as corpus** (Baroni & Ueyama 2006; Baroni et al. 2009; Jakubicek et al. 2013). ويستند المبدأ الأساس لهذه الطريقة إلى ملاحظة بسيطة في ظل ارتفاع شعبية الإنترنت وتزايدها، وهي أن كميات كبيرة من المواد اللغوية المثيرة للاهتمام قد أصبحت متاحة على الإنترنت. توسعت البيئة اللغوية عبر الإنترنت لتشمل مجموعة واسعة من سمات حياتنا اليومية من العمل والترفيه والتفاعل الاجتماعي وغير ذلك، ويمكننا بناء مدونات أكبر من أي وقت مضى بكشط الويب وجمع البيانات منه جميعاً منظماً وآلياً، وأسهل من الطريقة القائمة على جمع عينات نصية من غيره ثم تصنيفها حسب مجالاتها وأوعيتها (إطار العينة). ويُزعم أن للمدونات الكبيرة على الإنترنت القدرة على أن تكون أكثر تمثيلاً للاستخدام اللغوي الحالي، ولا سيما أن بيئة الإنترنت أصبحت أكثر شمولاً وتزايداً، وهو ما يجعلها تعكس الإنتاج اللغوي حتى في وضعية عدم الاتصال بالإنترنت كحال الكتب المطبوعة التي تتوافر في صورة كتب إلكترونية. وعلى الرغم من حجم

الإنترنت الكبير فإنه لا يمثل سوى نسبة معينة من إجمالي الإنتاج اللغوي ما عدا الحوارات غير الرسمية على سبيل المثال، التي تشكل جزءًا كبيرًا من الإنتاج اللغوي اليومي، وبناء على ذلك نحتاج إلى التفكير مليًا في العلاقة بين المدونة بوصفها عينة وبين الإنتاج اللغوي (المجموع اللغوي) الذي تمثله هذه العينة عند تحليل مدونة قائمة على الويب.

تكون المدونة عينة لغوية كما قلنا سابقا، وفي حالات معينة، يمكن أن تشمل المدونة كامل المجموع اللغوي. على سبيل المثال: في دراسات الأدب، تشمل المدونات التي تضم جميع أعمال مؤلف معين كامل المجموع اللغوي. إذا كنا نرغب في مقارنة كلام الأمير هاملت بكلام صديقه هوراشيو في رواية شكسبير التراجيدية الشهيرة، فسنعمل بكل الأدلة الموجودة حول السلوك اللغوي لهذين الشخصين. إذا جمعنا جميع المقالات الصحفية - على سبيل المثال أيضا - عن موضوع معين في فترة معينة، فسنبحث في كامل المجموع اللغوي للمقالات التي كُتبت عن هذا الموضوع. اقترح بيكر وآخرون (Baker et al. 2013) بمدونتهم البالغة ١٤٣ مليون كلمة من مقالات الصحف البريطانية التي تحتوي على كلمات تتعلق بالمسلمين والإسلام اقتراحًا كبيرًا من هذا النموذج المثالي. في هذه الحالات وما شابهها، فإن الطريقة الإحصائية المناسبة لتحليل المدونات القائمة على المجموع اللغوي *population-based corpora* هي الطريقة الوصفية (أي علم الإحصاء الوصفي). يمكن تخطيط عملية الاستدلال الإحصائي؛ لأننا سنكون قادرين على مراقبة الحالة الحقيقية للأمر.

بعد مناقشة جوانب مختلفة من بناء المدونة، يتبقى سؤال أساسي واحد ينبغي الإجابة عنه: ما الحجم الذي يتوجب أن تكون عليه المدونة؟ ما من إجابة شاملة عن هذا السؤال؛ لأن حجم المدونة يعتمد على سؤال البحث ونوع السمات اللغوية التي نريد تحليلها. بالنسبة للتحقيق في الأبنية النحوية مثل أبنية المبني للمجهول *passive*، وهي شائعة إلى حد ما، يمكن أن تكون حتى المدونات الصغيرة (على سبيل المثال البالغة مليون كلمة أو أقل) كافية. وفي المقابل، نجد العديد من المفردات المعجمية ومجموعاتها نادرة إلى حد ما حتى في المدونات الضخمة، ويسهل أن نواجه مشكلة تبعر البيانات *data sparsity*. لتوضيح هذه المشكلة، يوضح الجدول (٣، ١) خمسة تعبيرات وتكراراتها في مدونة من مليون و ١٠٠ مليون و ٢٠ مليار كلمة على التوالي. على الرغم من أن جميع هذه التعبيرات واضحة وصريحة تمامًا من جهة معناها فإنها لم تظهر في

مجموعة المليون كلمة على الإطلاق. ومن المثير للاهتمام، أن الظرف (غير صالح للسكن uninhabitably؛ كما في المثال الآتي: 44 fumes from this single hum will eventually render a km area of the lunar surface uninhabitably radioactive هذا الحريق وحده ستجعل في النهاية ٤٤ كم من سطح القمر مشعة وغير صالحة للسكن) قد تكرر أربع مرات فقط في ٢٠ مليار كلمة.

جدول (٣، ١). تكرار الكلمات والتعبيرات المختارة في ثلاث مدونات باللغة الإنجليزية.

التعبير	BE06 (1 مليون كلمة)	المدونة الوطنية البريطانية (100 مليون كلمة)	EnTenTen3 (20 مليار كلمة)
co-pilot ^a مساعد طيار (اسم)	0	69	7,887
rater ^a مقدر (اسم)	0	11	7,832
beautiful memories ذكريات جميلة	0	1	2,552
somewhat humorously مداعبٌ إلى حد ما	0	0	90
Uninhabitably غير صالح للسكن	0	0	4

^a تشمل تكراراتها صيغ الجمع

يمكننا استنباط قاعدة عامة من هذا المثال، وهي: ما لم تمثل المدونة كامل المجموع اللغوي فإن انعدام الدليل ليس دليلاً على الانعدام The absence of evidence is not the evidence of the absence. بعبارة أخرى، إذا لم يظهر تعبير في مدونة ما فإن هذا لا يعني أن هذا التعبير غير موجود في المجموع الكلي للغة. بصفتنا مستخدمين للمدونة، نحتاج إلى التفكير ملياً في طبيعة الأدلة التي توفرها المدونات من جهة جودتها (التمثيل والتوازن) وكذلك كميتها (حجم المدونة). ينبغي أن نذكر نبذة موجزة عن دور علم الإحصاء في تحليل المدونات، ولنبدأ بالنظر في بعض المبادئ العامة، تليها مناقشة تصميمات بحثية محددة. في عملية تحليل المدونة أربعة أبعاد منفصلة ولكن مترابطة (الجدول ٤، ١)، ويحتاج كلٌ منها إلى تفسير صحيح من أجل تحقيق نتيجة ذات دلالة. نبدأ عادةً في تحليلنا باستكشاف البيانات التي نلقي من خلالها نظرة على تكرارات المتغيرات اللغوية وتوزيعاتها (المبحث ٣، ١ عن المصطلحات الإحصائية)، ثم نقدم غالباً رسومات بيانية

تعرض الأنماط الرئيسية في تلك البيانات (انظر: المبحث ٥, ١ عن استكشاف البيانات وتصويرها). إذا كانت مدونتنا تمثل عينة ولا تمثل مجموعاً لغوياً وهو الحال عادةً، ينبغي أن نفكر في مقدار الأدلة التي لدينا من العينة. وبعبارة أخرى: يمكننا استخدام إحصاءات استدلالية للاستعلام عما إذا كان من الممكن تعميم التأثيرات الملحوظة والاختلافات بين المدونات (الفرعية) على المجموع اللغوي أو كل اللغة التي تمثلها المدونة أو المدونات. تُنتج الإحصاءات الاستدلالية قيم احتمال أو فئات ثقة، ونستخدم عبارات مثل (ذات دلالة إحصائية) أو (فئات ثقة ٩٥٪ غير متداخلة) لوصف تلك الإحصاءات الاستدلالية.

جدول (٤, ١). مستويات مختلفة من التحليل في لغويات المدونة الحاسوبية.

البُعد	السؤال الأساسي	المصطلحات الرئيسية
١- استكشاف البيانات	ما الاتجاهات الرئيسية في البيانات؟	الأشكال البيانية/ المتوسطات/ الانحرافات المعيارية
٢- الإحصاءات الاستدلالية: كمية الأدلة	هل لدينا أدلة كافية لرفض الفرضية الصفرية؟ هل التأثير الذي نراه في العينة ناتج عن الصدفة (خطأ في أخذ العينة) أم أنه يعكس شيئاً حقيقياً عن المجموع اللغوي؟ 	دال إحصائي/ قيم الاحتمال/ مجالات الثقة
٣-حجم التأثير	ما حجم التأثير في العينة؟ (قياس موحد)	حجم التأثير/ مؤشر كوهن-د/ قياس r
٤-التفسير اللغوي	هل التأثير ذو مغزى لغوي/ اجتماعي؟	

يدور نقاش في عدد من التخصصات مثل علم النفس وعلم الاجتماع واللغويات التطبيقية عن المكان الذي ينبغي أن تتبوأه الإحصاءات الاستنتاجية وقيم الاحتمال -على وجه الخصوص- في عملية

البحث.^(١١) ولسوء الحظ، لا يزال عددٌ من الباحثين عالقين في التفكير الثنائي حول نتائجهم من جهة كونها ذات دلالة إحصائية أو غير ذات دلالة إحصائية، وغالبًا ما يحدث ذلك خلطًا بين الدلالة الإحصائية وبين الأهمية التطبيقية العملية أو المغزى اللغوي والاجتماعي *linguistic and social meaningfulness*. ولزيد من المعلومات عن هذه النقطة انظر (Cumming 2012) الذي يقدم حججًا مقنعة لمصلحة ما يسميه (الإحصائيات الجديدة)، وهي إجراءات إحصائية تستند إلى أحجام التأثير وتقدير مجالات الثقة. وفي طيات هذه الدراسة، نكتفي بتأكيد أن (الدلالة الإحصائية) و(الأهمية التطبيقية العملية) و(المغزى اللغوي والاجتماعي) هي ثلاثة مفاهيم مختلفة. تجربنا الدلالة الإحصائية *statistical significance* ما إذا كان لدينا أدلة كافية في المدونة لرفض الفرضية الصفرية، وفي مدونة كبيرة جدًا يتبين عادة أن اختلافًا صغيرًا (لا يكاد يكون ملحوظًا) بين مجموعتين من النصوص الفردية أو المتحدثين يكون ذا دلالة إحصائية. وتستخدم الأهمية التطبيقية *practical importance* مقياس إحصائية موحدة للتعبير عن حجم التأثير مع ضرورة محاولة تقييم حجم التأثير (على سبيل المثال: ما حجم الفرق حقًا بين المجموعتين؟) أخيرًا، نحتاج إلى ربط التأثير الملحوظ مرة أخرى بما نعرفه عن اللغة والمجتمع، ونفسر النتائج في سياق النظرية اللغوية والاجتماعية، وتسعى هذه الخطوة الحاسمة إلى اكتشاف المغزى اللغوي والاجتماعي لما لاحظناه في المدونات.

نحتاج عند تحليل المدونات أيضًا إلى التفكير في نوع البيانات وشكلها الذي ينبغي الحصول عليه من المدونة للإجابة عن سؤال البحث الذي يعرف بـ: تصميم البحث *research design*. ترجع أهمية تصميم البحث إلى أن له آثارًا كبيرة في الإجراءات الإحصائية المحددة التي يمكننا استخدامها مع البيانات. يمكن التمييز بين ثلاثة أنواع رئيسة من أنواع تصميم البحث. الأول: تصميم مدونة بالكامل، والثاني: تصميم نص مكتوب أو منطوق فردي، والثالث: تصميم سمة لغوية. يوضح الشكل (٨، ١) بنية مجموعات البيانات بناءً على هذه التصميمات البحثية الثلاثة.

(١١) وصلت مجلة علم النفس الاجتماعي الأساسي والتطبيقي *Basic and Applied Social Psychology* إلى حدٍ حظر جميع الإحصاءات الاستدلالية في الأبحاث المقدمة إليها، ويزعم محررو هذه المجلة أن الإجراءات المتضمنة في الإحصاءات الاستدلالية كاختبار دلالة الفرضية الصفرية غير صالح (Trafimow & Marks 2015: 1).

(a)

الحالة (المدونة)	البنية للمجهول (تكرار مطلق)	البنية للمجهول (تكرار نسبي)
المدونة الوطنية البريطانية	1121436	11406.74

(b)

الحالة (المدونة)	البنية للمجهول (تكرار مطلق)	البنية للمجهول (تكرار نسبي)
A00	50	72.5
A01	81	99.8
A02	24	70.0
A03	369	184.8
A04	464	117.2
A05	580	137.1
A06	280	76.5
A07	424	106.2
A08	205	51.1

(c)

الحالة (السمّة)	البنية للمجهول (تكرار مطلق)	البنية للمجهول (تكرار نسبي)	نوع	مثال		
1	0	1	0	ong Hedging plants	are usually cut	back to half
2	0	1	1	regions, but it has	been deployed	under sector
3	0	1	1	BBC's recordings	aren't meant	for release c
4	1	1	1	One-way system. It	was caused	by the IRA, v
5	0	1	2	lopment projects	are scheduled	for the forth
6	0	1	3	ty grew and laws	were passed	for her prote
7	0	1	0	beral policies will	be implemented	in Peru at le
8	1	1	0	Romans, the Celts	were dismissed	by contemp
9	0	1	4	solar calendar by	being placed	at the winte
10	0	1	5	Final Invoice will	be issued	as appropria
11	0	1	6	he tissue samples	are taken	from the foe

شكل (٨, ١). التصميمات البحثية في لغويات المدونة الحاسوبية.

في تصميم المدونة بالكامل، تكون وحدة التحليل في العادة هي المدونة بالكامل، وأحياناً تكون المدونات الفرعية الكبيرة أيضاً. المدونة المعنية في الشكل (٨, ١a) هي المدونة الوطنية البريطانية، عندما نتبع تكرار تركيب المبنى للمجهول (على سبيل المثال: he was seen (رئي) بوصفه المتغير اللغوي الرئيس لدينا. تعني AF التكرار المطلق، و RF التكرار النسبي في كل مليون كلمة (انظر الفصل الثاني للاطلاع على مناقشة أوسع). يُعدّ تصميم المدونة بالكامل من أهم أنواع التصميم الأساسية التي توفر معلومات عامة عن تكرارات السمات اللغوية في المدونات دون الإشارة إلى التنوع الداخلي داخل المدونة، والهدف من ذلك هو الحصول على مثال عملي جيد في أثناء تصميم المدونة بالكامل (المبحث ٢, ٤). يتيح لنا تصميم النص المكتوب أو المنطوق الواحد إمكان تتبّع تكرار السمّة اللغوية في نص فردي أو أداء كلامي واحد. في الشكل (٨, ١b) يمكننا أن نرى تكرارات المبنى للمجهول في ملفات المدونة الوطنية البريطانية الفردية عندما يعرض الشكل الملفات التسعة الأولى فقط بالترتيب الأبجدي (انظر المبحث ٣, ٦ للاطلاع على مثال عملي لتصميم النص المكتوب أو المنطوق). يركز تصميم السمّة

اللغوية في السمة اللغوية بوصفها ملاحظة واحدة (أي: حالة واحدة). في الشكل (٨، ١٠) يمثل كل سطر حدثاً واحداً لتركيب المبني للمجهول في المدونة الوطنية البريطانية، والذي صُنّف وفقاً لخصائص البناء المختلفة وسياق الحدوث. على سبيل المثال، يمكن أن يكون تركيب المبني للمجهول قصيراً (مثل: تُقطع النباتات في العادة (plants are usually cut) ويُرمز له بالرمز 0، أو طويلاً مع مركب by (مثل It was caused by the IRA سُنت على أيدي الجيش الجمهوري الإيرلندي) ويُرمز له بالرمز 1. يمكن أن يظهر في الكلام (0) أو في الكتابة (1) وفي أنواع مختلفة مرمزة بـ: 0-6. يسمح لنا تصميم السمة اللغوية بالتحقيق في العوامل المختلفة التي لها دور في استخدام نوع معين من السمات اللغوية. على سبيل المثال: من مجموعة البيانات في الشكل (٨، ١٠ج) يُمكن البحث عما إذا كان أي نوع معين يفضل تركيبات المبني للمجهول الطويلة على القصيرة، والاهتمام بهيكل البيانات وتصميم البحث أمرٌ ضروري لنجاح التحليل الإحصائي (انظر المبحثين ٣، ٤ و ٤، ٤ للاطلاع على مثال عملي لتصميم السمة اللغوية). لاحظ أن أسئلة البحث المختلفة تتطلب تصميمات بحثية مختلفة تُنتج هياكل مختلفة من البيانات. سيناقش هذا الكتاب التصميم البحثي المناسب لأنواع مختلفة من التحليلات اللغوية، وسيوضح استخدام التقنيات الإحصائية مع دراسة حالات متعددة (تتوفر مجموعات البيانات الكاملة المستخدمة في هذا الكتاب في موقع الويب المرافق للكتاب).

وباختصار، فإن اختيار المدونة والإجراء التحليلي وتصميم المدونة تصميمًا صحيحًا كلها تُعدُّ الخطوة الأولى في التحليل الإحصائي الناجح للمدونة، وعلينا نحن الباحثين أن نفكر ملياً فيما تمثله المدونات التي نتعامل معها، وفي كيفية إمكان الكشف عن نتائج مثيرة للاهتمام عن اللغة والمجتمع.

(٥، ١) استكشاف البيانات وتمثيلها

فكّر في ...

قبل قراءة هذا المبحث، فكّر في الأسئلة الآتية:

- ١- لماذا يُعدّ تمحيص البيانات بأسلوب نقدي قبل التحليل أمراً مهماً؟
- ٢- ما أنواع الأخطاء التي يمكن أن نواجهها في مجموعة البيانات؟
- ٣- ما أنواع الأشكال البيانية التي تعرفها؟

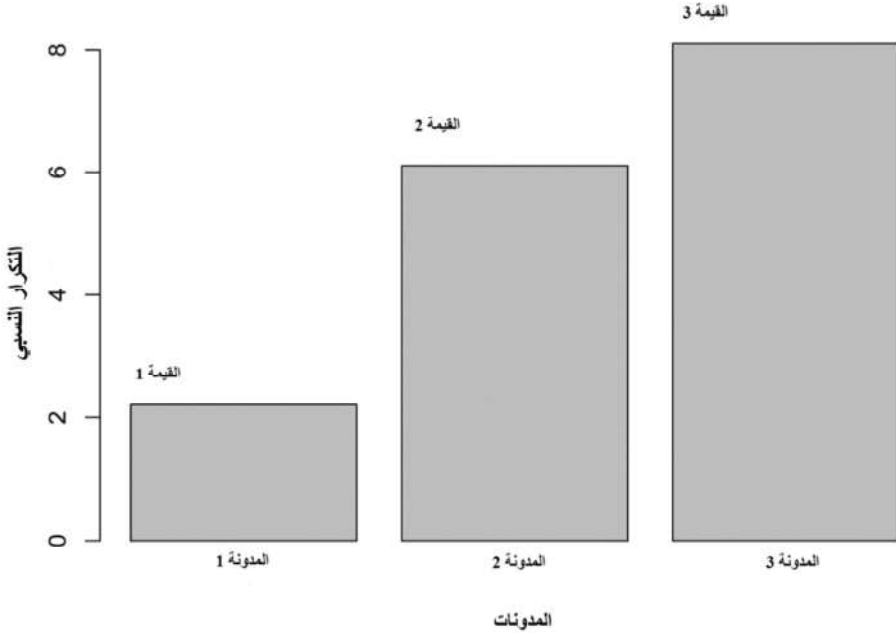
تؤدي البيانات السيئة إلى نتائج سيئة. بغض النظر عن مدى تطور تحليلنا الإحصائي، إذا أخطأنا في نقل الأرقام أو ارتكبنا خطأ عند نسخ البيانات من جدول البيانات إلى البرنامج الإحصائي فسوف ينتهي بنا المطاف إلى نتائج خطأ. من طرق تلافي هذه الأخطاء العرضية الاحتفاظٌ بدفتر بحثي يُسجّل فيه كل خطوة من خطوات التحليل لكي يمكن التحقق من الإجراء أو تكراره بسهولة. هناك جانب آخر ربما أكثر أهمية للتحليل الجيد للبيانات، وهو التساؤل المستمر عن صحة البيانات، وطرح أسئلة عدة: هل هذا هو الحجم المتوقع للمدونة أم أنني عددتُ أيضًا وسوم أجزاء الكلام بالخطأ؟ هل هذا التأثير منطقي بالنظر إلى ما أعرفه عن توزيعات الكلمات في النصوص؟ ما سبب نقطة البيانات غير العادية؟ يمكن أن يساعد التساؤل المستمر عن البيانات على تجنب ارتكاب أخطاء تافهة وإساءة تفسير النتائج.

ويعد تمثيل البيانات **data visualization** أمرًا بالغ الأهمية لفهم الاتجاهات الرئيسية في البيانات. يلخص تمثيل البيانات الفعال الأنماط في البيانات دون إخفاء السمات المهمة. يوضح المثال التالي مدى فعالية تمثيل البيانات في الشكل (٩، ١) والشكل (١٠، ١) اللذين يتضمنان مجموعة البيانات نفسها. تقارن مجموعة البيانات هذه بين ثلاث مدونات لغوية من جهة تكرارات المتغير اللغوي، إذ إنه مثال مضروب لأغراض توضيحية ولنصطلح على تسميتها (س). توضح خطاطة الأعمدة **bar chart** في الشكل (٩، ١) تكرارات (س) في كل مدونة. يبدو أن المدونات يختلف بعضها عن بعض بوضوح، إذ تغطي المدونة 3 بأكثر تكرار للمتغير (س).

والقول مختلف تمامًا في الشكل (١٠، ١) استنادًا إلى مجموعة البيانات نفسها، إذ يبدو أن المدونة ١ مختلفة بالفعل، لكن المدونة 2 و3 متشابهتان إلى حد كبير، على الرغم من أن المدونة 3 أكثر تشبّهًا من 2. تبرز على وجه الخصوص قيمة واحدة في المدونة 3 وهو ما نسميه القيمة المتطرفة **outlier**. في الشكل (١٠، ١) تُعرض البيانات في شكل **مخطط رسم مربع boxplot** مع المتوسطات المضافة ونقاط البيانات الفردية. يعرض مخططُ رسم المربع توزيعَ البيانات في كل مجموعة، ويعرض أيضًا القيم النهائية (الدنيا والقصى) والقيم المتطرفة عندما يقتضي التحليل ذلك. يمثل الجزء الداخلي من المربع المدى الربيعي (٥٠٪ من القيم)، ويمثل الخط الأفقي السميكة في المربع الوسيط (القيمة المتوسطة). تظهر الخطوط الدقيقة أعلى المربع وأسفله الحد الأدنى والحد الأقصى للقيم،

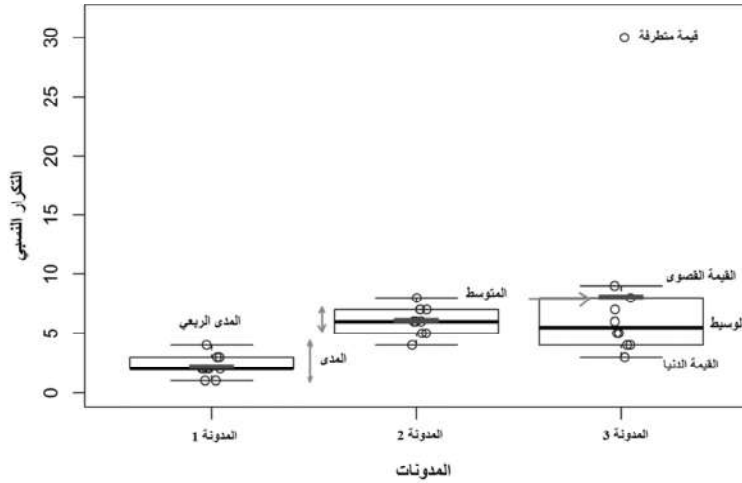
ولكنها تستبعد القيم المتطرفة التي تُعرض خارج نطاق الخطوط الدقيقة في صورة نقاط بيانات منفصلة. تُعرض أيضًا القيم الفردية (تكرارات س في نصوص مختلفة) في مخطط رسم المربع هذا مع الوسط (الخط الأفقي القصير).

يوفر مخطط رسم المربع باختصار معلومات أكثر عن البيانات مما توفره خطاطة الأعمدة البسيطة، وفي حين أن خطاطة الأعمدة في الشكل (٩، ١) لا تعرض سوى ثلاث قيم 2.2 و 6.1 و 8.1 يستخدم مخطط رسم المربع في الشكل (١٠، ١) المساحة نفسها لعرض ٤٦ معلومة مختلفة، بما في ذلك إحصائيات موجزة مهمة مثل المتوسط، والوسيط، والمدى، والمدى الربيعي^(١١)، ويساعدنا هذا على تفسير البيانات تفسيرًا مباشرًا وواضحًا، وعلى تجنب الاستنتاجات التي لا أساس لها.



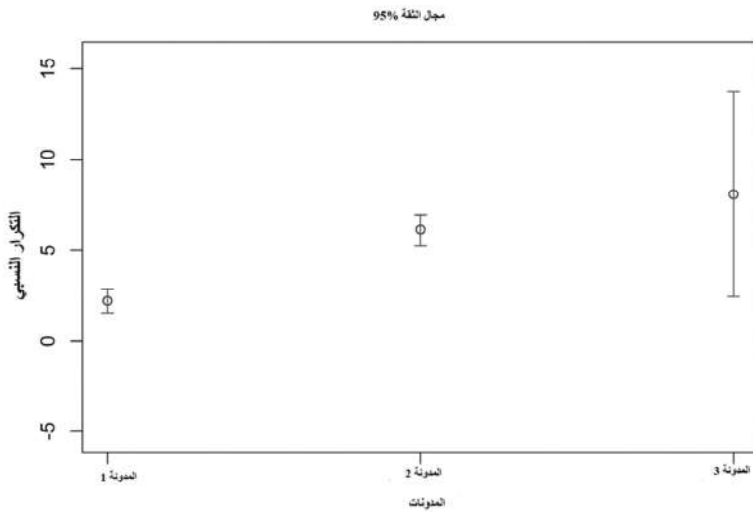
شكل (٩، ١). خطاطة أعمدة: المتغير س في ثلاث مدونات.

(١٢) عند تقديم البيانات في صورة مرئية ينبغي أن نهدف إلى عرض أشكال مفيدة (غنية بالمعلومات) ذات نسبة معلومات عالية. انظر: تافت (٢٠٠١، ٢٠٠٦) وهudson (٢٠١٥).



شكل (١٠, ١). مخطط رسم مربع: المتغير س في ثلاث مدونات.

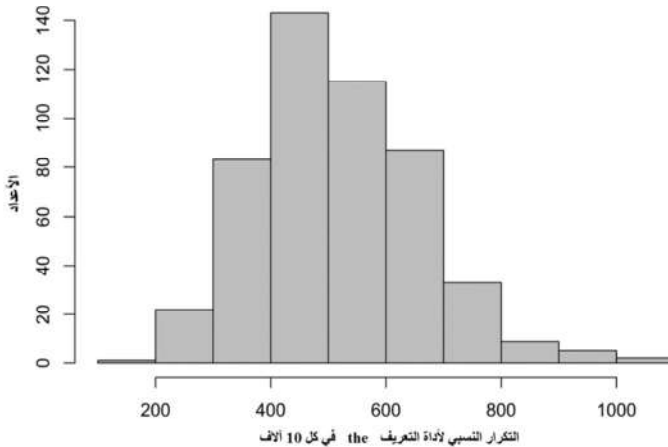
إذا أردنا تجاوز العينة وتعميمها على المجموع اللغوي فإنه يمكننا حساب مجالات الثقة ٩٥٪. للقيمة المتوسطة للمتغير س في المدونات الثلاثة، وتُعرض مجالات الثقة ٩٥٪ كأعمدة خطأ في الشكل (١١, ١).



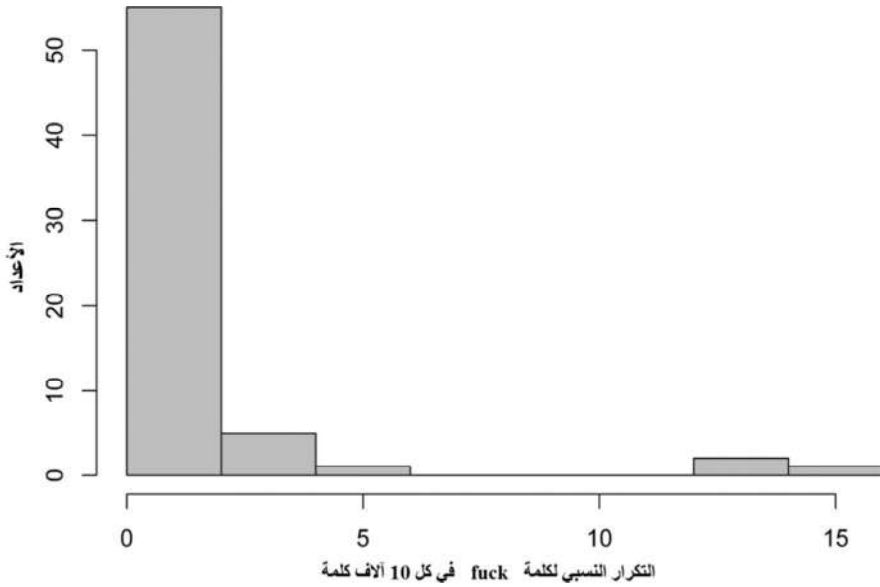
شكل (١١, ١). مخطط رسم مربع للمتغير س في ثلاث مدونات.

من أنواع تمثيل البيانات الأخرى المفيدة: المخطط الدرجي **histogram** (انظر إلى الشكلين ١, ١٢ و ١, ١٣)، والمخطط الدرجي هو رسم بياني يوضح توزيع تكرار المتغير اللغوي في شكل أعمدة، يُمثّل كلٌّ منها تكرار قيم المتغير اللغوي في فئة أو خانة (على سبيل المثال 20-30 و 31-40 وما إلى ذلك في الشكل ١, ١٢). نحن هنا مهتمون في المخطط الدرجي بشكل التوزيع الملحوظ للبيانات. توضح الأمثلة في الشكلين (١, ١٢ و ١, ١٣) التوزيع النسبي من كل 1000 كلمة لتوزيع المقال المحدد المكتوب، والتوزيع النسبي من كل 1000 كلمة لتوزيع كلمة معينة مع تصرفاتها في الكلام على التوالي. يمكننا في هذا السياق أن نرى توزيعين منحرفين إيجابياً، ويعد شكلاً نموذجياً للبيانات اللغوية. في حالة كلمة fuck وتصرفاتها يكون التوزيع منحرفاً للغاية لأنه لا يكاد أحد في المدونة يستخدم هذه الكلمة على الإطلاق.

نتعامل حتى الآن مع متغير لغوي واحد، وإذا كنا نريد مقارنة العلاقة بين متغيرات لغوية متعددة فإنه يمكننا استخدام مخطط التشتت scatterplot أو سلسلة من مخططات التشتت. وفيما يأتي مخططات تشتت استناداً إلى استخدام أداة التعريف، وضمير المتكلم وضمير المخاطب في مدونة BNC64، وهي مدونة فرعية تتضمن مليوناً ونصف المليون كلمة من المدونة الوطنية البريطانية المنطوقة. بالنظر إلى الشكل (١, ١٤) يمكننا أن نرى استخدام أداة التعريف *the* المُمثلة بيانياً مقابل ضمير المتكلم *I*. تمثل كل دائرة نقطة بيانات فردية، وفي هذه الحالة تمثل متحدثاً واحداً. من الشكل البياني، يمكننا أن نرى بوضوح كيفية استخدام المتحدثين الفرديين هذين المتغيرين اللغويين وطبيعة العلاقة بينهما.



شكل (١, ١٢). مخطط درجي: أداة التعريف في مدونة الإنجليزية البريطانية الفرعية BE06.



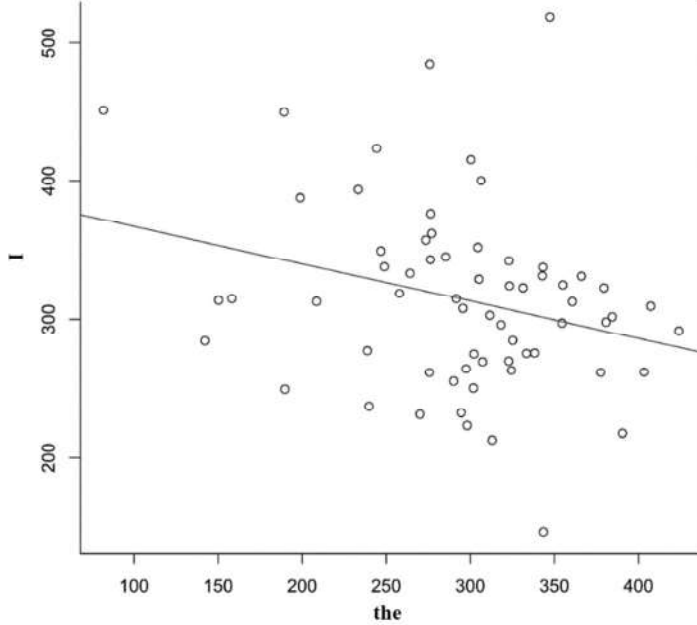
شكل (١٣، ١). مخطط درجي: كلمة fuck وتصريفاتها في المدونة الوطنية البريطانية الفرعية BNC64.

يُظهر خط الانحدار الذي يمر عبر منتصف (سحابة المتكلم speaker cloud) في مجموعة البيانات هذه أن استخدام أداة التعريف يتناسب تناسباً عكسياً مع استخدام ضمير المتكلم. عُرض كثيرٌ من مخططات التشتت في صورة مصفوفة مخطط تشتت في الشكل (١٥، ١)، وكل مخطط منها يقدم مزيجاً من اثنين من المتغيرات اللغوية الثلاثة.

المربع الأوسط في العمود الأول هو مخطط تشتت لأداة التعريف the وضمير المتكلم I، إذ يقع I على المحور السيني وthe على المحور الصادي. أفضل طريقة لقراءة مصفوفة مخطط التشتت تكون بإشارة العنوان (الكلمة) في العمود إلى ما يوجد على المحور السيني، في حين يشير العنوان (الكلمة) في الصف إلى ما يوجد على المحور الصادي.

هناك أيضاً أنواع أكثر تعقيداً من التمثيل يمكن استخدامها عند تمثيل البيانات اللغوية، ومنها: الخرائط الجغرافية geomapping على سبيل المثال التي تُعدّ طريقة فعالة لعرض إشارات إلى أماكن مختلفة في مدونة وتكراراتها على الخريطة. تعرض الخريطة أدناه الأماكن العشرة الأولى من

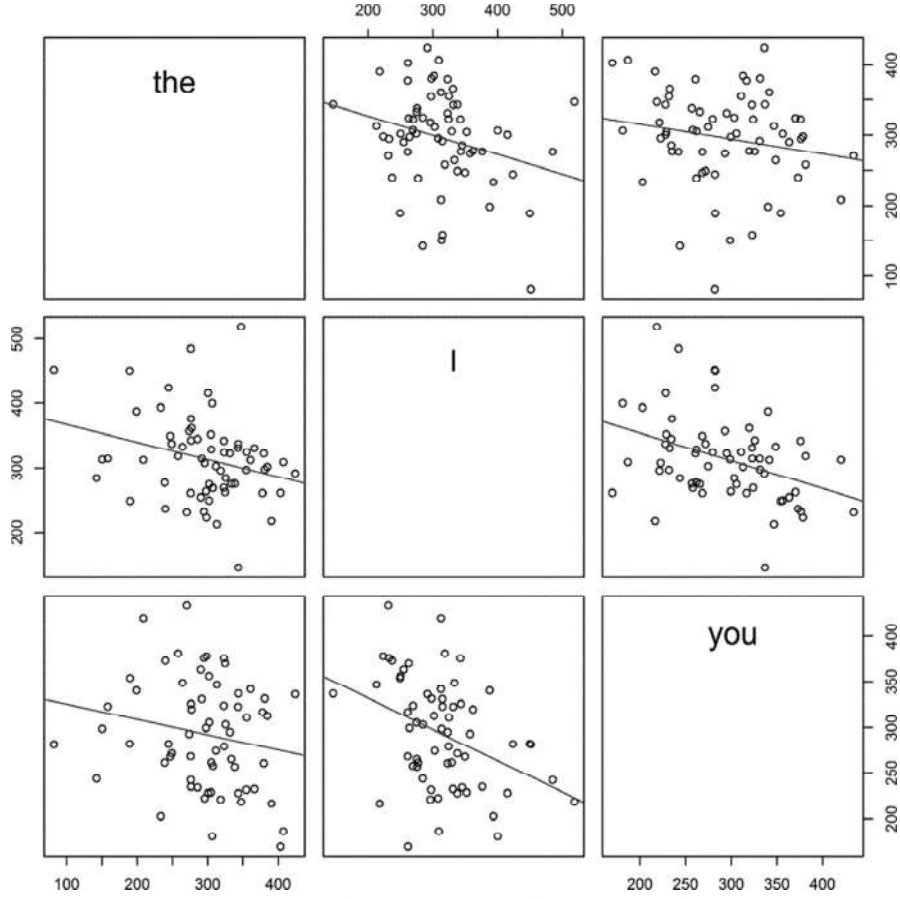
المدونة الوطنية البريطانية التي تتلازم مع الفعلين travel و go، وهذه الأماكن حسب الترتيب التكراري هي: لندن، وباريس، وأكسفورد، وروما، وكمبريدج، ومانشستر، ونيويورك، وليدز، وإدنبرة، وليفربول.



شكل (١٤، ١). مخطط تشتت لأداة التعريف the وضمير المتكلم I في مدونة BNC64.

ينبغي أن يوفر لنا نوع التمثيل الذي نستخدمه متوسطات لفهم الأنماط العامة في بيانات المدونة بنحو أفضل. يعتمد هذا إلى حد كبير على سؤال البحث ونوع الدراسة (تصميم البحث) الذي نتعامل معه. يوفر الشكل (١٧، ١) نظرة عامة إلى أنواع مختلفة من الرسومات البيانية الأكثر استخدامًا في هذا الكتاب مع إشارات إلى الفصول التي ناقشتها بالتفصيل.

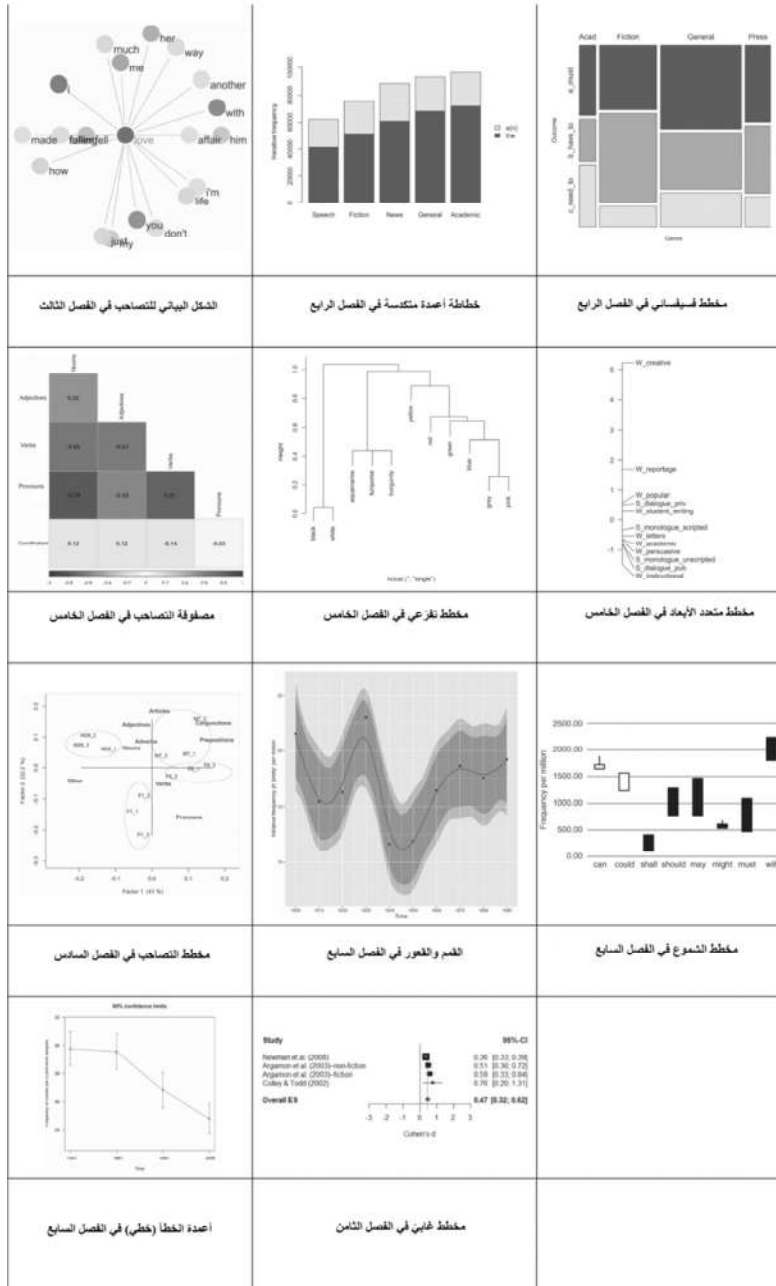
يمكن إنتاج جميع الرسوم البيانية المقدمة في هذا البحث بسهولة باستخدام أداة الشكل البياني من Lancaster Stats Tools عبر الإنترنت.



شكل (١٥). مخطط تشتت: the و I و you في مدونة BNC64.



شكل (١٦). أكثر الأماكن ارتباطاً باستخدام go و travel في المدونة الوطنية البريطانية.



شكل (١٧، ١). أنواع أخرى للتمثيل.

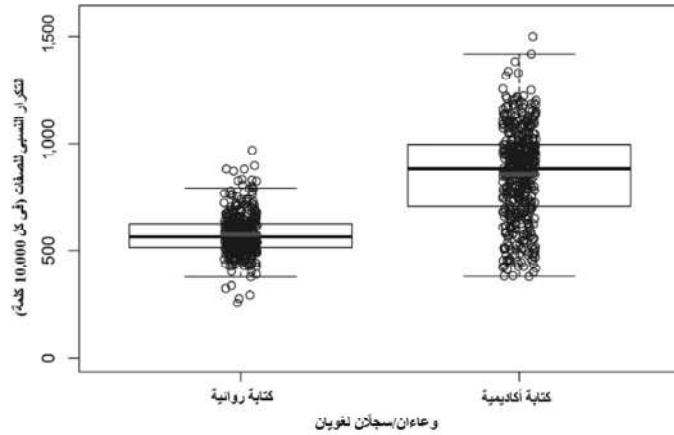
(٦، ١) التطبيق وأمثلة إضافية: هل يستخدم كُتّاب الرواية الصفات أكثر من الأكاديميين؟

أخبرتني صديقة لي، وهي كاتبة روائية ممتازة، أنها تعتقد أن الكتابة الأكاديمية جافة، لأن الأكاديميين يُقلّون جدًّا من استخدام الصفات. فكرتُ في هذا للحظة، وبدلاً من إصدار رد ذكي (لم أستطع التفكير فيه على أي حال)، قلت: "هذه مسألة تجريبية". قررت في حينها التحقق من فرضيتها في المدونة الوطنية البريطانية BNC التي تتضمن نصوصاً أكاديمية وروائية. وفيما يأتي تقرير قصير أعددتُه لتلك الصديقة، ونظراً لأنها غير ملمة بالمدونات وعلم الإحصاء استخدمت هوامش التقرير لوضع الملحوظات التوضيحية.

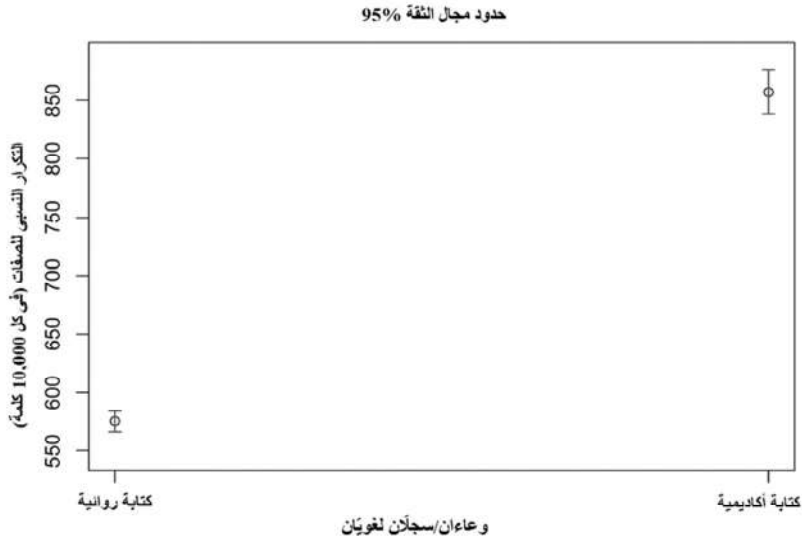
تقوم هذه الدراسة على المدونة الوطنية البريطانية. أُستخرجت على وجه الخصوص مدونتان فرعيتان من المدونة الوطنية البريطانية للتحقيق في الفرق في استخدام الصفات بين الكتابات الروائية والكتابات الأكاديمية. يبين الجدول (٥, ١) هاتين المدونتين الفرعيتين.	تمثل المدونة الوطنية البريطانية استخدام اللغة الإنجليزية البريطانية في أوائل تسعينيات القرن العشرين، ومن ثم فإن المدونة بالفعل قديمة إلى حد ما. على الرغم من ذلك، فإن استخدام الصفات متغير لغوي مستقر بدرجة كافية، لذلك لا داعي للقلق بشأن التمثيل كثيرًا. إذا كنت قلقًا بشأن هذا الأمر، فيمكن تكرار هذه الدراسة مع بعض البيانات الحديثة.									
جدول (٥, ١). المدونتان الفرعيتان في بحث مصغر.										
<table><tr><th>المدونة الفرعية</th><th>الرواية في المدونة الوطنية البريطانية</th><th>الكتابة الأكاديمية في المدونة الوطنية البريطانية</th></tr><tr><td>الكلمات (الكلمات الفعلية)</td><td>16,075,667</td><td>15,619,286</td></tr><tr><td>الملفات</td><td>432</td><td>501</td></tr></table>	المدونة الفرعية	الرواية في المدونة الوطنية البريطانية	الكتابة الأكاديمية في المدونة الوطنية البريطانية	الكلمات (الكلمات الفعلية)	16,075,667	15,619,286	الملفات	432	501	
المدونة الفرعية	الرواية في المدونة الوطنية البريطانية	الكتابة الأكاديمية في المدونة الوطنية البريطانية								
الكلمات (الكلمات الفعلية)	16,075,667	15,619,286								
الملفات	432	501								
في هذه الدراسة، يُمثل كل ملف نصًّا كتبه مؤلف واحد وبمنزلة نتيجة مشاهدة منفصلة. اختُبرت الفرضية التالية:	أي أنني نظرت إلى الفروق الفردية بين الكُتّاب، وليس في متوسطات المجموعة فقط.									

• الفرضية البديلة: يستخدم الأكاديميون صفات أقل من كُتاب الرواية.	هذه فرضيتك، تذكرين ذلك؟
استُكشفت البيانات لأول مرة باستخدام مخطط رسم المربع، ثم حُسبت مجالات الثقة ٩٥٪ للمدوتين الفرعيتين، واستُخدم قياس r^2 بوصفه إجراءً قياسياً لحجم التأثير. أخيراً، استُخدم اختبار-ت للعينات المستقلة لاختبار الفرضية الصفرية التي تفيد أنه لا فرق بين مجموعتي الكُتاب.	هذه تقنيات إحصائية مختلفة. لا حاجة إلى فهم التفاصيل.
الفرضية الصفرية: لا فرق بين استخدام الأكاديميين والكتاب الروائيين للصفات.	الفرضية الصفرية جزء من الإجراء الإحصائي الرسمي، وهي نفي فرضيتك.
تشير النتائج إلى وجود فرق بين الكُتاب الأكاديميين والكتاب الروائيين من جهة استخدامهم الصفات. ومع ذلك، كما توضح مخططات رسم المربع في الشكل (١٨، ١٩) فإن الفرق في مصلحة الكتاب الأكاديميين، أي إنهم يستخدمون صفات أكثر من كُتاب الرواية. يمكننا أيضاً ملاحظة تجانس أكبر بين كتاب الرواية أكثر من الكُتاب الأكاديميين.	يشار إلى التجانس الأكبر من خلال صندوق أصغر وخطوط دقيقة أقصر.
إذا تجاوزنا العينة، يمكن حساب مجالات الثقة ٩٥٪. ويكشف الشكل (١٩، ١٨) عن فئات ثقة غير متداخلة ومتباعدة بالنسبة لمجموعتي المقارنة. يمكن تفسير ذلك على أنه فرق واضح ودالّ إحصائياً للكتاب الأكاديميين الذين يستخدمون صفات أكثر بكثير من كتاب الرواية.	هذا أسلوب إحصائي لاختبار ما إذا كان هناك فرق أيضاً بالنسبة لكُتاب الرواية والأكاديميين عموماً وليس لكُتاب الرواية والأكاديميين الذين أخذت منهم عينات في المدونة الوطنية البريطانية (شريطة أن تكون المدونة الوطنية البريطانية عينة تمثيلية غير متحيزة).

أكد اختبار-ت للعينات المستقلة على وجود فرق ذي دلالة إحصائية بين كُتاب الرواية والكُتاب الأكاديميين. المزيد من الأساليب الإحصائية فقط لتأكيد النقطة نفسها.



شكل (١٨, ١). استخدام الصفات من قبل الكتاب الأكاديميين والروائيين: خطط رسم المربع.



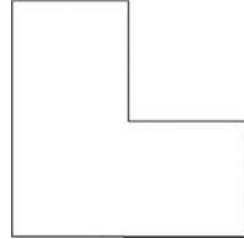
شكل (١٩, ١). استخدام الصفات من قبل الكتاب الروائيين والأكاديميين: أعمدة الخطأ.

الفرق ليس دالاً إحصائياً فقط بل يبدو أيضاً ذا مغزى لغوي أسلوبى.	فنتائج الكتاب هي: (ت = 67.714) = -26.78 - وقيمة الاحتمال $(r = 0.708)$ ، والذي يمكن ربطه بالواقع اللغوي وهو الفرق بين 857 صفة لكل 10,000 كلمة (مجال أكاديمي) و575 صفة لكل 10,000 كلمة (مجال روائي). وهذا يعني أن الكتاب الأكاديميين في المتوسط يستخدمون في كل 10 آلاف كلمة صفات أكثر من كتاب الرواية بما يزيد على 280 صفة. يبدو هذا فرقاً ذا مغزى من الناحية الأسلوبية، أي فرقاً ملحوظاً من القارئ.

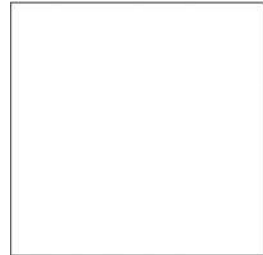
قدمتُ التقرير إلى صديقتي مع ابتسامة عريضة على وجهي. قالت: "إممم ممتع ..."، لكن ما زلت أعتقد أن الصفات التي يستخدمها كتاب الرواية هي أكثر ثراءً، ثم صرخت في سخط قائلة: "هذه مسألة بحثية مختلفة!" وبعد برهة قلتُ: "لعلك على حق، إذن ... لماذا لا نكتشف ذلك؟"

(١, ٧) تمارين

١ - لتنشيط الذهن (بشيء من التحدي): قسّم الشكل التالي الذي يمثل ثلاثة أرباع المربع إلى أربعة أشكال متماثلة. لا تتردد في تخطي هذا التمرين إذا كنت في عجلة إلى التركيز في الأساليب الإحصائية.



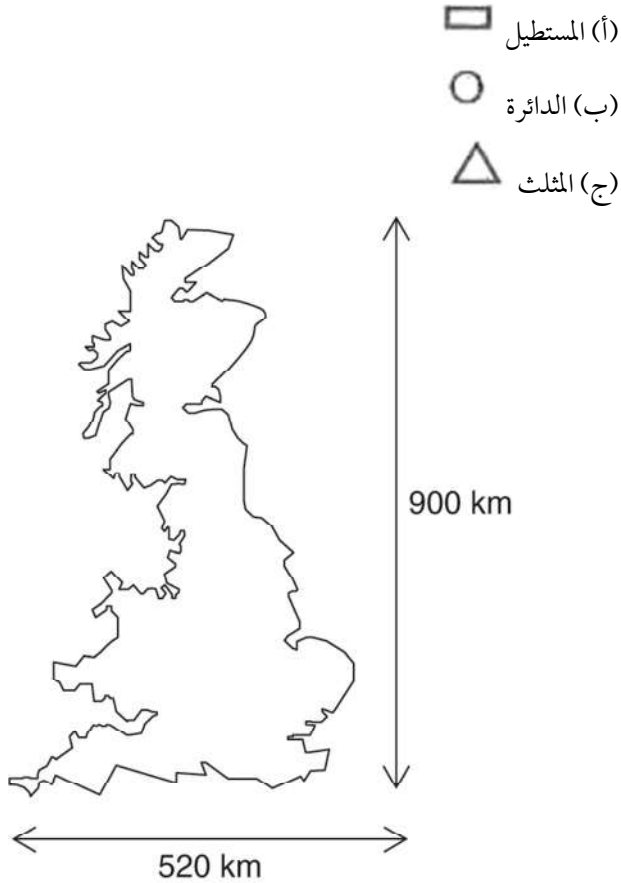
خذ مربعاً كاملاً كما في الشكل الآتي، ولكن هذه المرة قسّمه إلى خمسة أشكال متطابقة.



٢- احسب المتوسط للأرقام التالية: ٢٣٣٩ / ٢٠٩٨ / ٢٠٥٦ / ٢٢٧٦ / ٢٢٣٣ / ٢٠٥٦ / ٢٢٤١ / ١٩٩٥ / ٢٠٤٣ / ١٩٧٦ / ٢٠٦٢. هذه هي تكرارات الأفعال في النصوص الروائية للكتاب البريطانيين التي نوقشت في هذا الفصل (المبحث ٢، ١).

٣- ما النموذج في التفكير العلمي؟

٤- حدّد النموذج الهندسي الأكثر ملاءمة لمنطقة بريطانيا العظمى في الشكل (١، ٢٠) التي من شأنها أن تساعدك على استكشاف مساحة الجزيرة.



شكل (١، ٢٠). جزيرة بريطانيا العظمى.

٥- استخدم النموذج والأبعاد الواردة في الشكل (١, ٢٠) لحساب مساحة بريطانيا العظمى (الجزيرة الكبيرة فقط).

٦- اختبر معلوماتك عن المصطلحات الإحصائية الأساسية.

أولاً: ما الفرق بين المعدل average والمتوسط mean؟

(أ) لا فرق، فهما مصطلحان مترادفان.

(ب) المتوسط mean هو نوع من المعدل average وكذلك الحال بالنسبة للوسيط median

والمتوسط المرتب trimmed mean.

(ج) مفهوم المتوسط mean عادة أكبر من مفهوم المعدل average.

ثانياً: ما متوسط mean القيم التالية: 5, 10, 15, 20, 25؟

(أ) 15.0

(ب) 17.32

(ج) 25.3

ثالثاً: ما وسيط median القيم التالية: 5, 10, 15, 20, 25؟

(أ) 10

(ب) 15

(ج) 20

رابعاً: ما نوع متغير رتبة الكلمات في قائمة التكرارات؟

(أ) اسمي

(ب) ترتيبي

(ج) مقياسي

خامساً: ما التشتت في المدونة؟

(أ) توزيع المتغير (اللغوي) في أجزاء مختلفة من المدونة.

(ب) مصطلح آخر للانحراف المعياري.

(ج) انتشار عينة من المجموع اللغوي.

سادساً: إذا رسمت مجموعة من البيانات الموزعة طبيعياً، فما شكل الشكل البياني؟

(أ) مسطح

(ب) منحنى

(ج) على شكل جرس

سابعاً: ما مجال الثقة ٩٥٪؟

(أ) مجال يُوضح إمكان أن نكون واثقين بنسبة ٩٥٪ من صحة النتيجة في هذه الفئة.

(ب) مقياس لموضوعية نتائجنا.

(ج) مجال مبني حول قياس معين في عينة بطريقة ستقع فيها القيمة الحقيقية للمقياس في

المجموع اللغوي في هذه الفئة في ٩٥٪ من العينات.

ثامناً: ما قيمة الاحتمال؟

(أ) احتمال أن تكون الفرضية الصفرية صحيحة.

(ب) احتمال رؤية قيم لا تقل تطرفاً عما لو كانت الفرضية الصفرية صحيحة.

(ج) احتمال رؤية وحيد القرن في لانكستر.

٧- تخيل أن لديك ٥٠٠ نص (المجموع اللغوي المعني) ٢٥٠ نصاً منها كتبها ذكور و ٢٥٠ نصاً منها كتبها إناث، ومع ذلك، أنت لا تعرف أيها كتبها الذكور وأيها كتبها الإناث. في طيات دراستك، تريد اختيار عينة غير متحيزة من ٤٠ نصاً تمثل المجموع اللغوي. استخدم مؤلّد الأرقام العشوائية من أدوات الإحصاءات الخاصة بشبكة لانكستر عبر الإنترنت لإنشاء قائمة مكونة من ٤٠ رقماً عشوائياً بين ١ و ٥٠٠ ثم دوّن هذه الأرقام.

٨- في مبحث (الإجابة) على موقع الويب المصاحب، يمكنك التعرف على النصوص التي أنتجها الذكور والنصوص التي أنتجها الإناث. تحقّق من الإجابات في التمرين (٧) واحسب عدد المتحدثين من الذكور والإناث الذين وقع عليهم الاختيار.

----- المتحدثون الذكور في العينة: -----

----- المتحدثات الإناث في العينة: -----

هل حصلت على تمثيل متساوٍ تقريباً للمتحدثين من الذكور والإناث في العينة؟

٩- تكتشف لاحقاً أن نصف النصوص البالغ عددها ٥٠٠ نص أنتجها متحدثون صغار، ونصفها الآخر أنتجها متحدثون كبار. (راجع الإجابة عن السؤال في موقع الويب المصاحب لتمييز النصوص). راجع العينة من التمرين (٧) لمعرفة ما إذا كان لديك تمثيل متساوٍ تقريباً بين الجنسين والأعمار في عينتك المكونة من ٤٠ نصاً.

النوع	السن	متحدثون ذكور	متحدثون إناث
متحدثون صغار			
متحدثون كبار			

١٠- هل كان أخذ العينات العشوائية طريقة ناجحة؟ لماذا نعم؟ / لماذا لا؟

١١- ما أنواع التحيز التي نحتاج إلى تجنبها في تصميم المدونة؟

١٢- ما نوع تصميم البحث (مجموعة كاملة، أو سمة لغوية، أو نص/متحدث فردي) الذي تستخدمه في المواقف الآتية؟

(أ) لمعرفة أي الاسمين الموصولين: that و which أكثر تكراراً بين اللغتين المكتوبة والمنطوقة عندما يحل محل العلامة ϕ في المثال الآتي: (The second thing ϕ I want to say is) [الشيء الثاني الذي أودّ أن أقوله هو].

(ب) لمعرفة ما إذا كانت التحرّز الكلامي (على سبيل المثال، sort of, kind of) أكثر شيوعاً في خطاب الذكور أو الإناث.

(ج) لمعرفة تكرار أداة التعريف the في اللغة الإنجليزية المكتوبة.

(د) للتحقيق في الاختلاف القائم على السجل اللغوي لعدد كبير من السمات اللغوية بما في ذلك الأفعال الناقصة ومحدّدات الخطاب والأفعال الخاصة.

١٣- هل يمكنك اكتشاف ستة أخطاء في مجموعة البيانات التالية استناداً إلى مدونة BE06 المكونة من مليون كلمة تقريباً للغة الإنجليزية البريطانية المكتوبة؟

الكلمة أو التعبير	التكرار	التكرار كل مليون كلمة
The	5,896	5142.17
Of	30,666	26745.23
And	27,909	24340.72
To	26,188	2283.98
of the	6,887	6006.47
and the	19,530	17033.01
إجمالي الكلمات	2,293,194	-

١٤ - ما نوع التمثيل (الشكل البياني) الذي سيكون مناسباً في المواقف التالية؟

(أ) وصف توزيع التكرار frequency distribution لمتغير لغوي في مدونة واحدة.

(ب) مقارنة توزيع المتغير اللغوي في مدونتين.

(ج) إيجاد العلاقة بين متغيرين لغويين.

(د) تقدير ما إذا كانت الاختلافات بين تكرارات متغير لغوي قابلة للتعميم على المجموع اللغوي.

١٥ - استخدم أداة رسم بياني من أدوات Lancaster Stats Tools عبر الإنترنت والبيانات المقدمة هناك لإنشاء رسومات بيانية تعرض الأنماط الرئيسية في مجموعات البيانات هذه.

تذكر الآتي

- لغويات المدونة الحاسوبية هي طريقة علمية.
- التطبيق الناجح للأساليب الإحصائية يعتمد في لغويات المدونة الحاسوبية على استخدام المدونة الجيدة لإنشاء غير المتحيزة.
- علم الإحصاء يستخدم التعبيرات الرياضية لمساعدتنا على فهم البيانات الكمية.
- التمثيل الفعال يُلخّص الأنماط في البيانات دون إخفاء السمات المهمة.
- على الرغم من أن قيم الاحتمال أكثر استعماً فهي لا تشكل سوى جزء (صغير) من علم الإحصاء.
- (الدلالة الإحصائية) و(الأهمية التطبيقية) و(المغزى اللغوي) هي ثلاثة أبعاد منفصلة ينبغي عدم الخلط بينها.

قراءات إضافية

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, 5(4), 243-57.

- Biber, D. & Jones, K. (2009). Quantitative methods in corpus linguistics. In A. Lüdeling & M. Kytö (eds.), *Corpus linguistics: an international handbook*, vol. 2, pp. 1287–1304. Berlin: Walter de Gruyter.
- Cumming, G. (2012). *Understanding the new statistics: effect sizes, confidence intervals, and meta-analysis*. New York: Routledge.
- Diggle, R. J. & Chetwynd, A. G. (2011). *Statistics and scientific method: an introduction for students and researchers*. Oxford University Press.
- Gries, S. Th. (2006). Some proposals towards a more rigorous corpus linguistics. *Zeitschrift für Anglistik und Amerikanistik*, 54(2), 191-202.
- Leech, G. (1992). Corpora and theories of linguistic performance. In J. Svartvik (ed.), *Directions in corpus linguistics*, pp. 105-22. Berlin: Mouton de Gruyter.
- McEnery, T. & Hardie, A. (2011). *Corpus linguistics: method, theory and practice*. Cambridge University Press.
- Olcasha, S. (2002). *Philosophy of science: a very short introduction*. Oxford University Press.
- Salsburg, D. (2001). *The lady tasting tea: how statistics revolutionized science in the twentieth century*. London: Macmillan.
- Tufte, E. (2006). *Beautiful evidence*. Cheshire, CT: Graphics Press.
- Vickers, A. (2010). *What is p-value anyway? 34 stories to help you actually understand statistics*. Boston: Addison-Wesley.
- Yau, N. (2011) *Visualize this: the Flowing Data guide to design, visualization, and statistics*. Indianapolis: Wiley.

أدوات لانكستر الإحصائية

١- يمكن استنساخ جميع التحليلات المبينة في هذا الفصل باستخدام أدوات لانكستر الإحصائية. الأدوات المتاحة لهذا الفصل تشمل:

- حاسبة إحصائية.
- الموزع العشوائي.
- أداة الشكل البياني.

٢- يقدم الموقع أيضًا موادًا إضافية للطلاب والمعلمين.

الفصل الثاني

الكلمات: التكرار والتشتت والتنوع

(١, ٢) ما الذي يتناوله هذا الفصل؟

تتكون المدونات من أحجام كبيرة من آلاف أو ملايين أو حتى مليارات الكلمات. لا يقتصر العمل على المدونات على مجرد أوعية للكلمات نستخرج منها كلمة تلو الأخرى كحال الساحر الذي يستخرج الأرباب من القبة، بل الأمر بخلاف ذلك تماماً، فعند تحليل مدونة، نحتاج إلى التفكير ملياً في تكرارات الكلمات وتوزيعاتها لإيجاد أنماط ذات مغزى في استخدام اللغة. سننظر في هذا الفصل مقاييس إحصائية بسيطة من شأنها أن تساعدنا على وصف ورود الكلمات في النصوص والمدونات وذلك بالإجابة عن خمسة أسئلة رئيسة.

١- ما الكلمة؟ (المبحث ٢, ٢).

٢- كيف نقيس تكرارات الكلمات والعبارات؟ (المبحث ٣, ٢).

٣- كيف نقيس توزيعات الكلمات والعبارات؟ (المبحثان ٤, ٢ و ٥, ٢).

٤- كيف نقيس التنوع المعجمي؟ (المبحث ٦, ٢).

٥- كيف يمكن استخدام هذه القياسات في البحث اللغوي والاجتماعي؟ (المبحث ٧, ٢).

(٢, ٢) الكلمات الفعلية، والكلمات النوعية، والمواد المعجمية، واللكسيات

فكّر...

قبل قراءة هذا المبحث، فكر في الأسئلة التالية:

١- ما تعريفك للكلمة؟

٢- ما عدد الكلمات التي تتكون منها الجملتان التاليتان؟

Took them 26 years to win the title. During that time we won it 3 times despite being in the second division for half that time.

نتعامل في لغويات المدونة مع الكلمات والعبارات، وهذه في المصطلحات الإحصائية هي المتغيرات اللغوية التي نريد تحليلها (المبحث ٣، ١). نحسب في لغويات المدونة عادةً الكلمات في سياقات مختلفة، ثم نقارن هذه الأعداد لتحديد أنماط استخدامها في اللغة، وسنمر في هذا الكتاب على كثير من الأمثلة من هذا النوع مما يُسمى (اكتشاف النمط). ولحساب عدد الكلمات بطريقة موثوقة، نحتاج بالطبع إلى معرفة ما نحسبه، ومن ثم فعلينا أولاً تعريف الكلمة. قد يبدو هذا تافهًا إلى حد ما لكوننا نعلم معنى الكلمة، ولكن في الواقع، يعد تعريف الكلمة معقدًا جدًا. يمكن توضيح ذلك من خلال جمل المثال الوارد في مهمة (فكر):

13	12	11	10	9	8	7	6	5	4	3	2	1
won	we	time	that	During	title.	the	win	to	years	26	them	Took
26	25	24	23	22	21	20	19	18	17	16	15	14
time	that	half	for	division	second	the	in	being	despite	times	3	it

(المصدر: المدونة الوطنية البريطانية J1H، BNC)

ملاحظة: تشير الأرقام في المثال أعلاه إلى صيغ الكلمات الفردية مفصولة بمسافة.

كم كلمة في الجملتين؟ إذا سألت أحد لغويي المدونات، فستكون إجابته عن السؤال أعلاه غالبًا كما يأتي: ٢٦ كلمة فعلية tokens و ٢٣ كلمة نوعية types و ٢١ مادة معجمية lemmas و ٢٢ لكسيماً lexeme. دعونا نحاول فك شفرة هذه المعلومات: **الكلمة الفعلية** (token (running word هي الحدوث الفردي لصيغة من الصيغ التي تأخذ حيزًا مستقلًا في النص، وعليه فإن كل سلسلة متصلة من الأحرف أو الأرقام المفصولة بمسافة أو علامة ترقيم نُعدّها كلمة واحدة. عندما يسأل الطلاب أسئلة مثل: كم كلمة عليّ كتابتها في موضوع المقالة؟ يسألون بذلك عن عدد الكلمات الفعلية عادةً دون أن يعرفوا أي شيء عن مصطلح (الكلمة الفعلية). عندما نحسب عدد مرات كل حدوث فردي لصيغة من صيغ كلمة ما في الجملتين في مهمة (فكر)، فإننا نحصل على الرقم (26). تجدر الإشارة إلى أن التحديد الدقيق لما يُعد كلمة فعلية يختلف بمختلف تطبيقات أدوات التحليل الجاهزة للمدونات. يمثل هذا مشكلة حقيقية بالنسبة لموثوقية النتائج وتناحيثها (أي قابليتها لإعادة تطابق نتائجها)؛ نظرًا لأن عدد الكلمات الفعلية للمدونة نفسها، مثل المدونة الوطنية البريطانية، قد يختلف

بنسبة تصل إلى ١٧٪^(١) اعتماداً على الأداة المستخدمة (Brezina & Timperley 2017). ينبغي عند استخدام أدوات المدونة أن نكون على بينة كاملة من كيفية تطبيق تعريف الكلمة الفعلية بدقة، ومن قابلية مقارنة الأعداد عند الاعتماد على أدوات مختلفة للمقارنة نظراً لأن لأعداد الكلمات الفعلية تأثيراً كبيراً على جميع القياسات الإحصائية استناداً إلى تكرار الكلمة والعبارة (مثل التكرارات النسبية، وقياسات الكلمة الرئيسية، وقياسات المتصاحبات اللفظية، والاختبارات الإحصائية باستخدام التكرارات النسبية، وما إلى ذلك).

الكلمة النوعية عبارة عن صيغة فريدة للكلمة في المدونة. عندما نسأل عن أنواع الكلمة فإننا نسأل عن عدد صيغ الكلمة المختلفة الموجودة في النص/المدونة. في الجملتين الواردتين في مهمة (فكر)، هناك ثلاث صيغ: the (رقم 7 و20)، that (رقم 10 و25) و time (رقم 11 و26)؛ وقد ورد كل واحدة منها مرتين، ونُحسب كل واحدة منها باعتبارها كلمة نوعية واحدة على حدة، ولذلك كانت المحصلة ٢٣ كلمة فقط، وكلمة (أوقات times) لا تُعدّ مع كلمة (وقت time) كلمة واحدة وفقاً لتعريف الكلمة الفعلية.

تُحدّد كل من الكلمات الفعلية والكلمات النوعية بناءً على صيغة الكلمة ومظهرها الخارجي، ولتحديد المواد المعجمية واللكسيات، نحتاج أولاً إلى إجراء تحليل لغوي للنص، إذ تعتمد المواد المعجمية على التحليل الصرفي، في حين تعتمد اللكسيات على التحليل النحوي/الصرفي والدلالي. المادة المعجمية lemma هي مجموعة من جميع الصيغ التصريفية ذات الصلة بجذع ينتمي إلى فئة الكلمات نفسها (Kucera & Francis 1967: 1).^(٢) يبدو هذا التعريف الصرفي معقداً، لكن المبدأ بسيط

(١) تتضمن بعض الأدوات (مثل CQPweb و Sketch Engine) علامات الترقيم في تعداد الكلمات الفعلية، وتلجأ أدوات أخرى (مثل #LancsBox) لجوءاً كبيراً إلى التعريف البسيط والسطحي لـ (الكلمة الفعلية) دون حساب عدد علامات الترقيم. وفضلاً عن علامات الترقيم، تشمل مصادر التنوع الأخرى في عد الكلمات الفعلية ما يأتي: المتصلات (مثل 'll في 'he'll) والكلمات الموصولة بشرطة أو أكثر (well-known)، ونتائج التجريد tokenization استناداً إلى التحليل الصرفي (وسم أجزاء الكلام)، وكذلك التجريد للغات غير المقطعة (مثل الصينية).

(٢) يمكن القول إن هذا التعريف التقليدي لـ (المادة المعجمية) lemma هو الأكثر فائدة لتحليلات المدونة. في بعض التحليلات الإجرائية نجد مشكلة عدم معالجة (الانتهاء إلى قسم الكلمة نفسه)، ويؤدي هذا إلى حالات غير منطقية نوعاً ما، فعلى سبيل المثال: ستُصنّف (go) في: it took three goes تحت المادة المعجمية الفعلية كـ: they went في went.

لِلغاية، فنحن نجمع الصيغ التي لها صيغة الأساس نفسها، وتختلف فقط فيما يتعلق بالقواعد النحوية والصرفية، مثل صيغ المفرد والجمع للاسم نفسه، والحاضر والماضي للفعل نفسه، والصيغة الأصلية البسيطة للصفة وصيغة التفضيل الأعلى لها، وهلم جرا. أما اللكسيم ^(٣) lexeme فمادة معجمية لها معنى خاص مرتبط بها، وبها يُميز بين المشتركات الدلالية polysemous (الكلمات ذات المعاني المتعددة)، وأفضل طريقة لتمثيل اللكسيم هي بجعله مدخلاً فرعياً في قاموس، فعندما تفتح قاموساً ترى قائمة أبجدية للصيغ (الكلمات الرئيسية) التي ينقسم كثيرٌ منها إلى مداخل فرعية subentries حسب معناها، وهذه هي اللكسيمات. تسرد بعض القواميس أيضاً صيغاً صرفية متعلقة باللكسيمات، ففي جملة مهمة (فكر) ينتمي كل من كلمة win و won إلى المادة المعجمية نفسها؛ لأن won هي صيغة الماضي للفعل win، وعلى هذا المنوال فإن كلاً من time و times وحدتان لمادة معجمية واحدة لأن times هي صيغة الجمع لكلمة time. وبناء على ذلك، نحصل على ٢١ مادة معجمية. ولكن، كلمة time مشترك دلالي يمكن أن تعني عدداً من الأشياء المختلفة، فهي في المثال تشير إما إلى فترة زمنية (11 و 26) وإما إلى التكرار؛ أي: عدد مرات حدوث شيء ما (16). عندما نعد اللكسيمات، نحتاج إلى التمييز بين هذين الاستخدامين، وعليه يكون لدينا ٢٢ لكسيماً في المثال.

لنفكر في كيفية تأثير مفاهيم كل من الكلمة النوعية، والمادة المعجمية، واللكسيم في نوع التحليل الذي يمكننا إجراؤه في لغويات المدونة الحاسوبية. نجد أن الأسلوب الأبسط والأوضح يكمن في استخدام الكلمات النوعية استناداً إلى سهولة التمييز بين صيغ الكلمات المختلفة بغض النظر عن وظيفتها النحوية أو الصرفية أو معانيها. مع أن الكلمة النوعية فئة مفيدة للغاية، فإنها قد تحجب بعض الاختلافات المهمة، فعلى سبيل المثال استخدامات صيغة (نظيف) clean صفةً (قميص نظيف a clean shirt) مقابل الفعل (ينظف شيئاً to clean something). إذا ركزنا في المادة المعجمية بوصفها وحدة للتحليل نحتاج إلى معالجة المدونة آلياً من أجل تعيين القسم الكلامي من أقسام الكلام لكل صيغة، ثم إلى تجميع كل الصيغ التصريفية المرتبطة بكل قسم، وستعطي هذه المعالجة

(٣) يُستخدم المصطلح (لكسيم) lexeme أحياناً مترادفاً مع المصطلح (المادة المعجمية) lemma، وأوافق هنا باير وآخرين Biber et al. (١٩٩٩: ٥٤) في جعل اللكسيم على النحو الآتي: "مجموعة من الصيغ ذات الصلة تشترك في المعنى نفسه وتنتمي إلى قسم الكلمة نفسه".

نسبة معينة من الأخطاء. أما معالجة اللكسيات فهي تعين على الوسم الدلالي وإزالة غموضه، وتفيدنا بنسبة مئوية معينة من الأخطاء تكون أعلى دقة من توسيم أقسام الكلام، غير أن ذلك لا يمكن القيام به كاملاً بالمعالجة الآلية. يلخص الجدول (١, ٢) مزايا المفاهيم المختلفة وعيوبها (الكلمة).

جدول (١, ٢). الكلمة النوعية والمادة المعجمية واللكسيم: المزايا والعيوب.

تحديد الكلمة	المزايا	العيوب
كلمة نوعية type	فئة ذات استدلال منخفض	لا يوجد فرق بين الصيغ ذات الوظائف النحوية والصرفية و/أو المعاني المتعددة
مادة معجمية lemma	التفريق بين الصيغ ذات الوظائف النحوية والصرفية المتعددة	وسم أقسام الكلام وتجريد الكلمة المعنوية (مصادر) (تحتل الخطأ)
لكسيم lexeme	أخص فئة لكونها تراعي الفروق من جهة المعنى	فئة ذات استدلال مرتفع (مصادر محتملة للخطأ) وليست متاحة على أساس آلي بحث حتى الوقت الحالي

نظرًا لأن تعداد الكلمات (للكلمات الفعلية والنوعية في الغالب) جزء من كل معادلة إحصائية مطبقة في هذا الكتاب، من المهم أن تكون على فهم جيد لهذه التعريفات، وستساعدك التمارين في نهاية هذا الفصل مع الإجابات المتاحة في أدوات لانكستر على اختبار مدى فهمك لهذه المفاهيم الأساسية (المبحث ٨, ٢).

التقرير الإحصائي: الكلمات الفعلية، والكلمات النوعية، والمواد المعجمية واللكسيات

١ - ما يجب الإفصاح عنه

نحتاج عند الحديث عن الكلمات في لغويات المدونة الحاسوبية إلى تحديد ما إذا كنا نعني الكلمات الفعلية، أو الكلمات النوعية، أو المواد المعجمية، أو اللكسيات. عند وصف المدونات، وينبغي لنا دائمًا ألا نُغفل المعلومات المتعلقة بعدد الكلمات الفعلية؛ لأن عدد الكلمات الفعلية يختلف من أداة إلى أخرى، وينبغي لنا أيضًا توفير تفاصيل عن الأداة وطبيعة معالجتها لعدد الكلمات الفعلية.

٢- مثال الإفصاح في التقرير

- يتكون النص من ١٢٠ كلمة فعلية ويتضمن ٦٩ كلمة نوعية.
- استخدمنا المدونة الوطنية البريطانية التي تضم ١٠٠ مليون كلمة في دراستنا: (عدد الكلمات الفعلية بدقة: 98,313,429 مع استبعاد علامات الترقيم).

(٢, ٣) كلمات في قائمة التكرار

فكّر...
 قبل قراءة هذا المبحث، فكر في السؤال التالي:
 ما الكلمات الأكثر شيوعًا في اللغة الإنجليزية؟ حاول الخروج بقائمة بأكثر عشرة كلمات تكرارًا:
 (١) _____ (٢) _____ (٣) _____ (٤) _____
 _____ (٥) _____
 (٦) _____ (٧) _____ (٨) _____ (٩) _____
 _____ (١٠) _____

جدول (٢, ٢). أكثر ١٠ كلمات تكرارًا في المدونة الوطنية البريطانية.

المرتبة	الكلمة	التكرار المطلق	التكرار النسبي / مليون كلمة
1	the	6,041,234	61,448.72
2	of	3,042,376	30,945.68
3	and	2,616,708	26,615.98
4	to	2,593,729	26,382.25
5	a	2,164,238	22,013.66
6	in	1,937,819	19,710.62
7	that	1,118,985	11,381.81
8	it	1,054,279	10,723.65
9	is	990,281	10,072.69
10	was	881,473	8,965.95

في الجدول (٢, ٢) عرضٌ للكلمات العشر الأكثر تكرارًا في المدونة الوطنية البريطانية المتوازنة من حيث المجالات^(٤)، ومرتبة قياسيًا حسب التكرار المطلق والنسبي. التكرار المطلق (أو الخام) absolute frequency (AF) هو الإحصاء الأبسط والأوضح، وهو العدد الفعلي لجميع تكرارات كلمة معينة في المدونة. بالعودة إلى المصطلحات الواردة في المبحث (٢, ٢) فإن التكرار المطلق هو عدد الكلمات الفعلية في المدونة، وعليه يمكنك أن ترى أن أكثر أنواع الكلمات تكرارًا في الجدول (٢, ٢) هو أداة التعريف the. إذا تصفحت كل نص في المدونة الوطنية البريطانية وحسبت عدد تكرارات the (تقوم الأدوات الحاسوبية بالحساب في لمح البصر)، فستحصل على الرقم 6,041,234. يُعدّ التكرار المطلق للكلمات مقياسًا مفيدًا عندما ننظر في مدونة واحدة، ويُمكننا من فرز قائمة كلمات للحصول على العناصر الأكثر تكرارًا كما في الجدول (٢, ٢). أما عندما نريد مقارنة مدونتين أو أكثر^(٥)، نستخدم ما يسمى التكرار النسبي (أو المعير) relative (or normalized frequency (RF). يمكن حساب التكرار النسبي بسهولة بالغة على النحو الآتي:

$$(٢, ١) \quad \text{التكرار النسبي} = \frac{\text{التكرار المطلق}}{\text{عدد الكلمات في المدونة}} \times \text{أساس المعيرة}$$

على سبيل المثال: حُسب التكرار النسبي لأداة التعريف the، كما هو موضح في الجدول (٢, ٢)، على النحو الآتي:

$$(٢, ٢) \quad 61,448.72 = 1,000,000 \times \frac{6,041,234}{98,313,429} = \text{the أداة التعريف}$$

اخترنا المليون بوصفه أساس المعيرة، وهو الأساس المألوف في لغويات المدونة الحاسوبية. يعني ذلك أن المتوسط هناك هو أكثر من 61,000 حالة ورود لأداة التعريف the في كل مليون كلمة فعلية في المدونة، وليس بعيد أن تكون أداة التعريف the في أعلى قائمة الكلمات تكرارًا، وبمعدل

(٤) عدد الكلمات الدقيق للمدونة الوطنية البريطانية المستخدمة في حساب التكرار النسبي في الجدول ٢, ٢ هو 98,313,429.

(٥) من الصعب العثور على مدونات تحتوي على العدد نفسه من الكلمات، وذلك حتى في المدونات المصممة وفقًا لإطار العينات نفسها، مثل مدونات عائلة براون التي تختلف كل واحدة منها اختلافاً بسيطاً من حيث إجمالي عدد الكلمات. كقاعدة عامة، فالمطلوب هو استخدام التكرارات النسبية عند مقارنة تكرارات الكلمات أو العبارات بين مدونتين أو أكثر.

تكرار مطلق يساوي تقريباً ٦٪ من العدد الإجمالي للكلمات الفعلية في جميع مدونات اللغة الإنجليزية المكتوبة. يمكن اعتبار التكرار النسبي متوسط التكرار mean frequency؛ وهو مصطلح إحصائي لأحد أنواع المتوسطات، ويشير إلى متوسط تكرارات (س) في عينات افتراضية من الكلمات الفعلية؛ إذ إن س هو أساس المعيرة من كل مليون كلمة (المبحث ٣، ١). ستبين فائدة هذه الفكرة عن متوسط التكرار لاحقاً عند مناقشة المقاييس الإحصائية الأخرى.

تكون أسس المعيرة التي تقل عن مليون كلمة مناسبة أكثر لإيجاد المعيرة لكل 10,000 أو حتى 1000 كلمة؛ لأن التكرار النسبي لا يُستخدم لمجرد مقارنة التكرارات من نوع معين في مدونتين (أو أكثر)، بل أيضاً لتقديم أدلة عن تكرار الكلمات والعبارات في صيغة أسهل في الفهم على القارئ من التكرار المطلق. إذا اخترنا للمعيرة أساساً كبيراً جداً بالنسبة لحجم المدونة الفعلية فقد يؤدي ذلك إلى (نسف) أرقامنا بأيدينا وتشويه الأدلة المحدودة والمتوفرة لدينا. تصوّر لو أن لديك مدونة صغيرة جداً تحتوي على 11,000 كلمة فعلية بما في ذلك الكلمة النوعية (التوازن [التلقائي جسمياً]) homeostasis التي تظهر مرة واحدة في تلك المدونة. إذا اخترنا المليون أساساً للمعيرة، فسنحصل على أكثر من 90 تكرار نسبي لكلمة homeostasis لكل مليون كلمة. على الرغم من أن النسبة صحيحة من الناحية الرياضية إلا أنّها مضللة من زاوية طبيعة سلوك الكلمات النادرة كحال هذه الكلمة. عند ورود كلمة نادرة مثل هذه الكلمة لمرة واحدة فقط في مدونة صغيرة فمن غير المرجح أن تَرِد حسب المعادلة الرياضية 90 مرة في المليون كلمة. قد ترد مرة واحدة في مدونة ما تتضمن مليون كلمة أو قد ترد خمس مرات أو ربما لا ترد على الإطلاق. لا تعطينا المدونة الفعلية الأصغر ببساطة إشارة استدلالية كافية. عندما تعطينا المعالجة تكراراً نسبياً يبلغ نحو 0.9 في كل 10,000 كلمة فمن المهم أن نبتعد عن استخدام التكرارات النسبية التي تخفي حقيقة التكرارات المطلقة، وأن نصرح بهذين النوعين معاً من التكرارات.

من المهم أن نتذكر دائماً أن المدونة ما هي إلا عينة من اللغة (المبحث ٤، ١) توفر أدلة حول استخدام الكلمات والعبارات. من الممكن تقييد هذا الدليل في المدونات الكبيرة نوعاً ما، ففي المدونة الوطنية البريطانية المكونة من ١٠٠ مليون كلمة يرد أكثر من نصف الكلمات

النوعية مرة واحدة فقط، وتسمى الكلمات الفريدة^(٦) hapax legomena (hapaxes)، ويرد ١٣٪ من الكلمات النوعية مرتين فقط، ويرد أقل من ٥٪ من الكلمات النوعية مئة مرة أو أكثر في المدونة (أي بتكرار نسبي واحد لكل مليون أو أكثر). هذا التوزيع للكلمات لا يقتصر على المدونة الوطنية البريطانية، لكنه ينطبق على جميع المدونات. غالبًا ما يشار إلى التكرار المتناقص للكلمات باسم قانون زيف Zipf's law.

تُجبرنا هذا القانون بأنه عندما نبدأ بالكلمة الأكثر تكرارًا في قائمة الكلمات بغض النظر عن حجم المدونة فإن الكلمة الثانية الأكثر تكرارًا لن يكون لها سوى نصف تكرار الكلمة الأولى، وسيكون للكلمة الثالثة الأكثر شيوعًا ثلث تكرار الكلمة الأولى، وهلم جرا. يتناقص مقدار الأدلة التي يمكننا الحصول عليها من المدونات عن الكلمات بسرعة، ويمكن ملاحظة ذلك في الشكل (١، ٢) الذي يوضح توزيع تكرارات الكلمات وفقًا لترتيب مراتبها في المدونة الوطنية البريطانية على مقياس خطّي (فوقّي) وعلى مقياس لوغاريتمي (سفليّ). إن عرض التكرارات على المحور الصادي (Y) على مقياس لوغاريتمي يعني أن كل علامة على المحور الصادي هي العلامة السابقة مضروبة في 10. يُظهر هذا النوع من التمثيل التوزيع عبر نطاق كبير من قيم التكرار (1 - 5.4 M)، وهذا يوفر مزيدًا من الدقة عن نقطة بداية انحراف المنحنى بالضغط. من ناحية أخرى يُظهر المقياس الخطّي (العلوي) مدى الانخفاض الرهيب لتكرارات الكلمات في المدونة.

وعلى طريقة المتخصصين، يمكن التعبير عن قانون زيف على النحو الآتي:

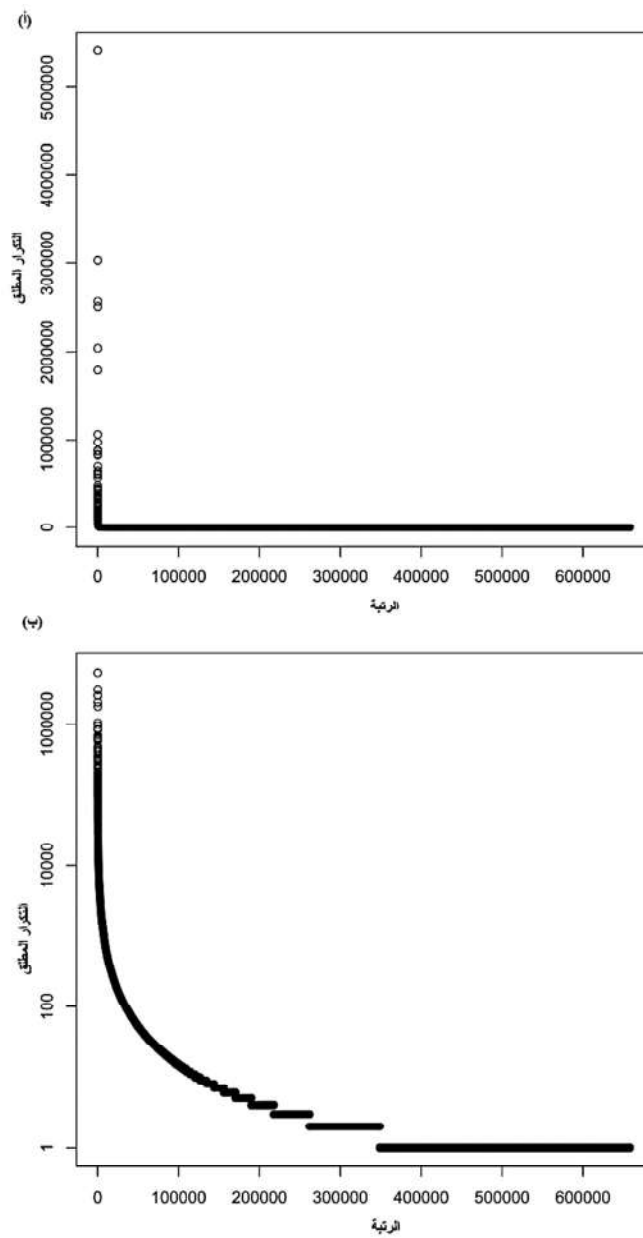
$$(٢, ٣) \quad \text{التكرار المطلق للكلمة } x \text{ مرتبتها في قائمة الكلمات} \approx \text{المقدار الثابت}$$

$$(٢, ٤)$$

$$\frac{\text{المقدار الثابت}}{\text{المرتبة في قائمة الكلمات}} \approx \text{التكرار المطلق}$$

والمقدار الثابت هو تكرار الكلمة الأولى في قائمة الكلمات.

(٦) يُعدّ مصطلح hapax legomenon يونانيًا ويمكن ترجمته إلى (اللفظ الوارد مرة واحدة).



شكل (١، ٢). توزيع تكرارات الكلمة في المدونة الوطنية البريطانية.

لاحظ أن قانون زيف يمثل قيمًا تقريبية. تختلف التكرارات الفعلية للكلمات في المدونة الحقيقية إلى حد ما عن تلك التي يتنبأ بها هذا النموذج. يترتب على معرفة قانون زيف على لغويات المدونة ما سيأتي. علينا أن نتعلم التقييم النقدي لمقدار ما لدينا من أدلة تؤيد ما ندّعيه. بالنسبة للكلمات الفريدة والكلمات القليلة التكرار، يكون هذا الدليل محدودًا بطبيعة الحال. للإجابة عن بعض الأسئلة البحثية نحتاج إلى مدونات كبيرة جدًا (تُقَدَّر بمليارات بدلًا من ملايين الكلمات) عندما تَرِدُ كلمات نادرة نوعًا ما عدة مرات. نحتاج أيضًا إلى البحث عن أدلة إضافية (قابلة للمقارنة) لسلوك الكلمات من مصادر متعددة (انظر المبحث ٣، ٨ حول التحليل البعدي meta-analysis).

التقرير الإحصائي: التكرارات المطلقة والنسبية

١- ما يجب الإفصاح عنه

ينبغي أن يشمل التقرير الذي نجريه عن تكرار الكلمات على كل من التكرار المطلق (الخام) والتكرار النسبي. يحتاج التكرار النسبي إلى المعيرة على الأساس المناسب الذي يتشابه في حجمه مع المدونة المعنية كلها أو الفرعية فيها.

٢- مثال الإفصاح في التقرير

- حرف الجر of هو ثاني أكثر العناصر تكرارًا في المدونة الوطنية البريطانية (التكرار المطلق له = 3,042,376 والتكرار النسبي له = 30,945.68).
- وردت كلمة corpus في النصوص ٢٠ مرة (بمعدل 3.13 في كل 1000 كلمة).

(٤, ٢) مشكلة حلزون البحر: التشتت

فكّر...

قبل قراءة هذا المبحث، فكر في الأسئلة الآتية:

١- هل تعرف ما المقصود بمشكلة حلزون البحر whelk problem؟

٢- تخنّ كم مرة نستخدم هذه الكلمة؟

نحتاج إلى تقديم مفهوم آخر بعد أن بحثنا موضوع تكرارات الكلمات. هذا المفهوم هو: التشتت الذي يفيد بتوصيف الكلمة أو العبارة توصيفاً كاملاً. يتضح هذا المفهوم وضوحاً أفضل مع ما يُعرف بـ (مشكلة حلزون البحر whelk problem). قدّم هذا المصطلح كيلجاريف Kilgariff (1997)، ويشير إلى التوزيع غير المتكافئ للكلمات في المدونات. تصوّر أن لديك مدونة إنجليزية تحتوي على عينة كبيرة من كتاب عن كائنات حلزونية بحرية صغيرة تشبه القواقع إلى حد ما وتعرف بـ whelks، ستظهر هذه الكلمة عدة مرات في هذا الكتاب على الرغم من عدم شيوعها في اللغة الإنجليزية العامة؛ لأنها خاصة بوعاء أو سجل لغوي محدد جداً. عندما نُكوّن قائمة تكرارات تستند إلى مدونتنا التي تتضمن كتاباً عن الكائنات البحرية قد تظهر كلمة whelk بين العناصر المتكررة إلى حد ما عدة مرات في نص واحد. الذي بان لنا هو أن هذه الكلمة قد ظهرت مرة واحدة في نصوص اللغة الإنجليزية العامة. إذا استندنا إلى تكرار الكلمة وحده فستكون نتائجنا مضللة للغاية، وعليه سنحتاج إلى حساب التشتت.

نخبرنا التشتت dispersion بوجه عام عن توزيع الكلمات أو العبارات في كل المدونة، فأداة التعريف the على سبيل المثال ليست مجرد كلمة شديدة التكرار بل موزّعة توزيعاً متساوياً إلى حد ما في النص، وذلك لأنها كلمة وظيفية لا يمكننا عادة تركيب الجمل دون استخدامها. أما الكلمات الأخرى الخاصة بسياق معين (مثل whelk و hashtag و corpus) فيكون توزيعها أقل تساوياً. من المهم الإشارة إلى أنه ليس للتشتت مقياس واحد، بل مجموعة من المقاييس المختلفة التي يمكن استخدامها للتحقيق في التباين داخل المدونات وإبراز الجوانب المختلفة للتشتت. يعتمد التشتت في المستوى النظري المجرد على فهمنا للمدونات وبنيتها لكونه يصف توزيع الكلمات والعبارات في المدونة أو في أجزاء مختلفة منها. يرتبط قرار اختيار مقياس التشتت الذي ينبغي استخدامه ارتباطاً وثيقاً بالتصميم البحثي لدراستنا وتصميم المدونة المستخدمة. بدلاً من الانخراط في نقاش عام عن أكثر مقاييس التشتت ملاءمةً في البحث المدوني (على سبيل المثال Biber et al. 2016; Gries 2008, 2010) نحتاج إلى فهم خصائص مقاييس التشتت الفردية المتاحة ونلائمها مع أغراض سؤالنا البحثي وفهمنا المحدد للتشتت. سننظر هنا في خمسة مقاييس تشتت مهمة (لإلقاء نظرة عامة على عدد من مقاييس التشتت المختلفة، انظر: Gries 2008). لنقل: لديك مدونة مكونة من مليون كلمة ومقسمة إلى ستة أجزاء بأحجام غير متساوية، وكل جزء يُمثل وعاء أو سجلاً لغوياً مختلفاً (الجدول ٣، ٢). سنستخدم هذه المدونة في شرح مقاييس التشتت المختلفة أدناه.

جدول (٢, ٣). مدونة مكونة من مليون كلمة فعلية.

المدونة بأكملها	الجزء 6	الجزء 5	الجزء 4	الجزء 3	الجزء 2	الجزء 1	
1,000,000	200,000	200,000	200,000	200,000	100,000	100,000	الكلمات الفعلية
50	10	24	0	2	4	10	التكرار المطلق لللمة
5	5	12	0	1	2	10	التكرار النسبي لللمة ك في كل 100 ألف كلمة
نعم	نعم	نعم	لا	نعم	نعم	نعم	هل تشمل على اللمة؟

المدى^(٧) $Range_2(R)$

هو مقياس تشتت أساسي للغاية وبسيط إلى حد ما، ونجربنا بعدد أجزاء المدونة التي ترد فيها كلمة أو عبارة ما بغض النظر عن حجمها. يمكن أن تكون الأجزاء مدونات فرعية قائمة على النوع أو المتكلم أو النصوص. يُعبّر عن المدة اصطلاحياً على النحو الآتي:

المدى₂ = عدد الأجزاء التي وردت فيها الكلمة (أو العبارة) (٢, ٥)

بالعودة إلى السؤال: (هل تشمل على الكلمة؟) في الجدول (٢, ٣) فإن مدى الكلمة هو (5) لأن الكلمة وردت في خمسة أجزاء (1 و 2 و 3 و 5 و 6). ويمكن التعبير عن ذلك على النحو الآتي:

المدى للكلمة = 5 (٢, ٦)

أحياناً يُحسب المدى₂ أيضاً بوصفه نسبة مئوية من إجمالي عدد أجزاء المدونة:

(٧) في علم الإحصاء عموماً: يُستخدم المصطلح (المدى) أيضاً بمعنى مختلف، وذلك للإشارة إلى الفرق بين القيم الكبرى والصغرى في أي مجموعة بيانات، وقد نوقش هذا الاستخدام في البحث (١, ٣) ويشار إليه بالرمز (المدى₁).

$$\text{المدى} \% = \frac{\text{عدد الأجزاء التي وردت فيها الكلمة (أو المركب)}}{\text{إجمالي عدد أجزاء المدونة}} \times 100 \quad (٢, ٧)$$

بالتطبيق على المثال الوارد في الجدول (٢, ٣) ينتج أن:

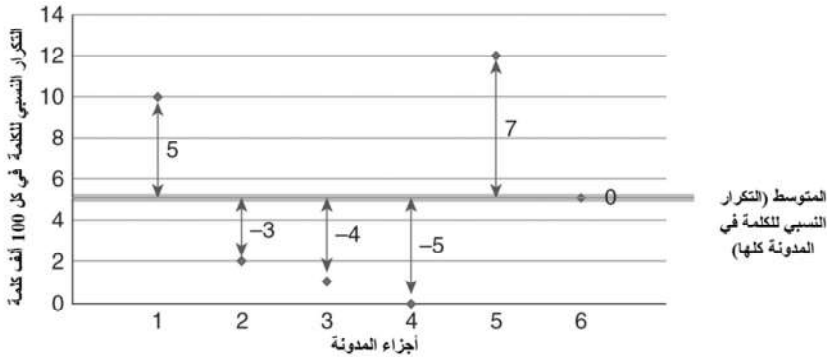
$$\text{المدى} \% = \frac{5}{6} \times 100 = 83.3\% \quad (٢, ٨)$$

يمكننا أن نقول إذن إن المدى_٢ للكلمة في المدونة المكونة من مليون كلمة في الجدول (٢, ٣) يزيد على 80% لأن الناتج 83.3% من أجزاء المدونة تتضمن هذه الكلمة.

لا يُعتبر المدى_٢ قياسًا جيدًا لمقدار التشتت عبر أجزاء المدونة الفردية لأنه يعتمد على قرار نعم/ لا البسيط بشأن وجود أو انعدام كلمة أو عبارة في كل جزء بغض النظر عن التكرارات الفعلية في الأجزاء المختلفة، ولا يراعي حجم الأجزاء. لتوضيح ذلك، تخيّل: كلمة أخرى لها التكرار المطلق نفسه في المدونة مثل الكلمة الأولى، ولنقل: إن التكرار هو (٥٠ مرة)، ولكن لها توزيع تكراري مختلف تمامًا في الأجزاء الستة من المدونة: 46 و 1 و 0 و 1 و 1. عندما نحسب المدى_٢ لتلك الكلمة الأخرى فإننا نحصل على النتيجة نفسها بالنسبة للكلمة الأولى (أي 5 من 6 أو 83.3%) على الرغم من أن الكلمة الأولى منحصرة في جزء واحد فقط من المدونة (الجزء 1) على عكس الأخرى المشتتة تشتتًا متساويًا في المدونة. هذا الانعدام في دقة المدى_٢ بوصفه مقياسًا تشتت هو أكبر قصور فيه. يمكن استخدام المدى_٢ لاستكشاف أولي (بسيط) لبيانات المدونة؛ ولكن لمزيد من التحليلات يفضل استخدام مقاييس التشتت الأكثر دقة.

الانحراف المعياري

مقياس مألوف للتشتت يستخدم بكثرة خارج لغويات المدونة الحاسوبية أيضًا، ويعبر عن مقدار القيم الفردية في مجموعة البيانات عن التكرار النسبي المتوسط (تتباين التكرارات النسبية للكلمة في الأجزاء الفردية للمدونة بالجدول ٢, ٣). يتضح هذا إذا عرضنا قيم التكرار النسبي من الجدول (٢, ٣) في شكل رسومات بيانية.



شكل (٢، ٢). حساب الانحراف المعياري لمدونة مكونة من ستة أجزاء.

نلاحظ في الشكل (٢، ٢) مع استثناء الجزء السادس أن كل قيم التكرارات النسبية للأجزاء الخمسة قد ظهرت على مسافات معينة من المتوسط (التكرار النسبي للكلمة في المدونة). المسافات الفردية من المتوسط هي 5 و-3 و-4 و-5 و7 و0. لاحظ أن المسافات السلبية تشير إلى قيم أصغر من المتوسط. يُحسب الانحراف المعياري بناءً على هذه المسافات من المتوسط. السؤال الجوهرى عند حساب الانحراف المعياري هو: ما مقدار تنوع المتوسط الذي يمكن أن نلاحظه في البيانات؟ يُعبّر عن الانحراف المعياري رياضياً كما هو آت:

$$(٢، ٩) \quad \text{الانحراف المعياري للجموع الغوي} = \sqrt{\frac{\text{مجموع مربعات المسافات من المتوسط}}{\text{مجموع عدد أجزاء المدونة}}}$$

يدل مصطلح الانحراف المعياري SD لوحده على شكل بسيط من أشكال القياس الخاصة به عند قياس كامل العينة، ويختلف في إجراءاته اختلافاً طفيفاً عن أشكال الانحراف المعياري الأخرى كالانحراف المعياري للجموع الغوي SD_p أو SD_p [سيجما]، ويُحسب الانحراف المعياري للجموع الغوي بالنسبة للبيانات الواردة في الجدول (٢، ٣) على النحو الآتي:

$$(٢، ١٠) \quad \begin{aligned} &= SD_p(w) \text{ (الانحراف المعياري للجموع الغوي للكلمة)} \\ SD_p(w) &= \sqrt{\frac{(10-5)^2 + (2-5)^2 + (1-5)^2 + (0-5)^2 + (12-5)^2 + (5-5)^2}{6}} \\ &= \sqrt{\frac{25+9+16+25+49+0}{6}} = 4.55 \end{aligned}$$

قد يبدو تربيع المسافات $(^2)$ ثم أخذ الجذر التربيعي $(\sqrt{})$ للرقم النهائي طريقةً معقدة لحساب التنوع في البيانات، ولكن يكمن سبب هذا الإجراء في كون بعض القيم أعلى من المتوسط؛ ومن ثم تكون المسافة موجبة، في حين أن القيم الأقل من المتوسط تكون مسافتها سالبة، ويجمع المسافات ستلغي القيم الإيجابية والسلبية بعضها بعضًا $(-5 - 3 - 4 + 5 + 7 + 0 = 0)$. إذن: تربيع العدد السليبي يجعل العدد موجبًا، وإجراء التربيع ثم الجذر التربيعي في الانحراف المعياري يُخلصنا من مشكلة السالب.

هناك طريقة أخرى معدلة بعض الشيء لحساب الانحراف المعياري سنستخدمها عندما نتناول الإحصاءات الاستدلالية بدلاً من الإحصاءات الوصفية التي ناقشناها هنا. يسمى هذا الانحراف المعياري للعينة sample standard deviation، ويُحسب بطريقة حساب الانحراف المعياري الموصوف في المعادلة (٩، ٢). يتمثل الاختلاف في أن مجموع المسافات المربعة تُقسم على ناتج [إجمالي عدد أجزاء المدونة - 1] وليس على إجمالي عدد أجزاء المدونة، ويرتبط سبب ذلك بمفهوم (درجات الحرية) الموضحة في المبحث (٣، ٦).

$$\text{الانحراف المعياري للعينة} = \sqrt{\frac{\text{مجموع مربعات المسافات من المتوسط}}{\text{مجموع عدد أجزاء المدونة} - 1}} \quad (٢, ١١)$$

هذا القياس مفيد عندما نريد أن نرى مدى تجانس أو تنوع توزيع الكلمة. ينبغي النظر إلى الانحراف المعياري دائمًا مقترنًا بالمتوسط (التكرار النسبي للكلمة في المدونة بأكملها). في مثالنا بالشكل (٢، ٢) كان مقدار الانحراف المعياري (4.55) كبيرًا مثل المتوسط (5) تقريبًا. يشير هذا إلى قدر كبير من التنوع عندما يظهر العديد من القيم الفردية على مسافة كبيرة من المتوسط.

من الضروري النظر إلى الانحراف المعياري مع المتوسط إلا أنه لا يمكننا استخدام هذا القياس لمقارنة تشتتات الكلمات (أو العبارات) المختلفة التي ترد بتكرارات مختلفة. في هذه الحالة، تكون مقاييس التشتت الأخرى مثل معامل التنوع coefficient of variation أو توزيع جولاند D أو انحراف النسب DP (deviation of proportions) أكثر ملاءمة.

يصف معامل التنوع (CV) مقدار التنوع بالنسبة لمتوسط التكرار النسبي للكلمة أو العبارة في المدونة. كلما زاد تنوع تكرارات الكلمة أو العبارة في الأجزاء الفردية زاد التشتت. ومعادلة معامل التنوع بسيطة جداً، وهي على النحو الآتي:

$$(٢, ١٢) \quad \text{معامل التنوع} = \frac{\sqrt{\frac{\text{الانحراف المعياري}}{\text{المتوسط}}}}$$

بالنسبة للكلمة في المدونة في الجدول (٣, ٢) يكون الحساب على النحو الآتي:

$$(٢, ١٣) \quad \text{معامل التنوع} = \frac{4.55}{5} = 0.91$$

ومعامل التنوع هو مقياس موحد؛ أي إنه يمكن مقارنته عبر كلمات وعبارات مختلفة في مدونة واحدة. وكلما اقترب المعامل من الصفر توزعت الكلمة أو العبارة توزيعاً متساوياً. تعتمد القيمة القصوى لمعامل التنوع على عدد الأجزاء في المدونة، وهي تساوي ناتج (الجذر التربيعي لعدد أجزاء المدونة - 1).

يُضرب معامل التنوع أحياناً على مئة ويُقدّم بوصفه نسبة مئوية للتنوع، غير أن هذا يمثل مشكلة عندما تكون نتيجة معامل التنوع أكبر من (1)، أو عندما يكون الانحراف المعياري أكبر من المتوسط، هو ما يعطي نسبة أعلى من (100). يمكن تحقيق تحويل النسبة بالنظر في أقصى تنوع ممكن في مدونة معينة. نحن نعلم أن الحد الأقصى لقيمة معامل التنوع يعتمد على عدد أجزاء المدونة ويمكن حسابه بالمعادلة التالية: (الجذر التربيعي لعدد أجزاء المدونة - 1). تحول المعادلة التالية معامل التنوع إلى نسبة مئوية وإلى الحد الأقصى الممكن للتنوع الملحوظ.

$$(٢, ١٤) \quad \text{معامل التنوع} \% = 100 \times \frac{\text{معامل التنوع}}{\sqrt{\text{عدد أجزاء المدونة} - 1}}$$

عندما نطبق هذا على الكلمة في المدونة من الجدول (٣, ٢) نحصل على الآتي:

$$(٢, ١٥) \quad \text{معامل التنوع} \% = 100 \times \frac{0.91}{\sqrt{6-1}} = 40.7\%$$

هذا يعني أن تشتت الكلمة كان أقل من 50% من أقصى تنوع ممكن. يمكن الوصول إلى الحد الأقصى لمستوى التنوع إذا وردت الكلمة في جزء واحد فقط من المدونة.

توزيع جولاند Juilland's D

هو مقياس للتشتت يعتمد على معامل التنوع، ويكون بين الصفر والواحد، ويشير (0) إلى توزيع متباين للغاية، ويشير (1) إلى توزيع متساوٍ تمامًا. وُضع توزيع جولاند في الأصل لاستخدامه في قواميس التكرار (Juilland & Chang-Rodriguez 1964; Juilland et al. 1970; Leech et al. 2001; Davies & Gardner 2010)، وتكون نتيجته عكس نتيجة معامل التنوع ومعامل التنوع المئوية، فعندما نخبرنا معامل التنوع عن مقدار التنوع في المدونة (كلما زاد معامل التنوع زاد التنوع في التكرارات)، نخبرنا توزيع جولاند عن تجانس التوزيع (أي: كلما زاد توزيع جولاند ازداد تساوي التوزيع وقل التنوع). تستخدم المعادلة التالية لحساب توزيع جولاند:

$$\text{توزيع جولاند} = \frac{\text{معامل التنوع}}{\sqrt{\text{عدد أجزاء المدونة} - 1}} \quad (٢, ١٦)$$

على سبيل المثال: سيُحسب توزيع جولاند من الجدول (٢, ٣) على النحو الآتي:

$$\text{توزيع جولاند} = \frac{0.91}{\sqrt{1-6}} - 1 = 0.59 \quad (٢, ١٧)$$

تدل هذه القيمة (0.59) على التوزيع المتذبذب uneven distribution وهو أقرب إلى 1 (التوزيع المتساوي تمامًا) منه إلى 0. لاقى توزيع جولاند انتقادًا في الأدبيات (Gries 2008) لأنه يسترجع القيم التي تقع خارج المدى المتوقع (0-1). كان هذا النقد قائمًا على معادلة مختلفة عن تلك المعروضة في هذا الفصل^(٨). أشار بايبر وآخرون (Biber et al ٢٠١٦) إلى مشكلة محتملة أخرى في

(٨) يستخدم غريس (2008) الانحراف المعياري للعينة (المعادلة ٢, ١١) لحساب معامل التنوع لمعادلة توزيع جولاند (المعادلة ٢, ١٦). لتجنب المشكلات المشار إليها في مقالة غريس فإنه ينبغي استخدام الانحراف المعياري البسيط للمجموع اللغوي الوصفي (انظر المعادلة ٢, ٩).

توزيع جولاند، إذ يَبَيَّنُ أن هذا المقياس يعتمد اعتمادًا كبيرًا على عدد أجزاء المدونة؛ فزيادة عدد أجزاء المدونة (على سبيل المثال: 1000) يصبح المدى العملي لتوزيع جولاند مقيدًا للغاية لأن التنوع الفعلي في المدونة كما هو مبين في المعادلة يُقاس على أقصى تنوع نظري ممكن في المدونة بالنظر إلى عدد أجزاء المدونة، فكلما زاد عدد أجزاء المدونة زاد التنوع النظري.

نُوقِش انحراف النسب (DP Deviation of Proportions) في هذا البحث، وهو مقياس اقترحه غريس (٢٠٠٨) بديلًا لتوزيع جولاند، ويقارن التوزيع المتوقع للكلمة أو العبارة في أجزاء مختلفة من المدونة مع التوزيع الفعلي، وتكون قيمته بين 0 و1، إذ يشير (0) إلى توزيع متساوٍ تمامًا ويشير (1) إلى توزيع متباين للغاية. لاحظ أن هذا المقياس هو عكس مقياس توزيع جولاند الذي يشير (0) فيه إلى توزيع متباين للغاية. يحسب انحراف النسب على النحو الآتي:

$$\text{انحراف النسب} = \frac{\text{مجموع القيم المطلقة لـ (النسب الملحوظة - النسب المتوقعة)}}{2} \quad (٢, ١٨)$$

تُحسب النسب المتوقعة بقسمة عدد التكرار المطلق للكلمات الفعلية من كل جزء من المدونة على إجمالي عدد الكلمات الفعلية في المدونة، ولو كان عدد تكرار الكلمة في كل أجزاء المدونة متساويًا فسنجري الخطوة نفسها في حساب النسبة المتوقعة والنسبة الملحوظة أيضًا. تُحسب النسب الملحوظة في كل جزء من المدونة بقسمة عدد التكرار المطلق للكلمة الفعلية من كل جزء على إجمالي عدد تكرار الكلمات الفعلية في المدونة، وذلك لتحديد مقدار ما يسهم به كل جزء من أجزاء المدونة في التكرار الكلي للكلمة الفعلية. وبمقارنة النسب الملحوظة والمتوقعة بقسمة ناتج النسبة المتوقعة على ناتج النسبة الملحوظة من كل جزء من أجزاء المدونة، ثم جمع نواتج كل التقسيم معًا، نحصل على قياس انحراف النسب.

فعلى سبيل التقريب، يُحسب انحراف النسب للقيم في الجدول (٢, ٣) كما هو موضح في الجدول (٢, ٤).

$$\text{انحراف النسب} = \frac{0+0.28+0.2+0.16+0.02+0.1}{2} = 0.38 \quad (٢, ١٩)$$

تشير هذه القيمة (0.38) إلى وجود توزيع متباين، وهي قيمة أقرب إلى (0) من (1)، وفي هذه الحالة يوفر انحراف النسب صورة مشابهة لتوزيع جولاند.

جدول (٤, ٢). حساب توزيع النسب في المدونة المكونة من ستة أجزاء [يشير 100K إلى مئة ألف كلمة، و1M إلى مليون كلمة].

	الجزء 1	الجزء 2	الجزء 3	الجزء 4	الجزء 5	الجزء 6	جميع المدونة
الكلمات الفعلية	100000	100000	200000	200000	200000	200000	1000000
التكرار المطلق للكلمة	10	4	2	0	24	10	50
النسبة المتوقعة	$\frac{100k}{1M} = 0.1$	$\frac{100k}{1M} = 0.1$	$\frac{200k}{1M} = 0.2$	$\frac{200k}{1M} = 0.2$	$\frac{200k}{1M} = 0.2$	$\frac{200k}{1M} = 0.2$	1
النسبة الملحوظة	$\frac{10}{50} = 0.2$	$\frac{4}{50} = 0.08$	$\frac{2}{50} = 0.04$	$\frac{0}{50} = 0$	$\frac{24}{50} = 0.48$	$\frac{10}{50} = 0.2$	1
الفروقات بين النسبتين	0.1	0.02	0.16	0.2	0.28	0	0.76

التقرير الإحصائي: مقاييس التشتت

١ - ما يجب الإفصاح عنه

عند إعداد تقرير عن التكرارات في المدونات ينبغي لنا أيضًا تضمين المعلومات عن التشتت. ينبغي أن يكون الدافع وراء قياس التشتت المُجرى عنه التقرير من ضمن أهداف البحث: ما جانب التشتت اللازم لتمكين من تفسير التكرار تفسيرًا صحيحًا؟ تتضمن الخيارات الرئيسية المدى (م)، ومدى النسبة المئوية (م/٪)، والانحراف المعياري للمجموع اللغوي (SD_p)، والانحراف المعياري للعينة (SD_s)، ومعامل التنوع coefficient of variation، والنسبة المئوية لمعامل التنوع، وتوزيع جولاند، وانحراف النسب.

٢ - مثال الإفصاح في التقرير

• ترد كلمة corpus 773 مرة في المدونة الوطنية البريطانية (6.9 مرة في كل مليون كلمة)، ولكن فقط في 201 نص (المدى. ٪ = 5٪).

- أداة التعريف the هي الكلمة (النوعية) الأكثر تكراراً في مدونة BE06، إذ وردت في النصوص بمتوسط تكرار نسبي 51.64 مرة لكل 1000 (الانحراف المعياري = 17.14).
- تتوزع الكلمات النابية توزعاً متبايناً في مدونة BNC64. فعلى سبيل المثال، ترد كلمة fuck 123 مرة في 14 فقط من أصل 64 عينة متكلمين، وظهرت قيمة انحراف النسب $DP = 0.85$.

(٥، ٢) ما الكلمات المهمة؟ تكرار المتوسط المقلص

فكّر...

قبل قراءة هذا المبحث، فكر في السؤال التالي:

كيف يمكنك معرفة الكلمات الإنجليزية التي ينبغي معرفتها؟

يمكن القول إن الكلمات التي تتكرر في عدد كبير من السياقات مهمة؛ لأنها من المحتمل جداً مصادفتها في مجموعة متنوعة من المواقف التواصلية. على سبيل المثال، بالنسبة لمتعلمي اللغة، من المهم معرفة العناصر المعجمية التي يلزمهم تعلمها أولاً. هذه الكلمات ليست غالباً كلمات ملموسة (كما قد نفترض خطأً) ولكنها مصطلحات مجردة عامة، فمثلاً يُعدّ الاسم الأكثر استخداماً في اللغة الإنجليزية هو (time؛ أي: الوقت) الذي يمكن مصادفته متكرراً في سياقات مختلفة في كل من اللغة المنطوقة والمكتوبة. نحن نتكلم (ونكتب) عن الوقت على الدوام. في الآونة الأخيرة، أجرى بريزينا وجابلاسوفا (Brezina & Gablasova 2015) بحثاً عن أهم كلمات اللغة الإنجليزية استناداً إلى مدونات متنوعة للغة الإنجليزية بلغ إجمالي حجمها أكثر من 12 مليار كلمة فعلية. كان المقياس الأساسي المستخدم في هذه الدراسة هو تكرار المتوسط المقلص (ARF) average reduced frequency. تكرار المتوسط المقلص هو مقياس يجمع بين التكرار والتشتت (Savický & Hlaváčová 2002). تتمثل فكرة تكرار المتوسط المقلص في إنتاج معامل استخدام يقيم أهمية الكلمات في مدونة من جهة تكرارها وتشتتها، وكلما كانت الكلمة أكثر تكراراً ومتوزعة بالتساوي دَلَّ ذلك على أن الكلمة بارزة. من مزايا تكرار المتوسط المقلص أنه لا يعتمد على تقسيم المدونة فعلياً إلى أجزاء مختلفة (مدونات فرعية). لحساب تكرار المتوسط المقلص نحتاج إلى المعلومات الثلاث التالية:

- التكرار المطلق للكلمة

• حجم المدونة (إجمالي عدد الكلمات الفعلية)

• مواضع الكلمة في المدونة

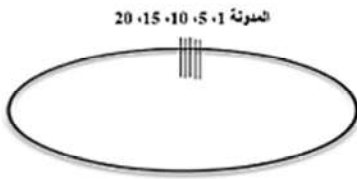
نوقش التكرار المطلق والعدد الإجمالي للكلمات الفعلية في المباحث السابقة (انظر: المبحث ٢, ٢ و ٢, ٣). تُمثِّل مواضع الكلمة في المدونة بأرقام تعبر عن الترتيب الذي تظهر به الكلمة في المدونة؛ ويُحسب هذا آلياً بأدوات تحليل المدونة. تخيل أنك تتصفح المدونة (نصاً تلو الآخر) وتعين كل كلمة فعلية برقم بدءاً من واحد. كلما واجهت الكلمة النوعية التي تبحث عنها دوّن رقم الكلمة الفعلية (أي موضعها في المدونة). استمر في القراءة حتى تصادف الكلمة مرة أخرى، ثم علّم رقم موضعها وهلم جرا. من خلال المواضع الفردية للكلمات بالمدونة، يمكننا بعد ذلك حساب المسافات بين عدد مرات ظهور الكلمة في المدونة. تُستخدم هذه المسافات في معادلة تكرار المتوسط المقلص كما يأتي:

$$\text{تكرار المتوسط المقلص} = \frac{1}{3} \times (\text{صغ}(المسافة_1, \text{ف}) + (\text{صغ}(المسافة_2, \text{ف}) + (\text{صغ}(المسافة_3, \text{ف}) + \dots))$$

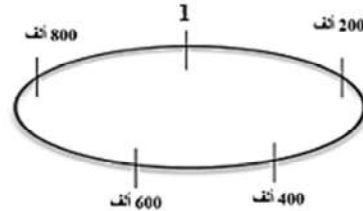
$$\text{تكرار المتوسط المقلص} = \frac{\text{إجمالي الكلمات الفعلية}}{\text{التكرار المطلق}} \times (\text{صغ}(المسافة_1, \text{ف}) + (\text{صغ}(المسافة_2, \text{ف}) + (\text{صغ}(المسافة_3, \text{ف}) + \dots))$$

$$\text{و تُشير ف إلى ناتج: } \frac{\text{إجمالي عدد الكلمات الفعلية}}{\text{التكرار المطلق للكلمة}}$$

يدل (صغ(المسافة_١, ف)) على القيمة الصغرى من قيمتين، الأولى: قيمة المسافة بين ظهور كلمتين، والثانية: قيمة (ف). تُجرى هذه الدالة على جميع المسافات بين ظهور كلمتين في المدونة.



توزيع الكلمة الأولى ك 1



توزيع الكلمة الثانية ك 2

شكل (٢, ٣). توزيع الكلمة الأولى والكلمة الثانية.

على الرغم من أن المعادلة تبدو معقدة فإن فكرة تكرار المتوسط المقلص بسيطة، وهذا لا يحصل فعلياً ولكنه جزء من المعالجة الآلية. نقسم المدونة نظرياً إلى أجزاء، وكل جزء يتكون من ناتج (ف)، وكل ناتج يُعدّ التكرار المطلق للكلمة المعنية، ويعني أن عدد الأجزاء النظرية وطولها (أي: ف) في المعادلة ٢٠, ٢) يعتمدان على تكرار الكلمة في المدونة. نحسب بعد ذلك ببساطة عدد الأجزاء النظرية التي تتضمن فيها الكلمة المعنية. نسمي هذا العدد (التكرار المقلص). الغرض من هذه الخطوات هو تجاهل تكرارات كلمة قريبة من بعضها وتقع داخل الجزء نفسه ويتطلب حسابها مرة واحدة فقط. لجعل هذا الإجراء محكماً نكرر عملية تحديد نقطة بداية ممكنة في المدونة، كأن نفكر في المدونة بوصفها دائرة بدلاً من كونها خطاً، ثم نحسب متوسط جميع قيم التكرار المقلص.

لنوضح هذا من خلال المثال التالي: لدينا مدونة تحتوي على مليون كلمة نبحت فيها عن كلمتين: ك_١ وك_٢. ترد الكلمتان خمس مرات في المدونة، أي بتكرار مطلق قدره (5). ترد الكلمة الأولى (ك_١) في نص واحد فقط في مواضع المدونة 1، و5، و10، و15 و20، في حين أن الكلمة الثانية (ك_٢) موزعة بالتساوي في المواضع 1، و200000، و400000، و600000 و800000. لاحظ أن الرقم في الشكل (٣, ٢) يشير إلى موضع ظهور الكلمة في المدونة.

يُحسب تكرار المتوسط المقلص لـ: ك_١ على النحو الآتي:

$$(٢, ٢١) \quad 200,000 = \frac{1,000,000}{5} \text{ : (ف) } \quad \text{نحدد طول الأجزاء النظرية لـ (ف):}$$

بعد ذلك نحسب المسافات بين التكرار الفردي للكلمة الأولى (ك_١) في المدونة، ونحتاج إلى استخدام مواضع المدونة. لاحظ جيداً كيف تُحسب المسافة بين الظهور الأخير والأول للكلمة ك_١ في المدونة.

$$\bullet \text{ المسافة}_1 = \text{الظهور الأول} + (\text{إجمالي الكلمات الفعلية بالمدونة} - \text{الظهور الأخير}) = 1 + 20 -$$

$$999,981 = (1,000,000$$

$$\bullet \text{ المسافة}_2 = \text{الظهور الثاني} - \text{الظهور الأول؛ أي: } 5 - 1 = 4$$

$$\bullet \text{ المسافة}_3 = \text{الظهور الثالث} - \text{الظهور الثاني؛ أي: } 10 - 5 = 5$$

$$\bullet \text{ المسافة}_4 = \text{الظهور الرابع} - \text{الظهور الثالث؛ أي: } 15 - 10 = 5$$

$$\bullet \text{ المسافة}_5 = \text{الظهور الأخير} - \text{الظهور الرابع؛ أي: } 20 - 15 = 5$$

٢- مثال الإفصاح في التقرير

- يوضح الجدول التالي أربع مواد معجمية في المدونة الوطنية البريطانية مرتبة حسب تكرار المتوسط المقلص والتكرار المطلق

التكرار المطلق	تكرار المتوسط المقلص	المدونة الوطنية البريطانية
6,050,229	3,839,770	أداة التعريف the
4,119,048	2,702,664	الفعل be
3,010,276	1,838,624	حرف الجر of
2,615,148	1,671,566	أداة العطف of

(٦, ٢) التنوع المعجمي: نسبة الكلمات النوعية إلى الكلمات الفعلية

والنسبة الموحدة والنسبة المتوسطة المتحركة بينهما

فكّر... قبل قراءة هذا المبحث، فكر في السؤال التالي: أي من هذين النصين القصيرين أكثر تنوعاً في المعجم؟	
النص أ [المدونة الوطنية البريطانية: KB7]	النص ب [المدونة الوطنية البريطانية: B25]
You want a cup of tea? Yeah I'm gonna put the kettle on. Yeah alright. Yeah make us a cup of coffee Stuart. Well we'll probably have our dinner first then I'll probably do it. تريد كوباً من الشاي؟ نعم سأوقد الغلاية. لا بأس. حسناً اصنع لي فنجاناً من القهوة ستوارت. حسناً، لنفعل ذلك بعد تناول العشاء.	All sciences - physics, agriculture, medicine and even sociology – go beyond the mere solution of immediate problems, whether these problems are of a 'pure' intellectual type, or an 'applied' practical sort. جميع العلوم - الفيزياء والزراعة والطب حتى علم الاجتماع - تتجاوز مجرد حل المشكلات العاجلة، سواء كانت هذه المشكلات من النوع الفكري البحث أم من نوع عملي تطبيقي.

عند النظر في المدونات يمكننا التفكير في كيفية استخدام كلمات (نوعية) مختلفة لتوصيل المعاني. يتكرر بعض الكلمات كثيرًا (خاصة الكلمات الوظيفية) ويتكرر بعضها الآخر بضع مرات فقط. لقياس ما إذا كانت المدونة أو النص يستخدمان نطاقًا واسعًا من المفردات أو نطاقًا محدودًا فقط من العناصر المعجمية التي يتكرر استخدامها بصيغ مختلفة فإنه يمكننا حساب إحصاء التنوع المعجمي (Jarvis, 2013).

أبسط إحصاء للتنوع المعجمي هو نسبة الكلمات النوعية إلى الكلمات الفعلية type/token ratio TTR (انظر المبحث ٢, ٢ للاطلاع على تعريف الكلمة النوعية والكلمة الفعلية). تعبر نسبة الكلمة النوعية إلى الكلمة الفعلية عن نسبة الكلمات النوعية (الصيغ المختلفة للكلمة) بالنسبة إلى نسبة الكلمات الفعلية (كل صيغ الكلمات للكلمة). تتمثل الفكرة في أن العدد الكبير من صيغ الكلمة المختلفة (الكلمات النوعية) بالنسبة إلى عدد كل الكلمات في النص (الكلمات الفعلية) يشير إلى نص أكثر تنوعًا من الناحية المعجمية. تُحسب نسبة الكلمة النوعية/الكلمة الفعلية على النحو الآتي:

$$\text{نسبة الكلمات النوعية إلى الكلمات الفعلية} = \frac{\text{عدد الكلمات النوعية في النص أو المدونة}}{\text{عدد الكلمات الفعلية في النص أو المدونة}} \quad (٢٣, ٢)$$

بالنسبة للنصوص الواردة في مهمة (فَكَّر) تكون نسبة الكلمات النوعية إلى الكلمات الفعلية على النحو الآتي:

$$0.8 = 35 \div 28$$

$$0.93 = 30 \div 28$$

يدل هذا على أن النص ب (النص الأكاديمي) أكثر تنوعًا من الناحية المعجمية من النص أ (خطاب غير رسمي)، وهي مقارنة مقبولة لكون حجم النص بينهما متقاربًا. ينبغي أن نتذكر أن نسبة الكلمات النوعية إلى الكلمات الفعلية حساسة جدًا للنصوص الطويلة إذ تتناقص كلما طال النص، ويُعاد استخدام المزيد من الكلمات وتدويرها مرة أخرى. يمكن استخدام نسبة الكلمات النوعية إلى الكلمات الفعلية البسيطة من المعادلة (٢٣, ٢) فقط لمقارنة النصوص ذات الأحجام المتساوية. لا يمكن الاعتماد على نسبة الكلمات النوعية إلى الكلمات الفعلية، وجعلها غير قابلة للتقلب الذي يكون

نتيجة طول أحجام النصوص (Covington & McFall 1998; Tweedie & Baayen 2010) على الرغم من بعض المزايم من أن بعض المقاييس (مثل: Yule's K، Guiraud's index إلخ.) يمكنها تحقيق ذلك على وجه الدقة. لا توجد صيغة سحرية من شأنها مَعْيَرَةُ نسبة الكلمة النوعية إلى الكلمة الفعلية، وجعلها قابلة للمقارنة عبر النصوص ذات الأحجام المختلفة. ولكن، اقترحت مقاييس مختلفة تعتمد على أخذ قيمة متوسط مجموعة من نسب الكلمات النوعية إلى الكلمات الفعلية من عينات نصية ذات طول قياسي: النسبة الموحدة للكلمات النوعية/الفعلية STTR standardized type/token ratio ونسبة متوسط الكلمة النوعية/الفعلية المتحرك MATTR moving average type/token ratio.

يستخدم سكوت (2004) Scott النسبة الموحدة للكلمات النوعية/الفعلية (STTR) للإشارة إلى مقياس يُعرف أيضًا باسم نسبة متوسط الكلمة الفعلية إلى النوعية القِطعي (MSTTR) mean segmental type/token ratio (Richards 2002 & Malvern). سنستخدم هنا مصطلح: النسبة الموحدة للكلمات النوعية/الفعلية (STTR) لاشتغالها في مجال لغويات المدونة بفضل أدوات وورد سميث WordSmith Tools (الإصدار 0.4) التي طبقت هذا المقياس تحت اسم STTR. يتميز حساب النسبة الموحدة للكلمات النوعية/الفعلية (STTR) ببساطته، فنحن نقسم النص إلى قطع موحدة الحجم (على سبيل المثال 1000 كلمة في كل قطعة)، ونحسب نسب الكلمات النوعية إلى الكلمات الفعلية (TTR) لكل قطعة ثم نأخذ القيمة المتوسطة لنسبة الكلمات النوعية إلى الكلمات الفعلية. ونظرًا لأن معظم النصوص لا تقسم إلى مقاطع ذات حجم موحد تمامًا، يُستبعد الجزء الأخير الأقصر من الحجم القياسي من الحسابات. بالنسبة للنصوص الأقصر ينبغي اختيار الحجم القياسي الأصغر للقطع، وقد أفادت تقارير بأن القطع الصغيرة جدًا (أصغر من 100 كلمة فعلية) تؤدي إلى تشوه النتائج (Richards 2002 & Malvern). تشبه نسبة متوسط الكلمة النوعية/الفعلية المتحرك moving average type/token ratio (MATTR) التي قدمها كوفينغتون ومكفول (Covington & McFall 2010) إلى حد ما نسبة الكلمات النوعية إلى الكلمات الفعلية الموحدة التي تحسب متوسط نسب الكلمات النوعية إلى الكلمات الفعلية في قطع متساوية الحجم لتجنب مشكلات اعتماد نسبة الكلمات النوعية إلى الكلمات الفعلية على طول النص. لكن بدلًا من تقسيم النص إلى قطع متعاقبة غير متداخلة تستخدم نسبة متوسط الكلمات النوعية إلى الكلمات الفعلية المتحرك نافذةً متداخلةً تتحرك بسلاسة عبر النص، لكل قطعة موضع نافذة تُحسب فيها نسبة الكلمات النوعية إلى الكلمات الفعلية

ثم تُحسب القيمة المتوسطة لنسب الكلمات النوعية إلى الكلمات الفعلية التي يُحصل عليها بهذه الطريقة، وعليه تكون نسبة متوسط الكلمات النوعية إلى الكلمات الفعلية المتحرك MATTR مقياساً متيناً وصلباً للتنوع المعجمي من النسبة الموحدة للكلمات النوعية/الفعلية STTR لأنها تأخذ في الحسبان جميع المقاطع الممكنة للنص.

التقرير الإحصائي: نسبة الكلمات النوعية إلى الكلمات الفعلية TTR ونسبتها الموحدة STTR ونسبة المتوسط المتحرك بينهما MATTR

١- ما يجب الإفصاح عنه

من المهم عند إعداد تقارير إحصائية مختلفة عن نسبة الكلمات النوعية إلى الكلمات الفعلية إعداد تقرير عن البارامترات parameters التي تؤثر في النتيجة. ينبغي بيان طول النص في التقرير عند قياس نسبة الكلمات النوعية إلى الكلمات الفعلية، وينبغي بيان حجم القطعة وحجم نافذة النص بعد ذلك عند قياس نسبة الكلمات النوعية إلى الكلمات الفعلية الموحدة ونسبة متوسط الكلمات النوعية إلى الكلمات الفعلية المتحرك.

٢- مثال الإفصاح في التقرير

- استُخدمت نسبة الكلمات النوعية إلى الكلمات الفعلية TTR لمقارنة النصوص لأنها كانت بالطول نفسه (2000 كلمة فعلية). وكانت قيم نسبة الكلمات النوعية إلى الكلمات الفعلية على النحو التالي: 0.36 للنص الأول، و0.33 للنص الثاني، و0.39 للنص الثالث.
- نسبة متوسط الكلمات النوعية إلى الكلمات الفعلية المتحرك MATTR (حجم النافذة: 100) من نص (عيد ميلاد كارول) Christmas Carol للمؤلف تشارلز ديكنز هي 0.67.

(٧، ٢) التطبيق وأمثلة أخرى: هل يتحدث البريطانيون عن الطقس طوال الوقت؟

في هذا البحث، سنتناول مثالاً لمشروع بحثي يطبق الإجراءات الإحصائية التي قدمناها في هذا الفصل. تصوّر أن مهمتنا هي استكشاف موضوعات نموذجية ككلمات المحتوى التي كثيراً ما تُناقش باللغة الإنجليزية البريطانية. يمكن أن يعطينا هذا نظرة ثاقبة للمبادئ العامة في خطاب المجتمع البريطاني. نريد هنا أن نعرف ما إذا كان الطقس weather أحد الموضوعات النموذجية

البارزة للثقافة البريطانية كما توحى الصورة النمطية السائدة. المدونة التي نريد استخدامها هي BE06، وهي مدونة باللغة الإنجليزية البريطانية الحديثة المكتوبة^(٩)، وتحتوي على عينات من ١٥ وعاءً مختلفاً، وتتكون من نحو مليون كلمة، ويتلخص تكوين المدونة في الجدول (٥، ٢).

قبل النظر في البيانات، يمكننا البدء في التخطيط لتحليلنا. نحن نعلم أن الوحدة المعجمية الأكثر تكراراً هي أداة التعريف في اللغة الإنجليزية المكتوبة، وتمثل نحو ٦٪ من جميع الكلمات الفعلية (المبحث ٣، ٢). يمكننا أن نتنبأ ببعض الثقة من أنه سيكون هناك قرابة 60,000 حالة ظهور لأداة التعريف في المدونة. لنفترض أيضاً أننا قررنا عند إجراء ملاحظات مدعومة دعماً صحيحاً عن الثقافة البريطانية الحاجة إلى تحليل الكلمات التي تظهر على الأقل ٣٠ مرة في المدونة لتفادي الادعاءات القائمة على كلمات وألفاظ فريدة ذات تكرار منخفض. اخترنا الحد الأدنى للتكرار: (٣٠)؛ نظراً لأن المدونة مكونة من ١٥ جزءاً، ونريد أن يكون لكل كلمة فرصة للظهور على الأقل مرتين في كل جزء (انظر: المبحث ٤، ١ عن أخذ العينات من المدونة).

جدول (٥، ٢). المدونة BE06.

الحجم (الكلمات الفعلية)	الفترة الزمنية	المحتوى: ١٥ وعاءً من اللغة الإنجليزية المكتوبة
مليون كلمة تقريباً	نحو عام 2006	(١) صحافة: تحقيق صحفي، (٢) صحافة: افتتاحية، (٣) صحافة: تقييمات نقدية، (٤) ديني، (٥) مهارات ومهن وهوايات، (٦) الثقافة الشعبية، (٧) الأدب البحث، والتراجم، والمقالات، (٨) متنوعات (وثائق حكومية، تقارير مؤسسية، تقارير صناعية، جدول جامعي، مجلة دار الصناعة)، (٩) كتابات مستنيرة وعلمية، (١٠) الرواية العامة، (١١) الغموض والرواية البوليسية، (١٢) الخيال العلمي، (١٣) المغامرات والروايات الغريبة، (١٤) القصص الرومانسية والغرامية، (١٥) الفكاهة

(٩) لاحظ أن المدونة المكتوبة تستخدم هنا لتمكيننا من استكشاف التنوع بين مجموعة من الأنواع / السجلات اللغوية. يمكن استكمال هذه الدراسة بالنتائج التي توصلت إليها مدونة اللغة المنطوقة مثل المدونة الوطنية البريطانية المنطوقة Spoken BNC أو المدونة الوطنية البريطانية المنطوقة لعام ٢٠١٤ م (Love et al. 2017).

يمكننا تطبيق قانون زيف Zipf لتحديد العدد التقريبي للعناصر التي تفي بهذا المعيار والتي يمكننا توقع العثور عليها في المدونة. السؤال الذي نطرحه هو: كم عدد الكلمات في قائمة تكرار الكلمات يبلغ عدد كلماتها 60,000 بعد استبعاد كل الكلمات التي تكررت ٢٩ مرة وأقل؟ عندما نحول معادلة قانون زيف للتعبير عن الترتيب نحصل على النتيجة الآتية:

$$(٢, ٢٤) \quad 2,068 \cong \frac{60,000}{29} = \frac{\text{الثابت}}{\text{التكرار المطلق}} \cong \text{الرتبة}$$

هذا يعني أنه يمكننا أن نتوقع أن قرابة 2000 أو 3000 كلمة ستكون داخلية ضمن تحليلنا، ويبدو أن هذا العدد كبير بما فيه الكفاية، ويُتوقع أن تكون هناك كلمات وظيفية مع الكلمات ذات المحتوى. لنكمل خطة تصميم هذا البحث التحليلي. في الخطوة التالية حري بنا بأن نقرر ما نريد تحليله بالفعل، فقد رأينا أن هناك تعريفات مختلفة لـ (الكلمة) (المبحث ٢, ٢) وعليه نحتاج إلى اتخاذ قرار بشأن التعريف الذي سنستخدمه. خياراتنا هي الكلمة النوعية أو المادة المعجمية أو اللكسيم. بعد النظر في الإيجابيات والسلبيات، لنفترض أننا قررنا التعامل مع المواد المعجمية في بحثنا. عادةً ما تكون المواد المعجمية اختيارًا جيدًا إذا كنا مهتمين بالثقافة أو الخطاب؛ نظرًا لأن الاختلافات التي تُخفى هي الاختلافات التصريفية التي لا ترتبط عمومًا بهذا النوع من البحث. تؤكد قائمة التكرار المستندة إلى المواد المعجمية افتراضنا المبدئي (استنادًا إلى قانون زيف) عن عدد الكلمات بواقع ٣٠ مرة للتكرار وصاعدًا: يوجد بالمدونة 3,196 من هذه المواد المعجمية، وبناء عليه فإن العدد الفعلي للمواد المعجمية أكبر من المتوقع، ويدلّ على أن قانون زيف لا يوفر سوى تقدير تقريبي لتوزيعات تكرار الكلمات.

جدول (٦, ٢). المواد المعجمية المتعلقة بالطقس في مدونة BE06.

الوحدات المعجمية المرتبطة بالطقس	التكرار المطلق	المدى	توزيع جولاند
سحابة cloud	40	8	0.7
برد cold	120	11	0.8
فيضان flood	54	10	0.2
حرارة heat	104	12	0.7

تابع جدول (٦, ٢).

الوحدات المعجمية المرتبطة بالطقس	التكرار المطلق	المدى	توزيع جولاند
hot حار	108	12	0.8
ice ثلج	63	13	0.7
rain مطر	37	12	0.8
storm عاصفة	37	12	0.8
sun شمس	71	9	0.7
temperature حرارة	66	11	0.7
warm دافئ	90	13	0.8
weather طقس	49	10	0.7
wind رياح	78	15	0.8

إذا نظرنا الآن عن كُتب إلى قائمة المواد المعجمية، نلاحظ أن من بين 3,196 مادة معجمية مما ينطبق عليها معيارنا لإدراجها في التحليل ثلاث عشرة وحدة معجمية متعلقة بالطقس. كان ينبغي إجراء هذا التحليل يدوياً بتصفح قائمة الوحدة المعجمية: cloud و cold و flood و heat و hot و ice و rain و storm و sun و temperature و warm و weather و wind. يتضح من الجدول (٦, ٢) أيضاً ظهور هذه المواد المعجمية في المدونة بتكرارات مطلقة تُراوح من 37 إلى 120 مرة. كما نعلم، ينبغي دائماً النظر في معلومات التكرار من عدسة التشتت، وبناء على ذلك يوفر الجدول (٦, ٢) أيضاً مقياسين للتشتت لـ: ١٣ وحدة معجمية متعلقة بالطقس؛ وهما: المدى وتوزيع جولاند.

نرى أن معظم الكلمات موزعة في المدونة بالتساوي إلى حد ما إذ تُراوح قيم المدى من 8 إلى 15 وقيم توزيع جولاند هي 0.7 أو 0.8. هناك استثناء واحد لكلمة flood (فيضان)، فعلى الرغم من ظهورها في معظم أجزاء المدونة (المدى 10) فتوزيع جولاند لها منخفض للغاية (0.2). يدل هذا على تفاوت التوزيع واختلافه في أوعية المدونة الخمسة عشر. يؤكد الفحص الدقيق لهذه الحالة على ظهور كلمة flood ظهوراً واضحاً في وعاء: (الوثائق الرسمية)، ومعظم مرات ظهورها كان من تقرير برلماني واحد عن الفيضانات، وبواقع 38 مرة من 54 مرة.

للإجابة عن سؤال البحث عن الموضوعات البارزة في الخطاب العام البريطاني نحتاج إلى النظر إلى مصطلحات الطقس في سياق الكلمات الأخرى في المدونة. نحتاج لذلك إلى فرز الكلمات وفقاً لبروزها في اللغة الإنجليزية البريطانية المكتوبة، ومن ثمّ النظر في وضع مصطلحات الطقس في هذه القائمة المرتبة. يمكننا حساب تكرار المتوسط المقلص الذي يجمع بين التكرار والتشتت، وترتيب المواد المعجمية وفقاً لقيم تكرار المتوسط المقلص الخاصة بها. يوضح الجدول (٧، ٢) الكلمات المتعلقة بالطقس والبالغ عددها ثلاث عشرة كلمة، ورُتبت القائمة على تكرار المتوسط المقلص في قائمة الكلمات التي تتضمن 3,196 كلمة.

جدول (٧، ٢). رُتب المواد المعجمية المتعلقة بالطقس في مدونة BE06

الوحدات المعجمية المتعلقة بالطقس	تكرار المتوسط المقلص	الرتبة القائمة على تكرار المتوسط المقلص
hot حار	53.8	835
cold بارد	48.3	940
warm دافئ	45	1,005
heat حرارة	38.2	1,195
wind رياح	34.6	1,333
sun شمس	30.1	1,527
weather طقس	24.2	1,867
ice ثلج	23.1	1,941
temperature حرارة	21.7	2,048
storm عاصفة	20.8	2,128
rain مطر	20.2	2,166
cloud سحابة	14.8	2,767
flood فيضان	14.1	2,888

يمكننا وفقاً لتكرار المتوسط المقلص ملاحظة أن كلمتين متعلقين بالطقس تنتميان إلى قائمة أول ألف كلمة، وتنتمي ست كلمات إلى قائمة ثاني ألف كلمة، وتنتمي بقية الكلمات الخمس إلى

قائمة الألف الثالث من العناصر الأكثر تكرارًا. يمكننا استنتاج أهمية الطقس بوصفه موضوعًا مُثَّل تمثيلاً بارزًا في الخطاب البريطاني رغم أنه ليس الموضوع الأكثر مناقشة على أيِّ حال. راجع موقع الويب المصاحب للتعرف على أكثر الموضوعات التي نوقشت على نطاق واسع في الخطاب البريطاني المكتوب.

(٨, ٢) تمارين

١- انظر إلى الجمل (أ) و(ب) و(ج) و(د) واحسب عدد الكلمات الفعلية، والكلمات النوعية، والمواد المعجمية (lemmas)، واللكسيات في كل منها.

- The City is braced for far worse figures to come in the coming months, unless the Government recovery package produces a startling turn round in optimism.
- Of 354 fifth- and sixth-formers who left Sharon's school in the summer of 1981 forty had found real jobs by 18 November, four of these having entered military service.
- Erm erm erm but, yeah and people er have great areas of that taken.
- Homonyms are headwords to different entries that are spelt in the same way, e.g. bow (the weapon), bow (the action), bow (the verb expressing the action).

(أ) تستعد المدينة للحصول على أرقام أسوأ بكثير في الأشهر المقبلة، ما لم تقدم حزمة الإنعاش الحكومية منعطفًا مذهلاً في التفاؤل (المصدر: BNC وCEN).

(ب) من بين ٣٥٤ من طلاب الصف الخامس والسادس الذين تركوا مدرسة شارون في صيف عام ١٩٨١، حصل ٤٠ منهم على وظائف حقيقية بحلول ١٨ نوفمبر، والتحق أربعة منهم بالخدمة العسكرية. (المصدر: BNC وJUR).

(ج) اممم لكن، نعم والناس آآآ قد أخذوا مناطق كبيرة من ذلك. (المصدر: BNC وKC3).

(د) المشتركات اللفظية هي كلمات رئيسية لإدخالات مختلفة مكتوبة بالطريقة نفسها، على سبيل المثال bow (السلاح)، bow (العمل)، وbow (الفعل الذي يعبر عن العمل)؛ (المصدر: BNC وEAT).

٢- استخدم حاسبة الكلمات Word Calculator على الإنترنت لمقارنة نتائجك من التمرين الأول عن عدد الكلمات الفعلية والنوعية والمواد المعجمية التي تولدت تلقائيًا. هل حصلت على النتائج نفسها؟ إن كانت الإجابة لا، فهل يمكنك تفسير الاختلافات؟

٣- استخدم حاسبة الكلمات على الإنترنت لمقارنة نصوص مختلفة من الإنترنت. احسب الكثافة المعجمية lexical diversity باستخدام المقاييس الثلاثة التي نوقشت في هذا الفصل: نسبة الكلمات النوعية إلى الكلمات الفعلية TTR والنسبة الموحدة للكلمات النوعية/ الفعلية STTR ونسبة متوسط الكلمات النوعية إلى الكلمات الفعلية المتحرك MATTR. قارن النتائج وفكر في أي المقاييس سيكون الأنسب استخدامًا مع النص. ما أسباب اختيارك هذا المقياس؟

٤- احسب التكرارات النسبية للعناصر التالية. اختر أساسًا مناسبًا للمعيرة في كل حالة.

(أ) كلمة: muggle

التكرار المطلق: 2

حجم المدونة: 100,000 كلمة

(ب) الكلمة: intriguingly

التكرار المطلق: 3035

حجم المدونة: 11,191,606,036 كلمة

(ج) الكلمة: worse

التكرار المطلق: 50

حجم المدونة: 1,007,299 كلمة

٥- انظر إلى قائمة التكرار أدناه. يظهر فيها 10 كلمات من المدونة الوطنية البريطانية مع رُتبها. استخدم قانون زيف للتنبؤ بالتكرار المطلق للعناصر المعروضة في الجدول.

الرتبة	الكلمة	التكرار المطلق
1	the	6,041,234
2	of	
3	and	
4	to	
5	a	
10	was	
50	so	
100	way	
1,000	limited	
10,000	conveniently	

٦- قارن نتائجك من السؤال الخامس بالتكرارات الفعلية الواردة في مبحث (الإجابات) بنتائج موقع الويب المصاحب. ما مدى إصابة توقع قانون زيف للتكرارات؟

٧- انظر إلى التكرارات المطلقة لأربع كلمات معينة في مجالات الأوعية المحددة من المدونة الوطنية البريطانية (الجدول ٨, ٢). تتوفر نسخة إلكترونية من هذا الجدول في الموقع الشبكي المرافق على الويب.

احسب لكل كلمة ما يأتي:

(أ) المدى

(ب) الانحراف المعياري

(ج) معامل التنوع

(د) توزيع جولاند

(هـ) انحراف النسب

جدول (٨, ٢). المدونة الوطنية البريطانية: توزيع أربع كلمات مختارة.

التكرار المطلق لكلمة chance	التكرار المطلق لكلمة theory	التكرار المطلق لكلمة smile	التكرار المطلق لكلمة some	إجمالي الكلمات الفعلية	مجال المدونة الوطنية البريطانية
2,645	347	5,498	24,616	16,143,913	القصص والشعر
2,589	266	304	10,520	9,412,174	الصحف
2,191	3,977	385	43,161	24,178,674	الشعر غير الأكاديمي والتراجم
923	6,588	58	30,297	15,778,028	الشعر الأكاديمي
3,323	1,268	488	37,867	22,390,782	مواد مكتوبة أخرى
1,138	363	112	20,589	10,409,858	الخطابات
12,809	12,809	6,848	167,050	98,313,429	المدونة كلها

٨- استخدم حاسبة التشتت dispersion calculator للتحقق من نتائجك في التمرين السابع.

٩- احسب تكرار المتوسط المقلص بالكلمات التالية في مدونة BE06 (985,628 كلمة فعلية)

- (أ) frigid: التكرار المطلق مرتان، رقم موقع الورود في المدونة: 840,797 ثم 848,280
- (ب) chemistry: التكرار المطلق سبع مرات، مواقع الورود في المدونة: 160,129 و 589,607 و 594,834 و 596,351 و 611,214 و 948,612 و 950,458
- (ج) porn: التكرار المطلق أربع عشرة مرة، مواقع الورود في المدونة: 16,602 و 16,792 و 28,191 و 49,606 و 161,929 و 170,396 و 268,155 و 497,891 و 497,916 و 498,146 و 498,205 و 498,216 و 498,246 و 498,361.
- ١٠- استخدم حاسبة تكرار المتوسط المقلص ARF calculator لمقارنة التكرار المطلق وتكرار المتوسط المقلص لكلمات مختلفة في نصوص من اختيارك من الإنترنت.

تذكر الآتي

- هناك مفاهيم مختلفة لـ (الكلمة): الكلمة الفعلية، والكلمة النوعية، والمادة المعجمية، واللكسيم.
- يصف قانون زيف توزيع الكلمات في المدونات وتكراراتها السريعة التناقص.
- لوصف كلمة في مدونة نحتاج إلى توفير كل من تكرار الكلمة وتشتتها.
- تُعدّ مقاييس التشتت المختلفة (المدى، والانحراف المعياري، ومعامل التنوع، ومعامل التنوع %، وتوزيع جولاند، وانحراف النسب) ملائمة في المواقف المختلفة.
- تكرار المتوسط المقلص (ARF) هو مقياس يجمع بين كل من التكرار والتشتت. يمكن استخدامه مع المدونات التي لا تنقسم إلى أجزاء مختلفة (مدونات فرعية).
- نسبة الكلمات النوعية إلى الكلمات الفعلية TTR هو مقياس للتنوع المعجمي، وهو قياس حساس لطول النص.
- النسبة الموحدة للكلمات النوعية/الفعلية STTR ونسبة متوسط الكلمات النوعية إلى الكلمات الفعلية المتحرك MATTR، وهما مقياسان بديلان للتنوع المعجمي، ويمكن استخدامهما مع النصوص المختلفة من ناحية الطول والحجم.

قراءات إضافية

- Baroni, M. (2009). Distributions in text. In A. Lüdeling & M. Kytö (eds.), *Corpus linguistics: an international handbook*, vol. 2, pp. 803-21. Berlin: Mouton de Gruyter.
- Covington, M. A. & McFall, J. D. (2010). Cutting the Gordian knot: the moving-average type-token ratio (MATTR). *Journal of Quantitative Linguistics*, 17(2), 94-100.
- Gardner, D. (2007). Validating the construct of word in applied corpus-based vocabulary research: a critical survey. *Applied Linguistics*, 28(2), 241-65.
- Gries, S. Th. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*, 13(4), 403-37.
- (2009). Dispersions and adjusted frequencies in corpora: further explorations. *Corpus linguistic applications: current studies, new directions*, Amsterdam: Rodopi.
- Hlaváčová, J. (2006). New approach to frequency dictionaries - Czech example. 5th edition of the International Conference on Language Resources and Evaluation, Genoa, 22-28 May. www.lrec-conf.org/proceedings/lrec2006/pdf/rll.pdf (accessed 22/6/2014).
- Savický, P. & Hlaváčová, J. (2002). Measures of word commonness. *Journal of Quantitative Linguistics*, 9(3), 215-31.
- Tweedie, F. & Baayen, R. H. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities*, 32, 323-52.

أدوات لانكستر الإحصائية

- ١ - يمكن استنساخ جميع التحليلات المبينة في هذا الفصل باستخدام أدوات لانكستر الإحصائية. تشمل الأدوات المتاحة لهذا الفصل:
- حاسبة الكلمات.
 - أداة قائمة الكلمات.
 - حاسبة التشتت.
 - حاسبة تكرار المتوسط المقلص.
- ٢ - يقدم الموقع أيضًا موادًا إضافية للطلاب والمعلمين.

الفصل الثالث

علم الدلالة والخطاب: النصاحبات اللفظية، والكلمات المفتاحية، وموثوقية الترميز اليدوي

(١، ٣) ما الذي يتناوله هذا الفصل؟

نستكشف هنا موضوع معاني الكلمات بعد أن تناولنا موضوع الكلمات. والبحث في معاني الكلمات وسياقاتها مهم للتحليلات اللغوية والاجتماعية. للموضوعات في هذا الفصل: النصاحبات اللفظية collocations، والكلمات المفتاحية keywords، وموثوقية الترميز اليدوي لأسطر الكشف السياقي أثر رئيس في دراسة علم الدلالة (معاني الكلمات) وفي تحليل الخطاب. يبدأ هذا الفصل بفرضية يسيرة تتمثل في أن أفضل طريقة لكشف معاني الكلمات تكون بتحليل الأنماط اللغوية المتكررة في اللغة. تساعدنا تقنيات مثل الكلمات المفتاحية keywords على لفت الانتباه إلى الكلمات المميزة في النصوص، ويمكن الكشف عنها باستخدام آلية النصاحب اللغوي وتوارد الكلمات، أو باستخدام آلية الكشف السياقي concordancing للنصوص واستخراج أمثلة الكلمات المميزة وفقاً للسياق. سنجيب في هذا الفصل عن الأسئلة الخمسة الآتية:

- ١- كيف يمكننا تحديد النصاحب اللفظي collocation؟ (المبحث ٢، ٣).
- ٢- ما المقصود بشبكات النصاحب اللفظي collocation networks؟ (المبحث ٣، ٣).
- ٣- كيف يمكننا تحديد الكلمات المفتاحية وكلمات الإيقاف lockwords؟ (المبحث ٤، ٣).
- ٤- كيف يمكن جعل الترميز اليدوي لأسطر الكشف اللغوي concordance أكثر موثوقية؟ (المبحث ٥، ٣).
- ٥- كيف يمكن استخدام كل هذه الآليات في البحث اللغوي والاجتماعي؟ (المبحث ٦، ٣).

(٢, ٣) التصاحبات اللفظية ومقاييس التصاحب

فكّر...

قبل قراءة هذا المبحث، فكر في الأسئلة التالية:

١ - ما التصاحبات التي تتبادر إلى ذهنك عندما ترى كلمة love؟

٢ - لماذا تعتقد أن للكلمة هذه المصاحبات؟

من المعلوم في لغويات المدونة الحاسوبية أن الكلمات ترد في مجموعات نسميها متصاحبات. منذ أكثر من خمسين عامًا، اقترح فيرث Firth (١٩٥٧: ٦) أنه ينبغي لنا البحث في (مُصاحبات الكلمات) التي أصبحت منذ ذلك الحين التعريف غير الرسمي للعلاقة التصاحبية (Gries 2013b; Brezina et al. 2015). المتصاحبات اللفظية هي مجموعات من الكلمات التي تتوارد معًا في النصوص والمدونات. يمكن أن تعتمد المتصاحبات اللفظية إمّا على التكرار وحده وإمّا على مقياس إحصائي يُسمّى قياس التصاحب كما هو الأكثر شيوعًا. قياسات التصاحب أو التجاور collocation/association measures هي مقاييس إحصائية تحسب قوة التصاحب بين الكلمات بناءً على جوانب مختلفة من علاقة التوارد. هناك العديد من قياسات التصاحب المختلفة، ويُنتج كل منها قائمةً مختلفة نوعًا ما من المتصاحبات اللفظية (Evert 2008; Gablasova et al. 2017b)، وليس هناك مقياس واحد يناسب جميع الأغراض وجميع أسئلة البحث. بناءً على ذلك فإننا بحاجة إلى فهم كيفية عمل قياسات التصاحب من أجل اختيار أفضل ما يسلط الضوء على جوانب العلاقة التصاحبية المعنية بالتحليل.

يوضح المثال التالي كيفية تحديد التصاحبات اللفظية تطبيقًا:

My love is like a red, red rose that's newly sprung in June: My love is like the melody that's sweetly played in tune. As fair art thou, my bonnie lass, so deep in love am I: And I will love thee still, my dear, till a' the seas gang dry, Till a' the seas gang dry, my dear, and the rocks melt wi' the sun: And I will love thee still, my dear, while the sands o' life shall run. And fare thee weel, my only love, and fare thee weel a while! And I will come again, my love, thou' it were ten thousand mile.

[الترجمة]: حبيتي مثل وردة حمراء، حمراء طلعت يانعة في يونيو. حبيتي مثل اللحن المعزوف متناغمًا بعدوبة. كفنّ جميل أنتِ، يا فتاتي الفاتنة، غارق في الحب أنا، ولا أفتأ أحبك، يا غاليتي، حتى تحفّ البحار. حتى تحفّ البحار، يا غاليتي، وتميع الحجارة في الشمس. ولا

أفتأ أحبك، يا غاليتي، ما انسابت لحظات العمر. والوداع، يا حبي الوحيد، الوداع إلى حين!
ولسوف أرجع من جديد، يا حبيبتي، ولو كان بيننا عشرة آلاف ميل.

(روبرت بيرنز، "A Red, Red Rose")

المثال أعلاه مأخوذ من نسخة إنجليزية من قصيدة بيرنز الشهيرة (مكتوبة أصلاً بالأسكتلندية) A Red, Red Rose. حلّقنا حول كل تكرارات كلمة love مع ما صاحبها من كلمة واحدة من اليسار واليمين، وعرضنا القصيدة دون فواصل سطرية لإنشاء فقرة واحدة من النص المتصل.

لنفترض أننا مهتمون باستخدام كلمة love في القصيدة، وتُسمى هذه الكلمة المعنية (عقدة). العقدة node هي الكلمة التي نريد البحث عنها وتحليلها، والكلمات عن العقدة هي كلمات مرشحة لتكون مُصاحبات لفظية. المُصاحبات اللفظية collocates هي الكلمات التي تتوارد مع العقدة في مدى span محدد عن العقدة يُسمى نافذة التصاحب اللفظي. يمكننا أن نُمثل هذا بـ (حقل مغناطيسي) عن العقدة التي تجذب مُصاحبات لفظية معينة كما يجذب المغناطيس الأجسام المعدنية. في هذه الحالة، فإن المدى (نافذة التصاحب اللفظي) هو كلمة واحدة إلى اليسار وكلمة واحدة إلى اليمين، ويُختصر في بعض الأحيان بـ: ايس، وايم. بالعودة إلى القصيدة السابقة يمكننا أن نلاحظ الكلمات التي تتوارد مع love من ضمن نافذة التصاحب اللفظي (ايس وايم) وهي: my (ثلاث مرات)، is (مرتين)، in (مرة واحدة)، am (مرة واحدة)، thee (مرتين)، only (مرتين)، and (مرة واحدة)، only (مرة واحدة) thou (مرة واحدة). لاحظ أنه في كل حالة بيناً تكرار التوارد بين قوسين، ونسمي هذه القيم تكرار التصاحب الملحوظ observed frequency of collocation. دعونا نتناول الكلمة الأكثر توارداً مع love في مثالنا، ألا وهي my، ثم نسأل: هل my حقاً مُصاحبة حقيقية لكلمة love في القصيدة؟ بعبارة أخرى، هل ترتبط my ارتباطاً قوياً حقاً مع love؟ لمعرفة ذلك نحتاج إلى إيجاد طريقة لتقييم التكرار الملحوظ، وذلك بثلاثة خيارات أساسية:

١- بدون خط أساس baseline (أي: دون الاعتماد على مجموعة قياس أو قياسات محددة): نقارن التكرارات الملحوظة لكل الكلمات الفردية التي تتوارد مع العقدة ونضع قائمة ترتيبية.

٢- خط أساس التوارد العشوائي (نموذج القرعة): نقارن التكرارات الملحوظة بالتكرارات المتوقعة بالصدفة فقط، ونقيّم قوة التصاحب باستخدام معادلة رياضية تركز على جانب معين من جوانب العلاقة التصاحبية.

٣- خط أساس تنافس الكلمات: من التوارد العشوائي نستخدم نوعًا مختلفًا من خط الأساس يُدمج في المعادلة لإبراز جانب معين من جوانب العلاقة التصاحبية.

لا يتضمن الخيار الأول (الأبسط) أي حساب إحصائي. نحن ننتج فقط قائمة ترتيبية من الكلمات التي تتوارد مع العقدة بناءً على التكرار كما في مثالنا السابق: my (3)، is (2)، thee (2)، will (2)، إلخ، والكلمات التي في أعلى القائمة هي أقوى المصاحبات حسب عدد تكرارها. يتمثل عيب هذا النهج في أن أبرز المصاحبات تكون عادةً كلمات وظيفية تتوارد مع العقدة ومع غيرها لكثرة تكرّرها في كل المدونة، لذلك فإن المصاحبات اللفظية القائمة على التكرار عامةً نوعًا ما، ويمكننا توقع مجموعة مماثلة من المصاحبات لأي عقدة تقريبًا تكون محدودة الفائدة.

يتضمن الخيار الثاني مقارنة بخط أساس التوارد العشوائي. فنسأل ما إذا كان من الممكن أن تتكرر مجموعة الكلمات المعنية (مثل my و love) بمحض الصدفة. تأمل ما يأتي:

١- تحتوي القصيدة على 107 كلمات فعلية (انظر المبحث ٢, ٢). للاطلاع على تعريف الكلمة (الفعلية).

٢- تظهر كلمة love 7 مرات في القصيدة كلها.

٣- تظهر كلمة my 8 مرات في القصيدة كلها.

٤- تظهر كلمة my 3 مرات مقترنة بكلمة love و 4 مرات مقترنة بكلمات أخرى.

تحليل أيضًا أنه لا يوجد ارتباط بين الكلمات الموجودة في القصيدة، والكلمات التي تظهر عشوائيًا في النص. يتضح ذلك في المثال الوارد أدناه، والذي يعرض كلمات قصيدة بيرنز بترتيب عشوائي:

fare art And like red, sweetly in love love, And gang wi' played like dear, life shall rocks sprung the Till deep my my And still, weel, again, ten the the while! is till And As I: a' only come were sands sun: dry, and gang it a' the still, My thee will in my bonnie My red is a run. my love thee thou, melt the seas and thou' I the I lass, I melody thee a my am rose love dear, that's love newly love fare love, will o' so dry. fair thee will that's in while June: my seas tune. mile. thousand weel dear,

[الترجمة]: كالوردة الحمراء أنت يا حبيبتى لما تبدّت تحفةً، يا حلوتي في الحب، فالحياة تذوب وقد دبّت في داخلي لمرات ولم تزل كذلك في أعماقي يا ربّتي. لأعشقنك وإن غدت شمس الرمال جافة وظلت، فأنت يا حسناء ما زلت أنت في حسني وردةً تهرب، يا حسني الجميل، يا حبيبتى ذبنا في البحار وصرت أنا في الصمود، يا ربتي ظللتُ لحناً بديعاً كوردة الصباح، هو الحب، هو الحب الجديد المتجدد الجفيف، لأكن بخير في شهر يونيو، فبحاري ترقص في الوسيح ميلاً وألفاً يا ربّتي

كم مرة تتوقع أن تتوارد كلمتا my و love بمحض الصدفة؟ في المثال العشوائي أعلاه، تظهر my love مرة واحدة. في الواقع، إذا أجرينا المحاكاة العشوائية عدة مرات سنحصل على رقم متوسط قريب من واحد. نسمي هذه العملية تحديد خط أساس التوارد العشوائي ونسمى القيمة الناتجة تكرار التصاحب المتوقع expected frequency of collocation. لا ينبغي تحديد تكرار التصاحب المتوقع تجريبياً، ولكن يمكن حسابه على النحو الآتي:

$$\text{تكرار التصاحب المتوقع} = \frac{\text{تكرار العقدة} \times \text{تكرار المصاحبة اللفظية}}{\text{عدد الكلمات الفعلية في النص أو المدونة}} \quad (١, ٣)$$

عند التفكير في امتداد نافذة التصاحب التي تزيد على متصاحب واحد، يمكن تطبيق تصحيح لتفسير وجود فرصة أكبر للكلمات التي تظهر عشوائياً مع العقدة. يُحسب تكرار التصاحب المتوقع المصحح على النحو الآتي:

$$\text{تكرار التصاحب المتوقع} = \frac{\text{تكرار العقدة} \times \text{تكرار المصاحبة} \times \text{حجم النافذة}}{\text{عدد الكلمات الفعلية في النص أو المدونة}} \quad (٢, ٣)$$

في مثالنا السابق يُحسب تكرار التصاحب المتوقع لكلمتي my و love في المدى 1 يس-1 يم كما يأتي:

$$\text{تكرار التصاحب المتوقع (my و love: المصححة)} = \frac{(1+1) \times 8 \times 7}{107} = 1.05 \quad (3, 3)$$

نلاحظ أن تكرار التصاحب الملحوظ للكلمتين (3 مرات) أكبر من التكرار المتوقع (1.05). لمقارنة الفرق بين القيمتين، يمكن تطبيق مقاييس تصاحبية مختلفة (الجدول ٣, ٣). تستند جميع مقاييس التصاحب التي تتضمن التكرار المتوقع في المعادلة (مت^{١١}) إلى خط أساس التوارد العشوائي. العيب المحتمل لهذه المقاييس هو أنها تفترض نموذجاً معيناً للغة (نموذج القرعة) محفوفاً بالمشكلات (Stubbs 2001: 73-4). يُشبّه هذا النموذج المدونة بصندوق فيه كلمات مكتوبة لنا على بطاقات صغيرة منفصلة، ويُهرّ هذا الصندوق بالكامل للحصول على المنطلق (أو الخط الأساس) baseline، ومن ثمّ فالنموذج ينطوي على قدر كبير من السهولة أمام واقع اللغة الذي يُعدّ مرتباً ومنظماً.

ولتجنب نموذج (القرعة) اللغوي المحفوف بالإشكالات، تعمل بعض مقاييس التصاحب على نوع مختلف من المنطلق لا يكون في معادلتها الـ (مت^{١١}). ينبغي فهم تلك الأنواع المختلفة من المنطلق وفهم معادلاتها في قياس التصاحب خطوة خطوة. لنتمكن من فهم المعادلات نحتاج إلى النظر في المصطلحات التي تنطوي عليها هذه المعادلات. أفضل وسيلة هي استعمال الجدولة المتقاطعة (إظهار جميع التصاحبات المحتملة تقاطعياً لتوارد الكلمة). يعرض الجدول (٣, ١) التكرارات الملحوظة (مل^{١١}): تكرارات التوارد التي يمكننا ملاحظتها في المدونات)، في حين يحسب الجدول (٣, ٢) التكرارات المتوقعة، ويُعدّ الجدول ذا صلة فقط بالمقاييس التي تستخدم منطلق التوارد العشوائي^(١).

جدول (٣, ١). التكرارات الملحوظة.

ظهور المُصاحبة (affair)	غياب المُصاحبة	الإجمالي	
ظهور العقدة (love)	مل ^{١١}	ص ^١ × حجم النافذة	
غياب العقدة	مل ^{٢١}	ص ^٢	
الإجمالي	ع ^١	ع ^٢	ن

(١) التدوين في الجداول المتقاطعة قائم على إيفرت (Evert 2008).

في الجدولة المتقاطعة يرمز (مل) إلى التكرار الملحوظ، أما (مت) فيرمز إلى التكرار المتوقع، ويرمز (ع) إلى إجمالي العمود ويرمز (ص) إلى إجمالي الصف. تشير الأرقام بعد (مل) و(مت) إلى الصف والعمود ذي الصلة. على سبيل المثال، يشير (مل₁₂) إلى التكرار الملحوظ في الصف الأول والعمود الثاني من الجدولة المتقاطعة.

يُستمد جدول التكرارات المتوقعة بالكامل من جدول التكرارات الملحوظة، باستخدام المعادلات المبينة في الجدول (٣، ٢). تمثل الخلايا المظللة في الجدول (٣، ١) القيم التي نحتاج إلى جمعها مباشرة من المدونة (باستخدام برنامج مناسب). وهي:

- (ن): عدد الكلمات الفعلية في المدونة بأكملها.
- (ص1): تكرار العقدة في المدونة بأكملها.
- (١ع): تكرار المصاحبة في المدونة بأكملها.
- (مل11): تكرار التصاحب (مثل العقدة + المصاحبة) في نافذة التصاحب.
- حجم نافذة التصاحب.

تُستخدم المعادلات الواردة في الجدول (٣، ٣) لحساب مجموعة من قياسات التصاحب، وتشير المصطلحات إلى جدولي تصادُف: (الجدول ١، ٣ والجدول ٢، ٣).

هناك سؤال آخر يحتاج إلى إجابة، وهو: ما قياس التصاحب الأفضل؟ يشير إحباط بعضهم الجواب الآتي: يعتمد ذلك على جوانب العلاقة التصاحبية التي نريد تسليط الضوء عليها. تُبرز بعض القياسات التصاحبية مثل المعلومات المتبادلة (MI) mutual information الطابع الحصري النادر للعلاقة التصاحبية، وتفضيل المصاحبات التي تتوارد توارداً حصرياً تقريبياً مع العقدة على الرغم من أن هذا قد لا يحدث سوى مرة أو مرتين فقط في المدونة بأكملها. تفضل المقاييس الأخرى، مثل الدايس Dice واللوج دايس log Dice والمعلومات المتبادلة، MI₂ المصاحبات التي تتوارد حصرياً ولكن دون اشتراط أن تكون نادرة. يمكن للآخرين أن يأخذوا في الحسبان الاتجاهية directionality (دلتا ب Delta P) أو التشتت (كوهن -Cohen's d). لِنَلْقِ نظرة موجزة على الاتجاهية والتشتت. عند النظر في الاتجاهية نطرح السؤال الآتي: ما مدى تبادل الجذب بين العقدة والمصاحبة؟ تكون قوة جذب المصاحبة اللفظية للعقدة قريبة من قوة جذب العقدة للمصاحبة اللفظية، والكلمتان red light (الإشارة الحمراء) مثال على هذا التناظر

التصاحبي. ومن الناحية الأخرى: تكون الجاذبية أقوى بكثير في اتجاه واحد وفي علاقة غير متناظرة. على سبيل المثال، في التصاحب اللفظي red herring (الرّنجة الحمراء) تكون جاذبية كلمة herring لكلمة red أقوى بكثير من العكس (McEnergy 2006: 18)، أي إنه عندما نرى كلمة herring في نص باللغة الإنجليزية يكون هناك احتمال كبير (أكثر من 20%) في أن تسبقها كلمة red.

جدول (٢، ٣). التكرارات المتوقعة: منطلق التوارد العشوائي.

ظهور المصاحبة (affair)	غياب المصاحبة	الإجمالي
مت $_{11} = \frac{ص1 \times ع1}{ن}$	مت $_{12} = \frac{ص1 \times ع2}{ن}$	ص $_1 \times$ حجم النافذة
مت $_{21} = \frac{ص2 \times ع1}{ن}$	مت $_{22} = \frac{ص2 \times ع2}{ن}$	ص $_2$
ع $_1$	ع $_2$	ن

ومن الناحية الأخرى: عندما نرى كلمة red في نص ما لا نملك هذه الثقة لتنبأ بأن تتبعها كلمة herring لأن كلمة red يمكن أن يتبعها عدد كبير جداً من الأسماء المختلفة، فمجيئها قبل كلمة herring لا يتجاوز 0.3%. من الناحية التطبيقية، يُظهر مقياس اتجاهي مثل دلتا ب احتمالين، أحدهما لكلا اتجاهي العلاقة التصاحبية. أما المقاييس غير الاتجاهية (جميع المقاييس ما عدا دلتا ب من الجدول ٣، ٣) فلا تُظهر إلا قيمة واحدة فقط، لذلك لا يمكن استخدامها لاستكشاف الاتجاهية. أما قياس التشتت الذي يمكن استخدامه في كشف العلاقة التصاحبية فهو توزيع المصاحبات في ملفات المدونة أو أجزائها (المبحث ٣، ١ و ٤، ٢). يمكن القول إنه كلما كانت المصاحبة متساوية التشتت في المدونة ازدادت أهميتها في المدونة.

كيف نختار مقياس التصاحب تجريبياً؟ انظر الجدول (٤، ٣) الذي يوضح الأداء النموذجي للمقاييس الأربعة عشر الواردة في الجدول (٣، ٣). جُعِلت مجموعة من الكلمات بدءاً من أداة التعريف the كثيرة التكرار إلى الكلمات قليلة التكرار مثل ex-teacher أو zealand مَصاحِبَات للصفة new (التكرار المطلق = 1,233) في المدونة الفرعية BE06 المكونة من مليون كلمة. توضح الأرقام الواردة في الجدول كيف ترتب مقاييس التصاحب الفردية المصاحبات، إذ يمثل الرقم 1 الترتيب الأكثر أهمية في مجموعة المصاحبات الثمانية وفقاً للإحصاء المحدد.

جدول (٣, ٣). مقاييس التصاحب: نظرة عامة.

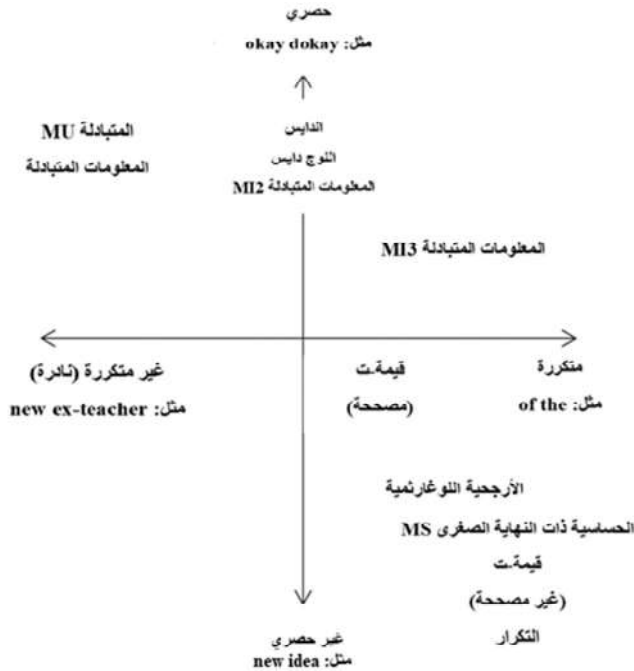
الإحصاء	المعادلة
١- تكرار التوارد frequency of co-occurrence	مل 11
٢- التبادل	$\frac{\text{مل 11}}{\text{مت 11}}$
٣- المعلومات المتبادلة (MI)	$\text{لوج}^2 \frac{\text{مل 11}}{\text{مت 11}}$
٤- المعلومات المتبادلة ₂ (MI ₂)	$\text{لوج}^2 \left(\frac{\text{مل 11}}{\text{مت 11}} \right)^2$
٥- المعلومات المتبادلة ₃ (MI ₃)	$\text{لوج}^3 \left(\frac{\text{مل 11}}{\text{مت 11}} \right)^3$
٦- الأرجحية اللوغارتمية LL	$2 \times \left(\text{لوج} \frac{\text{مل 21}}{\text{مت 21}} + \text{لوج} \frac{\text{مل 12}}{\text{مت 12}} + \text{لوج} \frac{\text{مل 22}}{\text{مت 22}} \right) - \left(\text{لوج} \frac{\text{مل 11}}{\text{مت 11}} + \text{لوج} \frac{\text{مل 21}}{\text{مت 21}} + \text{لوج} \frac{\text{مل 12}}{\text{مت 12}} + \text{لوج} \frac{\text{مل 22}}{\text{مت 22}} \right)$
٧- قيمة-ز	$\frac{\text{مل 11} - \text{مت 11}}{\sqrt{\frac{\text{مت 11}}{11}}}$
٨- قيمة-ت	$\frac{\text{مل 11} - \text{مت 11}}{\sqrt{\frac{\text{مل 11}}{11}}}$
٩- الدائيس	$\frac{2 \times \text{مل 11}}{\text{ص 1ع} + 1}$
١٠- اللوج دايس	$14 + \text{لوج} \frac{2 \times \text{مل 11}}{\text{ص 1ع} + 1}$
١١- نسبة اللوج	$\text{لوج} \frac{\text{مل 11} \times 2}{\text{مل 21} \times \text{ص 1ع}}$
١٢- حساسية النهاية الصغرى (MS)	النهاية الصغرى $\left(\frac{\text{مت 11}}{\text{ص 1ع}}, \frac{\text{مت 11}}{\text{ص 1ع}} \right)$
١٣- دلتا ب	$\frac{\text{مل 11}}{\text{ص 1ع}} - \frac{\text{مل 21}}{\text{ص 2ع}} ; \frac{\text{مل 11}}{\text{ص 1ع}} - \frac{\text{مل 21}}{\text{ص 2ع}}$
١٤- كوهن-د	$\frac{\text{المتوسط داخل النافذة} - \text{المتوسط خارج النافذة}}{\text{الانحراف المعياري المُجمّع } pooled SD}$

جدول (٤، ٣). ترتيب مُصاحبات (جديد: BE06 في new) (يس 3-يم 3).

الغاصبة	ع ^١	مل ^٢	تكرار البراد	البدال	المعلومات البدالة	المعلومات البدالة	المعلومات البدالة	الارجحية اللوعا لدية	قيمة-ن	قيمة-ت (مصححة)	قيمة-ت (غير مصححة)†	اللايس	لوح دليس	نسبة اللوح	احصائية ذات النهاية المعزى	دلائل	كوهن-د
the	58,951	447	1	7	7	8	4	2	8	7	1	2	2	7	2	3	8
and	27,917	203	2	8	8	5	5	3	7	8	2	3	3	8	3	8	7
new york	100	83	3	3	3	1	1	1	1	1	3	1	1	2	1	1	1
year	708	25	4	6	6	8	6	5	5	2	4	4	4	6	4	2	5
system	285	17	5	5	5	6	7	4	5	5	5	5	5	5	5	3	3
technologies	31	14	5	4	3	3	4	3	3	4	7	5	5	4	5	3	2
zealand	14	14	6	4	2	4	2	2	2	3	6	5	6	1	5	3	4
ex-teather	1	1	8	1	1	7	7	8	4	6	8	8	8	3	8	7	6

† تبيين قيمة-ت اختلافات كبيرة بين النتيجة المصححة وغير المصححة. تعمد القياس الأخرى مثل المعلومات البدالة مستقرة إلى حد كبير إذ تعرض النتائج المصححة قتيًا أصغر قليلًا من تلك غير المصححة.

لاختيار مقياس تصاحبي معين علينا أولاً تحديد نوع التصاحبات المعنية (بناءً على سؤالنا البحثي). يمكننا أن نفكر في معظم مقاييس التصاحب بوصفها تُسلط الضوء على التصاحبات على مدى بُعدين رئيسيين: التكرار والحصرية. يشير التكرار إلى عدد الحالات التي تتوارد فيها العقدة والمُصاحبة معاً في المدونة، في حين تشير الحصرية إلى جانب محدد من علاقة التصاحب عندما تظهر الكلمات حصرياً أو في الأغلبية العظمى معاً. وفقاً لكيفية ترتيب مقاييس التصاحب للمُصاحبات الفردية، يمكن وضع مقاييس التصاحب على مقياس ثنائي الأبعاد. تُسلط مقاييس التصاحب الضوء على تكرار أو حصرية العلاقة التصاحبية أو كليهما (الشكل ١، ٣). هذا الشكل لا يتضمن نسبة اللوج (نسبة الأرجحية) $\log \text{ratio}$ أو دلتا ب أو كوهن-د. نسبة اللوج هي مقياس يفترض مسبقاً تصفية البيانات (عادةً بالأرجحية اللوغارتمية LL $\log\text{-likelihood}$) قبل تطبيق معادلة نسبة اللوج؛ تأخذ دلتا ب الاتجاهية في الحسبان، أما كوهن-د فتحسب التششت. تعمل هذه المقاييس على أبعاد منفصلة، ومن ثم فإنها لا تُعرض في الشكل (١، ٣).



شكل (١، ٣). مدى التكرار والحصرية.

التقرير الإحصائي

١- ما يجب الإفصاح عنه

لتحقيق تطابق النتائج، ينبغي أن يشتمل التقرير على جميع البارامترات parameters الرئيسية التي يمكن أن تؤثر في تحديد المصاحبة. ولهذا الغرض أدخل بريزينا وآخرون (Brezina et al. 2015) مجموعة بارامترات التصاحب التي تُحدّد جميع البارامترات المطلوبة لتحديد المصاحبة اللفظية (الجدول ٥، ٣).

جدول (٥، ٣). مجموعة بارامترات المصاحبة collocation parameters notation.

رمز العملية الإحصائية	اسم العملية الإحصائية	القيمة الفاصلة للعملية الإحصائية	المدى الأيسر والأيمن	الحد الأدنى لتكرار المصاحبة (C)	الحد الأدنى لتكرار التصاحب (NC)	الفلتر
4ب	المعلومات المتبادلة ٢	3	يس-5 يم5	5	1	مثال { أزيلت الكلمات الوظيفية
4ب-المعلومات المتبادلة: (3) يس-5، يم5، C5-NC1؛ الكلمات الوظيفية أزيلت.						

تحتوي مجموعة بارامترات التصاحب على سبعة بارامترات مختلفة. يشير الرقم ٤ في (رمز العملية الإحصائية) إلى رقم التسلسل في الجدول (٣، ٣)، ويشير (أ) إلى أن العملية غير مصححة ويشير (ب) إلى أن العملية مصححة من العملية الإحصائية نفسها، وتشير هذه التفاصيل إلى تصحيح أحجام النوافذ التي يزيد حجمها على الحجم المحدد في الجدول (٥، ٣). ويتبع ذلك اسم العملية الإحصائية والقيمة الفاصلة للعملية الإحصائية المستخدمة، ومدى السياق الأيسر واليمين، والحد الأدنى لتكرار المصاحبة في المدونة بأكملها، والحد الأدنى لتكرار التصاحب (أي توازد العقدة والمصاحبة). يحدد البارامتر الأخير (المرشح) إجراءات أخرى في عملية استخراج التصاحب كإزالة بعض الكلمات من النتائج، أو إزالة الحد الأدنى لقيمة التشتت.

٢- مثال الإفصاح في التقرير

* حُدِدت الكلمات التالية بوصفها أبرز خمس مُصاحبات للصفة new (جديد) في مدونة BE06 وذلك باستخدام إحصاء المعلومات المتبادلة (قيمة لمعلومات المتبادلة الدنيا ٥، يم3-يس3، تكرار المتصاحب ٥ وتكرار التصاحب ٥ بدون تحديد أي قيم أخرى لبارامترات التصاحب): zealand وmobilities وyork وtechnologies وtestament.

(٣, ٣) مخططات التصاحب والشبكات: استكشاف المتجاورات المتقاطعة

فكّر...

قبل قراءة هذا المبحث، فكر في الأسئلة التالية:

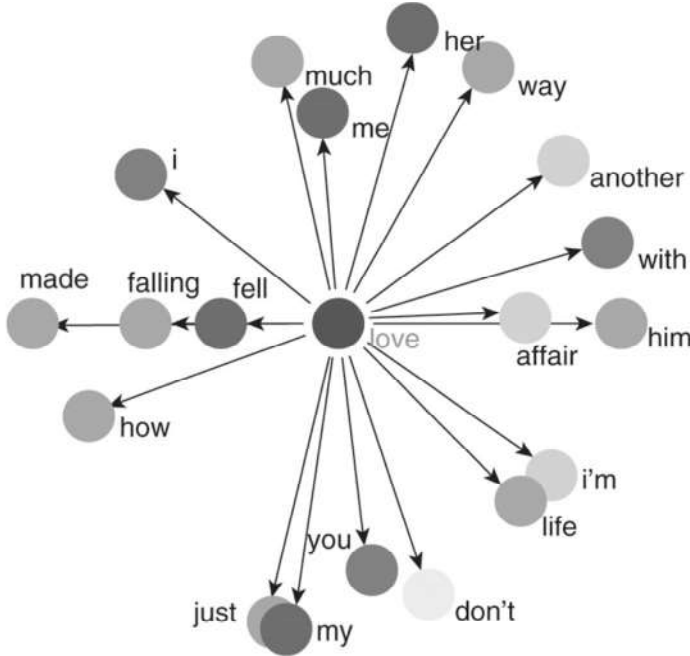
١- ما الكلمات التي تتبادر إلى ذهنك عندما ترى كلمة university؟ اكتب ما لا يقل عن خمس مُصاحبات.

٢- راجع القائمة وَّضَعْ خطأً تحت تلك الكلمات التي تعتقد أنها ترتبط ارتباطاً وثيقاً بكلمة university أكثر من بقية الكلمات الموجودة في القائمة.

تعتمد الرسوم البيانية وشبكات التصاحب على فكرة التصاحب الواردة في المبحث (٢, ٣). يمثل الشكل البياني للتصاحب تمثيلاً بصرياً للعلاقة التصاحبية بين العقدة ومصاحباتها. بدلاً من قائمة المُصاحبات المجدولة يُظهر الشكل البياني العلاقة بين العقدة ومصاحباتها بعرض المصاحبات الأقرب إلى العقدة أو الأبعد عنها.

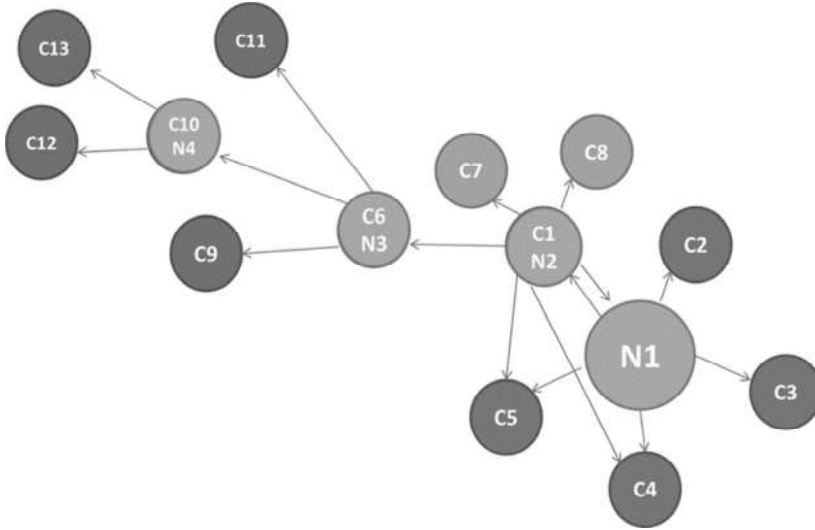
يوضح الشكل البياني في الشكل (٢, ٣) مثالاً للتصاحب من مدونة BE06 المكونة من مليون كلمة باستخدام اللوج دايس مقياساً تصاحبياً. يشغل مركز الشكل البياني العقدة love والمتصاحبات من حولها، ويعرض الشكل البياني ثلاثة أبعاد لعلاقة التصاحب، الأول: قوة التصاحب، والثاني: تكرار التصاحب، والثالث: موضع المتصاحب في النص. يُعبّر عن قوة التصاحب بمقياس اللوج دايس (المبحث ٢, ٣) بمدى قوة الرابط بين العقدة والمصاحبة، فكلما اقتربت المصاحبة من العقدة كانت العلاقة أقوى. كلما كان اللون أغمق للمتصاحب دل ذلك على

كونه أكثر تكرارًا من أقرانه. يظهر في الشكل البياني موضع التصاحب في النص سواء كان قبل العقدة -وهو الأغلب- أو بعدها، فمثلاً: تظهر المصاحبة fall و falling و fell دائماً قبل العقدة love أما you فتظهر قبل وبعد love بالتكرار نفسه تقريباً.



شكل (٢، ٣). الشكل البياني للتصاحب love (الحب) في مدونة BE06 (قيمة لوج دايس الدنيا ٧، يم-3-يس3، تكرار المصاحبة ٥ وتكرار التصاحب ٥).

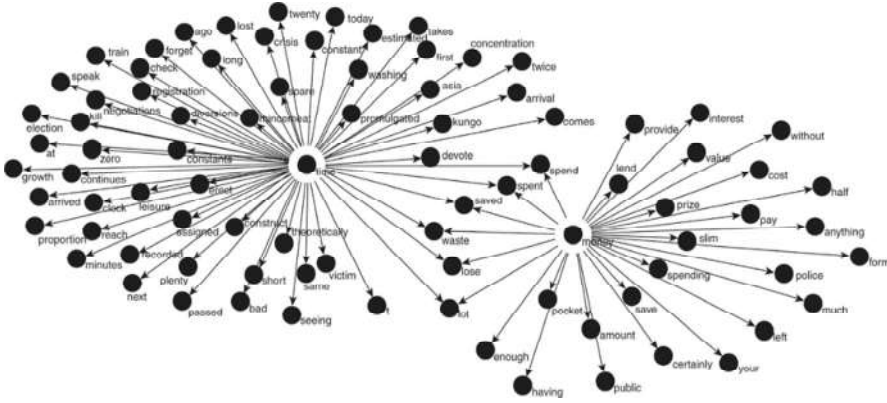
إذن، فشبكات التصاحب رسومات بيانية تصاحبية موسّعة، تُظهر أنماطاً تصاحبية أوسع من تلك التي تظهر من خلال المصاحبات المباشرة التي ناقشناها حتى الآن (Phillips 1985; Williams 1998; Brezina et al. 2015). شبكة التصاحب، كما يوحي المصطلح، هو شبكة من التصاحبات المرتبطة (الشكل ٣، ٣) تبدأ بالعقدة الأولى (N1)، وتُحدّد حولها مجموعة من المصاحبات من الدرجة الترتيبية الأولى (المتصاحب الأول C1-المتصاحب الخامس C5)، ويمكن معالجة أيّ من هذه المصاحبات بصفتها عقدة جديدة (مثل: N2 و N4 و N4 في الشكل ٣، ٣) تُحدّد حولها أيضاً مجموعة أخرى من المصاحبات.



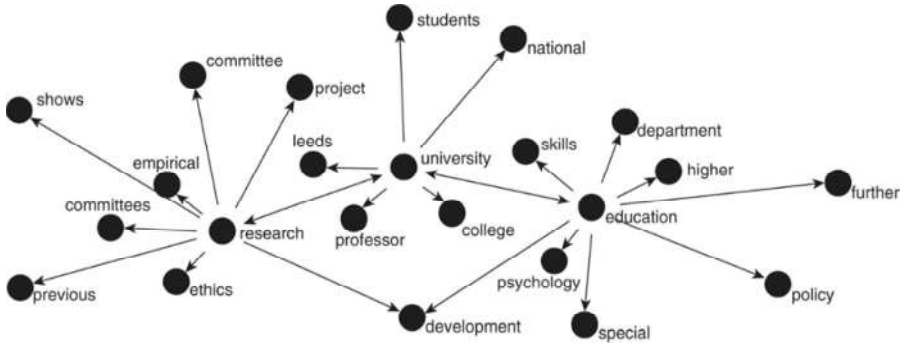
شكل (٣, ٣). شبكات التصاحب: توضيح المفهوم [يشير الـ N إلى node الذي يعني العقدة أو الكلمة الأساس، ويشير C إلى collocate الذي يعني المتصاحب].

يمكن تكرار هذه العملية عدة مرات، وقد أجريَت العملية أربع مرات في الشكل (٣, ٣) لإنشاء شبكة شاملة توضح كيفية ارتباط الكلمات التقاطعية. تعمل شبكات التصاحب بافتراض بسيط ينص على الآتي: يمكن للمتصاحبات البعيدة أن تشارك في معنى الكلمة المعنية بالتحليل بمساحة مفاهيمية ماثلة من خلال ارتباط المتصاحبات بكلمات أخرى ترتبط مباشرةً بالكلمة المعنية. من الأفضل توضيح هذه الفكرة بمثال لشبكة تصاحبية بسيطة نسبياً تستند إلى مدونة LOB المكونة من مليون كلمة (الشكل ٤, ٣) من خلال بارامتر محدد. في هذا الشكل شبكة علاقة تصاحبية بين كلمة time وكلمة money. على الرغم من أن كلمة money لا تتصاحب مباشرة مع time والعكس صحيح أيضاً إلا أنه يمكننا أن نرى ارتباطهما عبر ارتباطاتهما بكلمات مثل lot (a)، lose، وwaste، وsaved، وspent، وspend. يمكن أن تُلَمَّح هذه الارتباطات عن وجود استعمال استعارى مفهومي كما: الوقت مأل TIME IS MONEY الذي اقترحه لاکوف وجونسون (Lakoff & Johnson 1980). يمكننا أن نزع معهما بوجود هذه القرينة بين كلمة time و money كوننا نربط كثيراً ربطاً ضمناً بين الوقت والمال.

لنضرب مثالاً آخر يوضح مفهوم شبكات التصاحب، ولتكن كلمة university ومصاحباتها التقاطعية. يمكنك مقارنة شبكة التصاحب هذه بإجاباتك عن السؤالين الأول والثاني من مهمة (فكر) في أول هذا المبحث. سنرى أنه على الرغم من أن كلمة university نفسها قد يظهر معها فقط سبع متصاحبات مباشرة بمتعاير التحديد البارامترات التصاحبية المحددة فإن المساحة المفاهيمية التي تشغلها هذه الكلمة تعد أكبر بكثير عند توسيع نطاق المتصاحبات كما في الشكل (٥، ٣).



شكل (٤، ٣). مُصاحبات الدرجة الترتيبية الثالثة لكلمة time (الوقت) في مدونة LOB (قيمة المعلومات المتبادلة الدنيا ٥، ٥-٥ يس: لم تُحدد أي نقاط قيم فاصلة).



شكل (٥، ٣). الشبكة التصاحبية لكلمة university (جامعة) استناداً إلى مدونة BE06 (قيمة المعلومات المتبادلة الدنيا ٣، ٥-٥ يس: ٥).

وأخيرًا فالرسومات البيانية للتصاحب والشبكات التصاحبية هي ملخصات مفيدة للمعاني المعقدة للكلمات في النصوص والمدونات. يمكن أن توفر هذه الشبكات معلومات مفيدة عن الموضوعات الرئيسية في النصوص والخطاب. من أجل البناء الفعال لشبكات التصاحب، من الضروري استعمال برامج متخصصة قادرة على إجراء مقارنات متعددة لتجاوز الكلمات وتواردها ثم تمثيل بياناتها. هذه الأداة #LancsBox متاحة الآن للتنزيل المجاني في الموقع المرافق (Brezina et al. 2015)، وتُمكن المستخدم من رفع المدونات الخاصة به، وتُعرف شبكات التصاحب باستخدام مجموعة بارامترات التصاحب Collocation Parameters Notation (CPN) المناسب. وتُشغّل الأداة جميع مقاييس التصاحب الواردة بالجدول (٣, ٣) وتتيح للمستخدم تعديل مقاييس التصاحب وإضافة مقاييس جديدة.

(٣, ٤) الكلمات المفتاحية وكلمات الإيقاف

فكّر...

قبل قراءة هذا المبحث، فكر في السؤال التالي:
أيّ من القوائم في الجدول (٣, ٦) تجسد الكلمات التي تظهر متكررةً في اللغة الإنجليزية الأمريكية دون البريطانية؟ برّر أسباب اختيارك.

تُعد معالجة الكلمات المفتاحية أحد الأساليب الأساسية في لغويات المدونة (Scott 1997) ومن الإجراءات التي يساء فهمها في الأغلب أيضًا. تُعرّف الكلمات المفتاحية بالكلمات الأكثر تكرارًا في مدونة دون مدونة أخرى، لذلك يمكننا أن نقول إن الكلمات المفتاحية هي كلمات نموذجية للمدونة الرئيسية عند مقارنتها بمدونة أو مدونات أخرى.

جدول (٣, ٦). الكلمات المفتاحية الإنجليزية الأمريكية من مدونة AmE06.

قائمة الكلمات المفتاحية (١)	قائمة الكلمات المفتاحية (٢)	قائمة الكلمات المفتاحية (٣)	قائمة الكلمات المفتاحية (٤)
U.S. الولايات الأمريكية المتحدة	U.S. الولايات الأمريكية المتحدة	LABOR العمل	TOWARD أمام
PERCENT نسبة مئوية	PERCENT نسبة مئوية	NEIGHBORHOOD الحي (سكن)	NEIGHBORHOOD الحي (سكن)

تابع جدول (٦، ٣).

قائمة الكلمات المفتاحية (١)	قائمة الكلمات المفتاحية (٢)	قائمة الكلمات المفتاحية (٣)	قائمة الكلمات المفتاحية (٤)
AMERICAN أمريكي	PROGRAM برنامج	DEFENSE دفاع	RECOGNIZE تعرّف على
PROGRAM برنامج	TOWARD أمام	CONGRESSIONAL كونغرس	NEIGHBORS جيران
TOWARD أمام	AMERICAN أمريكي	ATLANTA أتلانتا	COLORED ملون
STATES ولايات	BUSH بوش	PGF2A بروتاجلاندين ف ٢ ألفا	MANHATTAN منهاتن
FEDERAL فدرالي	FEDERAL فدرالي	MACDOWELL ماكديويل	FAVORITE مفضل
BUSH بوش	STATES ولايات	MRNA الرنا المرسال (حمض نووي)	RECOGNIZED متعرف عليه
PRESIDENT رئيس	CENTER مركز	NEIGHBORS جيران	CENTER مركز
CENTER مركز	MR. سيد	ABBY آبي (مسلسل تلفزيوني)	REALIZE أدرك
MR. سيد	PRESIDENT رئيس	GENOME جينوم	RECOGNIZING يتعرف على
PROGRAMS برامج	PROGRAMS برامج	FLORIDA فلوريدا	TRAVELED سافر
UNITED المتحدة	UNITED المتحدة	9-11 ١١ سبتمبر	SIGNALLED أشار
STATE ولاية	WASHINGTON واشنطن	DOE وزارة الطاقة	COLOR لون
CONGRESS كونغرس	CONGRESS كونغرس	POE بوليول أستر (مادة)	CALIFORNIA كاليفورنيا
WASHINGTON واشنطن	AMERICANS أمريكيون	ROUSSEAU روسو (الفيلسوف)	GOTTEN متحصل عليه
AMERICANS أمريكيون	STATE ولاية	NSI فيروس	LABOR عمل
DEFENSE دفاع	CALIFORNIA كاليفورنيا	REZKO ريزكو (رجل أعمال)	FAVOR مفضل
CALIFORNIA كاليفورنيا	AMERICA أمريكا	MITCH ميتشل (سيناتور أمريكي)	FINALLY أخيرا
WAR حرب	DEFENSE دفاع	ADDITIVES مواد إضافية	CENTERS مراكز

من المهم أن نتذكر أن مصطلح (الكلمات المفتاحية) نسبي، ويعتمد على الاختلافات في التكرارات المعجمية بين مدونتين حاسوبيتين، وتكمن أهمية مفهومه الإجرائي في تحديد الفروقات اللغوية بين الخطابات أو المفردات في الأوعية والمجالات أو بين حقبة زمنية معينة. يُعدّ مصطلح (كلمات الإيقاف) lockwords مكملًا لمفهوم (الكلمات المفتاحية)، وقد قدمه بيكر (Baker 2011)، ويشير إلى الكلمات التي تظهر بتكرارات ماثلة في مدونتين نقارن بينها. تعتمد الكلمات المفتاحية وكلمات الإيقاف على قدرتنا على مقارنة مدونتين بطريقة ذات هدف محدد، ويُعدّ ذلك أكثر تعقيدًا مما يبدو، ويؤدي إلى بعض الجدل حول ماهية أفضل الأساليب التحليلية.

لتفصيل تحليل الكلمة المفتاحية: تُقَارَن المدونة الرئيسية corpus of interest بالمدونة المرجعية reference corpus باستخدام مقياس إحصائي يُحدّد الكلمات التي تُستخدم إما أكثر وإما أقل في المدونة الرئيسية عند مقارنتها بالمدونة المرجعية. تُعرف المدونة الرئيسية بمصطلح كمصطلح focus corpus وفقاً لـ: كيلجاريف Kilgariff (٢٠١٢)، ومصطلح node corpus وفقاً لـ: سكوت Scott (١٩٩٧). إذا ظهرت كلمة ما في المدونة الرئيسية أكثر من المدونة المرجعية فإننا نسميها كلمة مفتاحية إيجابية (+). إذا ظهرت كلمة ما في المدونة الرئيسية أقل من المدونة المرجعية فإننا نسميها كلمة مفتاحية سلبية (-). إذا استخدمت كلمة ما في المدونة الرئيسية بتكرار يقارب تكرار استخدامها في المدونة المرجعية فإنها تسمى كلمة إيقاف (0). في التطبيق العملي يحصل ذلك تلقائياً بأدوات معالجة اللغة التي تُنشئ قائمتين من الكلمات، تستند الأولى إلى المدونة الرئيسية وتستند الأخرى إلى المدونة المرجعية، ثم تُقارن بين المواد المعجمية في قائمتين لتحديد الفئة التي تنتمي إليها الكلمة من بين الكلمات المقارنة. في الجدول (٧، ٣) تلخيص لإجراء استخلاص قائمة الكلمات المفتاحية.

جدول (٧، ٣). قرارات إجراء الكلمة المفتاحية: خيارات أساسية.

القرار	المدونة المرجعية	المدونة الرئيسية
متكررة	غير متكررة	+
غير متكررة	متكررة	-
التكرار في كلا المدونتين متقارب	0	(كلمة إيقاف)

يبدو مبدأ المقارنة حتى الآن واضحاً، لكننا نحتاج إلى اتخاذ عدة قرارات محددة عند تفعيل هذا المبدأ العام. تتمثل هذه القرارات في ضرورة استكشاف خيارات مهمة عن كيفية اختيار مدونة مرجعية، وكيفية التعامل مع الكلمات الغائبة، ومعرفة الأساليب الإحصائية لإجراء المقارنة بين المدونات.

١ - كيف نختار مدونة مرجعية؟

عادةً ما تكون المدونة المرجعية أكبر من أو تساوي المدونة الرئيسية حجماً. الغرض من ذلك هو توفير قدر كافٍ من الأدلة عن تكرارات الكلمات (انظر: السؤال الثاني أدناه). كلما كانت المدونة المرجعية أكبر من المدونة الرئيسية كانت المقارنة أكثر موثوقية، وأن الاختلاف بينهما يُبرز لنا عدداً من

الجوانب المتغايرة وبدرجات متفاوتة عند معالجة الكلمات المفتاحية. لتوضيح نقطة الالتجاس النسبي هذا بين أي مدونتين، دعونا نُقارن بين قطعتين نصيّتين من مدونة أمريكية (AmE06) ومدونة بريطانية (BE06) على التوالي، يتكون كل مقتطف من 100 كلمة بالضبط.

النص أ AmE06	النص ب BE06
Democrats call those shifts too little, too late. "Changing direction in Iraq starts with changing the people we send to Washington," Shays' challenger, Diane Farrell, said Saturday in the Democratic response to Bush's radio address. Democrats, who since the Vietnam War have battled voter perceptions that they are soft on defense, are finding a more receptive audience for the argument that they could do a better job of protecting America and conducting its foreign policy. In the poll, 52% say the Iraq war has made the USA less safe from terrorism. Nonetheless, Republicans continue to view the issue of terrorism as ... يصف الديمقراطيون تلك التحولات بأنها ضئيلة للغاية ومتأخرة للغاية. وقال ديان فاريل، منافس شيس، في رد فعل ديمقراطي على خطاب بوش الإذاعي: "يبدأ تغيير الواقع في العراق بتغيير الأشخاص الذين نرسلهم إلى واشنطن". يجد الديمقراطيون، الذين حاربوا منذ حرب فيتنام تصوّر الناخبين بولوعهم بالدفاع عن أنفسهم، جمهوراً أكثر تقبلاً للحجة التي تنصّ على إمكانهم القيام بعمل أفضل لحماية أمريكا وإدارة سياستها الخارجية. قال ٥٢٪ في الاستطلاع إن حرب العراق جعلت الولايات المتحدة أقل أماناً من الإرهاب. ومع ذلك، لا يزال الجمهوريون ينظرون إلى قضية الإرهاب على أنها ...	Something behind him went 'gloink'. It was a small, subtle and yet curiously intrusive sound, and it accompanied the appearance, on a shelf above Rincewind's desk, of a beer bottle where no beer bottle had hitherto been. He took it down and stared at it. It had recently contained a pint of Winkle's Old Peculiar. There was absolutely nothing ethereal about it, except that it was blue. The label was the wrong colour and full of spelling mistakes but it was mostly there, right down to the warning in tiny, tiny print: May Contain Nuts. Now it contained a note. شيء ما وراء أصدر صوت (جلوينك)، وقد كان صوتاً ضئيلاً ومسموعاً، لكنه مزعج وغريب، ورافق ذلك ظهور زجاجة جعة على رفٍ فوق طاولة رينسويند [شخصية في الرواية]، إذ لم تكن هناك زجاجة جعة حتى في ذلك الوقت. أخذها ونظر إليها، إذ كانت تحتوي على نصف لتر من جعة أولد بيكيليار من شركة وينكل. لم يكن فيها ما يلفت الانتباه سوى ما غلب على النظر من لون المصق الأزرق الغريب والمليء بالأخطاء الإملائية. في المصق تحذير مطبوع بحروف صغيرة جداً تتضمن هذه العبارة: يحتوي على مكسّرات، وعليه لزم التنويه.

يمكننا البدء في تعداد الاختلافات بين النصين من خلال الإشارة إلى أن النص (أ) يتعلق بالسياسة الأمريكية ومأخوذ من صحيفة أما النص (ب) فهو قصة عن شخص يُسمّى (رينسويند) ومقتبس من رواية. يستخدم النص (أ) قواعد الإملاء الأمريكية (كما في تهجئة defense) أما النص (ب) فيلتزم بقواعد الإملاء البريطانية (كما في تهجئة colour). قد نبدأ أيضًا في ملاحظة الاختلافات المعجمية التي تعكس موضوعات النصين، كأن يُلاحظ بعض القراء الكلمة غير المعتادة التي ينم جرسها عن معناها gloink التي تظهر في النص (ب) والكلمات المتعلقة بحرب العراق (Bush، Iraq، و terrorism، war، إلخ) في النص (أ). هذه بعض الاختلافات التي توضح الجوانب التي قد يختلف فيها نصان عن بعضهما، وينطبق الشيء نفسه على نطاق واسع على المدونتين.

والسؤال الحاسم الذي ينبغي طرحه: ما موضوع اللغة الذي تمثله المدونة الرئيسية والمدونة المرجعية، وما الذي تعكسه تركيبة كل مدونة من المقارنة؟ (انظر: المبحث ٤، ١ للاطلاع على مناقشة تمثيلية المدونة) يلزم في التطبيق العملي أن ترتبط كل الكلمات المفتاحية (أو كلمات الإيقاف) المحددة بما نعرفه عن تركيبة المدونتين الرئيسية والمرجعية وعن مصادر الاختلاف المتعددة بينهما. ينبغي أن نفكر أيضًا في الكلمات التي سُمِّيت بوصفها كلمات مفتاحية إذا اخترنا مدونة مرجعية مختلفة. في الواقع، يمكننا اختيار عدة مدونات مرجعية وتطبيق معالجة الكلمات المفتاحية معها ثم مقارنة النتائج.

٢- كيف نتعامل مع الكلمات الغائبة؟

تُصادف عند مقارنة مدونتين تغيب الكلمة التي تتكرر في إحداها عن الأخرى. كيف نتعامل مع هذه الكلمات عند معالجة الكلمات المفتاحية؟ نحن ندرك أن انعدام الدليل في المدونة (أي: انعدام ظهور الكلمة) ليس دليلًا على الانعدام وإن كانت المدونة تُمثّل المجموع اللغوي كله (المبحث ٤، ١). للإجابة عن السؤال السابق نعتمد إلى حد كبير على الأحجام النسبية للمدونتين اللتين نقارنهما، وعلى مدى وعينا بما تمثله عينة المدونة. كقاعدة عامة، ينبغي أن نقيم بعناية دائمة مقدار الأدلة التي لدينا في المدونتين المعنيتين لمصلحة الادعاء التجريبي من أن الكلمة هي كلمة مفتاحية (إيجابية أو سلبية)، وينبغي أيضًا طرح الأسئلة التطبيقية الآتية:

- هل ص كلمة مفتاحية إيجابية أم أن المدونة المرجعية ليست كبيرة بدرجة كافية؟
- هل ص كلمة مفتاحية سلبية أم أن المدونة الرئيسية ليست كبيرة بدرجة كافية؟

في التطبيق العملي، تتيح لنا بعض أدوات المدونة تعيين الحد الأدنى للحدود الفاصلة لتكرارات الكلمات في المدونتين الرئيسة والمرجعية قبل النظر فيها في أثناء معالجة الكلمات المفتاحية وذلك يتطلب عناية فائقة.

٣- ما المقياس الإحصائي الواجب استخدامه لمقارنة المدونات؟

للإجابة عن هذا السؤال دعونا نعود إلى النصين السابقين من المدونتين الفرعية الأمريكية والبريطانية التي يتكون كل واحدة منهما من مليون كلمة تقريباً. لنفكر في خمسة عناصر معجمية تظهر مختلفة في النصين: war، defense، pint، Rincewind، و gloink، إضافة إلى كلمتين وظيفيتين عامة، وهما: the و he. ستكون المقارنة هذه المرة بين كل المدونتين المقتبس منهما النصان (أ) و(ب) على التوالي. لنفترض أيضاً أن AmE06 هي مدونتنا الرئيسية وأن BE06 هي المدونة المرجعية. يمكننا أن نرى في الجدول (٨، ٣) أن الكلمتين war و defense (بقواعد الإملاء الأمريكية) تتكرران وفق الاستعمال في الإنجليزية الأمريكية أكثر من الإنجليزية البريطانية، وتكرر كلمتي Rincewind و pint في الإنجليزية البريطانية أكثر. أما كلمة gloink فلفظة فريدة ظهرت في النص (ب)، ويُعبّر عنها بمصطلح hapax legomena الذي يعني: الكلمة التي تظهر مرة واحدة فقط في المدونة. أما the و he فهما كلمتان وظيفيتان شائعتان تظهران بتكرارات متقاربة نسبياً. إذا ترجمنا هذه المشاهدات التي رأيناها إلى قرارات عن الكلمات المفتاحية يمكننا افتراض وجود كلمتين مفتاحيتين إيجابيتين (+) وكلمة مفتاحية سلبية واحدة (-) وكلمتي إيقاف (0). أما بالنسبة لكلمتي Rincewind و gloink فما من أدلة كافية في المجموعة الأمريكية لاتخاذ قرار بشأن ندرتهما في النص (ب).

للمساعدة على اتخاذ هذه القرارات تُستخدم عادةً إحصائية الأرجحية اللوغارتمية -log likelihood (LL) لتحديد ما إذا كان من المحتمل أن تكون الاختلافات بين المدونتين الرئيسة والمرجعية ناتجة عن الصدفة أم أنها ذات دلالة إحصائية.

جدول (٨، ٣). مقارنة عناصر معجمية مختارة في مدونتي BE06 و Ame06.

الفرضية	RF ratio Ame06/BE06	R: BE06 AF (التكرار النسبي)	C: Ame06 AF (التكرار النسبي)	الكلمات
+	2.30	267	620	war حرب
		(265.00)	(609.11)	
+	118.78	1	120	defense دفاع
		(0.99)	(117.89)	
-	0.06	16	1	pint وحدة كمية
		(15.88)	(0.98)	
لا تنطبق	0.00	10	0	Rincewind رينسويند
		(9.93)	(0)	
لا تنطبق	0.00	1	0	gloink جلوينك
		(0.99)	(0)	
0	1.01	58,960	59,901	the (أل) التعريف
		(58,519.23)	(58,848.84)	
0	1.06	6,827	7,310	he هو
		(6,775.96)	(7,181.60)	

RF: التكرار النسبي؛ AF: التكرار المطلق؛ C: المدونة الرئيسية؛ R: المدونة المرجعية؛ أ: التكرار النسبي في كل مليون كلمة؛ ب: نسبة التكرار النسبي بين المدونتين Ame06 و BE06

يُستخدم الجدول المتقاطع للتكرارات الملحوظة (مل) والتكرارات المتوقعة (مت)، مع الأرجحية اللوغارتمية (LL) كما في المبحث (٢، ٣). تُدخَل القيم في معادلة الأرجحية اللوغارتمية وفق المعادلة الآتية^(٢):

$$(٣، ٤) \quad \left(\begin{matrix} \text{مل} \\ 21 \\ 21 \end{matrix} \text{ لوج} \times 21 \text{ مل} + \begin{matrix} \text{مل} \\ 11 \\ 11 \end{matrix} \text{ لوج} \times 11 \text{ مل} \right) \times 2 = \text{الأرجحية اللوغارتمية القصيرة}$$

(٢) هناك نوعان لمعادلة الأرجحية اللوغارتمية LL: المعادلة القصيرة والمعادلة الطويلة، وتختلفان قليلاً فقط في الناتج النهائي. يمكن جعل معادلة الأرجحية اللوغارتمية القصيرة بدايةً لتسهيل معادلة الأرجحية اللوغارتمية الطويلة في تحليل الكلمات المفتاحية. أُدرجت الصيغة الطويلة لمعادلة الأرجحية اللوغارتمية في الجدول (٣، ٣) مع مقاييس التصاحب.

وتدل (مل₁₁) على تكرار الكلمة الملحوظة المعنية في المدونة الرئيسية، أما (مل₂₁) فتعني تكرار الكلمة نفسها في المدونة المرجعية. أما (مت₁₁) و(مت₂₁) فتعنيان التكرارات التي نتوقعها نظرياً بالصدفة في المدونتين الرئيسة والمرجعية على التوالي، إذا لم يكن هناك فرق في تكرار الكلمة في المدونتين. تُحسب التكرارات المتوقعة على النحو الآتي:

$$\text{مت}_{11} = \frac{\text{الكلمات الفعلية في المدونة الرئيسية} \times (\text{تكرار الكلمة في المدونة الرئيسية} + \text{تكرار الكلمة في المدونة المرجعية})}{\text{إجمالي عدد الكلمات الفعلية في المدونتين الرئيسة والمرجعية}}$$

$$\text{مت}_{21} = \frac{\text{الكلمات الفعلية في المدونة المرجعية} \times (\text{تكرار الكلمة في المدونة الرئيسية} + \text{تكرار الكلمة في المدونة المرجعية})}{\text{إجمالي عدد الكلمات الفعلية في المدونتين الرئيسة والمرجعية}}$$

على سبيل المثال، لحساب إحصائية الأرجحية اللوغارتمية لكلمة war من الجدول (٨، ٣)، نستخدم التكرارات المطلقة (AF) وأحجام المدونات التالية: AmE06: 1,017,879 كلمة فعلية؛ ومدونة BE06: 1007532 كلمة فعلية.

$$\text{مت}_{11} = \frac{(267 + 620) \times 1,017,879}{2,025,411} = 445.77$$

$$\text{مت}_{21} = \frac{(267 + 620) \times 1,007,532}{2,025,411} = 441.23$$

$$\text{الأرجحية اللوغارتمية الصغيرة} = 2 \times \left(\frac{267}{441.23} \times \text{لوج} + \frac{620}{445.77} \times \text{لوج} \right) = 140.87 \quad (٥، ٣)$$

إذا كانت قيمة إحصاء الأرجحية اللوغارتمية تزيد على (3.84) التي تُعدّ نقطة الفصل للدلالة الإحصائية عند مستوى $p > 0.05$ (يمكن العثور عليها في الجداول الإحصائية أو إنتاجها مباشرة باستخدام البرمجة) يمكننا أن نستنتج أن الفرق بين استخدام كلمة war في المدونتين AmE06 وBE06 ذو دلالة إحصائية. هذا يعني أن لدينا أدلة كافية في البيانات لرفض الفرضية الصفرية التي تنص على أنه لا يوجد فرق بين تكرارات كلمة war في كل من المدونتين AmE06 وBE06 (المبحث ٣، ١). يمكننا أن نذكر في التقرير أن قيمة الاحتمال p-value أقل من 0.0001 ($p < 0.0001$) لأن قيمة الأرجحية اللوغارتمية أكبر من (15.13). يمكننا إذن أن نستنتج أن كلمة war هي كلمة مفتاحية إيجابية لمدونة AmE06 (الإنجليزية الأمريكية). على الرغم من أن الأرجحية اللوغارتمية قد تساعدنا على تقرير ما إذا كانت لدينا أدلة كافية في المدونة لادعاء أن تكرارات الكلمة نفسها تختلف بين المدونتين الرئيسة والمرجعية، وهناك أدلة أقوى (Brezina & Meyerhoff 2014; Bestgen 2014) على

أن الأرجحية اللوغارتمية المستخدمة في مقارنة المدونات تميل إلى تحديد عدد كبير جداً من الكلمات المفتاحية ومن ثم إلى نتائج خاطئة، ومن أسباب ذلك أن فروق التكرار الصغيرة نسبياً بين المدونتين الرئيسة والمرجعية يمكن أن تصل إلى حد الدلالة الإحصائية في المدونات الكبيرة. لحل مشكلة العدد الكبير من الكلمات المفتاحية المحددة ينبغي أن تُفَرَّز الكلمات المفتاحية عادةً وفقاً لإحصاء الأرجحية اللوغارتمية، ثم يُنظر فقط في تلك الكلمات ذات قيم الأرجحية اللوغارتمية الأكبر (مثل أبرز 10 أو 50 أو 100 كلمة مفتاحية) لمزيد من التحليل. يُثير هذا الإجراء الأخير ما إذا كانت إحصائيات الأرجحية اللوغارتمية هي أفضل مقياس لإظهار حجم الفرق في تكرارات الكلمات في المدونتين الرئيسة والمرجعية وكيفية تفسير النتائج تفسيراً علمياً منطلقاً من مغزى تجريبي كتفسير قيمة الأرجحية اللوغارتمية الصغيرة (140.87) التي ظهرت بعد المعالجة.

وبديلاً عن استخدام قيم إحصائيات الأرجحية اللوغارتمية التي يصعب تفسيرها أحياناً يقترح كيلجاريف Kilgariff (٢٠٠٩) النظر في النسبة بين التكرارات النسبية للكلمات في المدونتين الرئيسة والمرجعية. ونظراً لأن النسبة يمكن أن تُحسب فقط إذا كانت القيم في المدونة المرجعية أكبر من الصفر (القسمة على الصفر غير مُعرَّفة في الرياضيات) يقترح كيلجاريف إضافة ثابت k (الثابت k constant k رمز لقيمة أو عدد يُمكن توليده) لكلا التكرارين النسبيين قبل حساب النسبة. يمكن أن يكون الثابت أي عدد موجب من اختيارنا، ولكن عادة ما يكون 1 أو 10 أو 100 أو 1000. يسمى القياس الناتج البارامتر الرياضي البسيط (SMP) simple maths parameter ويُحسب على النحو الآتي:

$$(٣, ٦) \quad \frac{\text{التكرار النسبي للكلمة في المدونة الرئيسة} + k}{\text{التكرار النسبي للكلمة في المدونة المرجعية} + k} = \text{البارامتر الرياضي البسيط}$$

يعمل الثابت k في الوقت نفسه كفلتر يسمح بالتركيز في كلمات تتجاوز تكرارات نسبية معينة في المدونة. على سبيل المثال، إذا جعلنا القيمة (1) هي الثابتة فسوف نبرز الكلمات الفريدة منخفضة التكرار، أما لو جعلناها (100) فستُخرج الكلمات التي تظهر بتكرار نسبي أقل من 100 في كل مليون كلمة إذا استُخدمت خاصية (التكرار النسبي لكل مليون كلمة). يُحسب البارامتر الرياضي البسيط لكلمة war (حيث $k = 100$) على النحو الآتي:

$$1.96 = \frac{100 + 620}{100 + 267} = (\text{war}) \text{ البارامتر الرياضي البسيط للكلمة (war) } (3, 7)$$

نلاحظ أن تفسير قيمة البارامتر الرياضي البسيط أوضح من تفسير إحصاء الأرجحية اللوغارتمية لأن التكرارات النسبية للكلمة أكبر من 100 التي حددناها عند اختيار k . يمكننا القول إن كلمة war تظهر في المدونة الرئيسية نحو ضعفي عدد مرات ظهورها في المدونة المرجعية. إذن فمسألة الإحصاء الأكثر ملاءمة لتحديد الكلمات المفتاحية في الوقت الحالي هي مسألة مفتوحة. من الاقتراحات القياسية الأخرى لمبادئ الفرز: /ديف %DIFF (Gabrielatos & Marchi 2012) ونسبة اللوج log ratio (Hardie 2014) وكوهن -d Cohen's (Brezina 2014)، والأخير يأخذ التشتت في الحسبان. يمكن أيضًا تطبيق معالجة الكلمات المفتاحية على فئات ذات مستوى أعلى من التجريد النظري مثل المواد المعجمية أو المجالات الدلالية الرئيسية (Rayson 2008) فيمكننا مثلاً تحديد مفاهيم معجمية أكبر أو مجالات دلالية نموذجية لنوع من الخطاب (المدونة الرئيسية) مقارنةً بنوع آخر من الخطاب (المدونة المرجعية).

جدول (٩، ٣). الكلمات المفتاحية في اللغة الإنجليزية الأمريكية: تطبيقات مختلفة لتحديد الكلمة المفتاحية.

الأرجحية اللوغارتمية	البارامتر الرياضي البسيط	نسبة اللوج	كوهن -د
الولايات الأمريكية U.S.	الولايات الأمريكية U.S.	العمل LABOR	أمام TOWARD
نسبة مئوية PERCENT	نسبة مئوية PERCENT	الحي NEIGHBORHOOD	الحي NEIGHBORHOOD
أمريكي AMERICAN	برنامج PROGRAM	دفاع DEFENSE	تعرف على RECOGNIZE
برنامج PROGRAM	أمام TOWARD	كونغرس CONGRESSIONAL	جيران NEIGHBORS
أمام TOWARD	أمريكي AMERICAN	أتلانتا ATLANTA	ملون COLORED
ولايات STATES	بوش BUSH	بروستاجلاندين ف2 ألفا PGF2A	منهاتن MANHATTAN
فدرالي FEDERAL	فدرالي FEDERAL	ماكدويل MACDOWELL	مفضل FAVORITE
بوش BUSH	ولايات STATES	الرنا المرسال (حمض MRNA	متعرف عليه RECOGNIZED
رئيس PRESIDENT	مركز CENTER	جيران NEIGHBORS	مركز CENTER

تابع جدول (٩, ٣).

الأرجحية اللوغارتمية	البارامتر الرياضي البسيط	نسبة اللوج	كوهن-د
مركز CENTER	سيد MR.	آبي (مسلسل تلفزيوني) ABBY	أدرك REALIZE
سيد MR.	رئيس PRESIDENT	جينوم GENOME	يتعرف على RECOGNIZING
برامج PROGRAMS	برامج PROGRAMS	فلوريدا FLORIDA	سافر TRAVELED
المتحدة UNITED	المتحدة UNITED	١١ سبتمبر 9-11	أشار SIGNED
ولاية STATE	واشنطن WASHINGTON	وزارة الطاقة DOE	لون COLOR
كونغرس CONGRESS	كونغرس CONGRESS	بوليول استر (مادة) POE	كاليفورنيا CALIFORNIA
واشنطن WASHINGTON	أمريكيون AMERICANS	روسو ROUSSEAU (الفيلسوف)	متحصل عليه GOTTEN
أمريكيون AMERICANS	ولاية STATE	فيروس NSI	عمل LABOR
دفاع DEFENSE	كاليفورنيا CALIFORNIA	ريزكو (رجل أعمال) REZKO	مفضل FAVOR
كاليفورنيا CALIFORNIA	أمريكا AMERICA	ميتشل (سيناتور MITCH أمريكي)	أخيرا FINALLY
حرب WAR	دفاع DEFENSE	مواد إضافية ADDITIVES	مراكز CENTERS

قد يكون مصطلح (الكلمات المفتاحية) غامضًا بعض الشيء؛ لأنه يشير إلى وجود مجموعة واحدة من الكلمات التي تتميز بها مدونة معينة. القائمة هي نتيجة قرارات متعددة في عملية تبدأ باختيار المدونة المرجعية، وتنتهي باختيار عملية إحصائية معينة. دعونا نعود إلى السؤال الوارد في مهمة (فكر) في بداية الفصل. إذا اخترت قائمة الكلمات المفتاحية المفضلة للغة الإنجليزية الأمريكية فقد تكون مهمتها بمعرفة الإجراء الذي استخدم لتحديد هذه الكلمات المفتاحية. يوضح الجدول (٩, ٣) عدم وجود إجابة واحدة للسؤال الذي يبحث عن الكلمات المفتاحية في بيانات المدونات اللغوية.

التقرير الإحصائي

١- ما يجب الإفصاح عنه

تتأثر نتائج معالجات الكلمات المفتاحية بثلاثة معايير أساسية: (١) اختيار المدونة المرجعية، (٢) تطبيق الحد الأدنى للنقاط الفاصلة للتكرار، و(٣) اختيار المقياس الإحصائي. من المعتاد أيضًا أن يشمل التقرير على ما إذا كانت جميع الكلمات المفتاحية المحددة قد استخدمت أو فقط أهم 10

أو 50 أو 100 منها. جرت العادة التحليلية أن يشتمل التقرير على البارامترات المذكورة في المبحث السابق. على أي حال، ينبغي استكمال تلك البارامترات بمعية الوصف الدقيق للمدونة الرئيسية كما هو مذكور في المبحث المنهجي السابق.

٢- مثال الإفصاح في التقرير

البيانات

• استُخدمت مدونة AmE06 التي تمثل عينة من الإنجليزية الأمريكية المكتوبة في عام 2006. تتكون مدونة AmE06 من ١٥ قسمًا فرعيًا قائمًا على النوع، ويمكن تصنيفها إلى أربعة أصناف أوسع من حيث الوعاء، وهي: الصحف، والنثر، والنص الأكاديمي، والكتابات الروائية.

الإجراءات

• حُدِّدَت الكلمات الفريدة للغة الإنجليزية الأمريكية بالمقارنة بين مدونتي AmE06 و BE06 باستخدام إجراء الكلمات المفتاحية. أستخدمت مدونة BE06 بوصفها مرجعية إذ أنشئت باستخدام إطار عينات عائلة براون الذي يُمثل عدة مجالات للغة الإنجليزية المكتوبة، وأُخذت عينات المجالات نفسها من العام الميلادي (2006) من مدونة AmE06. استُخدمت البارامترات الرياضية البسيطة لكيلجاريف Kilgariff (٢٠٠٩) لتحديد الكلمات المفتاحية مع الثابت 100 ولم تطبَّق نقاط التكرار الفاصلة. أُجري بعد ذلك تحليل لأبرز 20 كلمة مفتاحية إيجابية.

(٣, ٥) مقاييس ملاءمة المُقدَّر البيني

فكّر...

قبل قراءة هذا المبحث، فكر في السؤال التالي:

أيُّ من أسطر الكشافات السياقية في الجدول (٣, ١٠) يتضمن معنى إيجابيا لكلمة (religion)؟
دوّن إجاباتك.

دعونا نركز في مجال لم يحظَ بعدُ بالعناية الكافية في دراسات الخطاب القائمة على المدونة، وهو ملاءمة المُقدَّر البيني. يشير ملاءمة المُقدَّر البيني inter-rater agreement إلى مدى موثوقية الترميز

واتساقه في الدراسات التي تتناول المتغير التقييمي judgement variable. هو متغير يشتمل على تصنيف أو تقييم الحالات (مثل أسطر الكشف السياقي) من خلال المحلل الذي قد يُدخل فيه عنصر الذاتية. كلما زاد عنصر الذاتية زادت الازدواجية في الترميز، ويلزم الإقرار بذلك في التقرير الإحصائي. على سبيل المثال، إذا كنا نرغب في تصنيف تكرارات كلمة time في 10 فئات دلالية مختلفة (فكر في تعريفات القاموس) فإن التصنيف سيتأثر بالذاتية لأن الدلالات غامضة كما لا يخفى. هل ينبغي تصنيف كلمة time في سياق معين على أنها س أم هل ينبغي أن تندرج تحت ص؟ من الأمثلة الأخرى: التصنيف المذكور في مهمة (فكر) إذ طُلب منك تحديد السياق الإيجابي والسياس السلبى لكل سطر من أسطر الكشف السياقي الذي يمثل استخدام كلمة religion.

من الواضح أن لطبيعة فهمك لكلمة (الدين religion) وتقييمك لها تأثيراً في الحكم بالإيجاب أو بالسلب على الأمثلة التي ترد فيها. هناك حالات لا تنطبق عليها الذاتية ولا تتطلب ترميزاً مزدوجاً كتصنيف الجمل وفقاً للفئات النحوية للأفعال إلى مبنية للمعلوم ومبنية للمجهول، وعادة ما يُمكن معالجة مثل هذه الحالات بطرق آلية تلقائية كمدونة موسومة بوسوم أجزاء الكلام ويمكن البحث فيها عن التركيبات المبنية للمعلوم والمبنية للمجهول وحسابها تلقائياً.

جدول (١٠، ٣). مدونة BE06: كشافات سياقية مختارة للعقدة (الدين: religion): أي الكشافات السياقية إيجابية وأيا سلبية؟.

السياق الأيسر	العقدة	السياق الأيمن
١	استخدمها لتلبية احتياجاتهم الخاصة، ولا تلقِ اللوم على	الدين. واجه كثير من أصدقائي العنصرية من البيض
٢	لا غرو أن الطائفة الضالة اتجهت إلى	الدين. الأصولي، الذي لا يحيد عن الانحراف عن حقيقته الجامدة. و
٣	الحرب على الإرهاب كما هو الحال في الصراعات الداخلية للطبقة	والدين. والعرق. المجتمعات الإسلامية حين تُدرس عن كُتب،
٤	حل مفاهيمنا المحدودة التي نعتز بها عن	الدين. والديمقراطية والعلمانية والرأسمالية. إنهم
٥	الذين هم أيضاً شخصيات متنفذة داخل	الدين. منظم، لهم الحق في التحدث بحرية في الساحة العامة
٦	التمييز وفقاً لتعاليم	(الدين) أو المعتقد في العدالة في العمل. إقبال سكراني، السكرتير

تابع جدول (١٠, ٣).

السياق الأيمن	العقدة	السياق الأيسر	
الضحية الفعلية أو المتصورة هو الإسلام. انتقادات للذين	الدين	الجريمة، في ٥٠٪ من الحالات، كان	٧
قديم أو جديد، أكد على عظمة الكون	دين،	وأريد منه أن يبقى هكذا".	٨
لكنه يتكلم وكله نشوة عن الطبيعة والكون	الدين،	"علم إيجابي لا يتعارض مع	٩
حقيقيا هو ذلك التقوى أو الخشوع الذي يخرج من	دينا	موقفان مختلفان تجاه العبادة. ٢٠	١٠

كيف يكون حساب ملاءمة المقدّر البيئي؟ نحتاج قبل ذلك إلى العثور على مُقدّر ثانٍ أو زميل على استعداد للمساعدة. الخطوة الأولى هي شرح نظام الترميز له بالتفصيل، ويكون على مخطط ترميزي تفصيلي. نطلب من المقدّر الثاني أن يرمز مجموعة البيانات أو العينة العشوائية المأخوذة من مجموعة البيانات ترميزًا مستقلاً. باستخدام هذا الإجراء يمكننا تقدير مدى موثوقية ترميزنا، أو بعبارة أخرى، مقدار الذاتية التي ينطوي عليها الترميز.

قبل مناقشة تفاصيل الإجراءات الإحصائية، دعونا نلقي نظرة على بعض البيانات. لنفترض أننا طلبنا من مُقدّرَيْن اثنين -الأول متدين والآخر ملحد- ترميزَ أسطر الكشف السياقي من الجدول (٣, ١٠). يمكن رؤية النتائج في الجدول (٣, ١١)، وفيه يرمز الرقم (1) إلى التقييم الإيجابي ويرمز الرقم (0) إلى التصنيف السلبي لمثال كل كشف سياقي.

جدول (٣, ١١). الترميز المزدوج للكشافات السياقية.

يوافق/ لا يوافق	ملحد	متدين	رقم الكشف السياقي
يوافق	1	1	1
يوافق	0	0	2
لا يوافق	0	1	3
يوافق	0	0	4
لا يوافق	0	1	5

تابع جدول (١١، ٣).

رقم الكشف السياقي	متدين	ملحد	يوافق/ لا يوافق
6	1	1	يوافق
7	1	0	لا يوافق
8	1	1	يوافق
9	0	0	يوافق
10	1	0	لا يوافق
إيجابي	7	3	-
سلبي	3	7	-
الإجمالي	10	10	

اتَّفَق المتدينّ والمُلحد في ست حالات واختلفا في أربع. الحالات الأربع هذه لا تتضمن أدلة كافية سياقيًا لغرض التقييم، ومن ثم يعتمد المقدّران على فهمهما الخاص. أول إحصاء يمكننا حسابه بسهولة هو المواءمة الخام. المواءمة الخام raw agreement هي مقياس يُعبّر عنه عادةً بنسبة مئوية، وتوفر نسبة حالات الموافقة في جميع الحالات. تُحسب المواءمة الخام على النحو الآتي:

$$\text{المواءمة الخام} = \frac{\text{حالات الموافقة}}{\text{إجمالي عدد الحالات}} \quad (٣، ٨)$$

بالنسبة للمقدّرين المذكورين فالمواءمة الخام هي على النحو الآتي:

$$\text{المواءمة الخام} = \frac{6}{10} = 0.6 \quad (٣، ٩)$$

يمكننا القول إذن بأن المواءمة الخام هي 0.6 أو 60% وهي نسبة منخفضة نسبيًا. لتحقيق المثالية نسعى إلى موافقة تبلغ نسبتها 80% فما فوق^(٣). مثل المتغير التقييمي judgement variable المستخدم في المثال مشكلة كبيرة بسبب انطوائه على قدر كبير من الذاتية التي أمكن ملاحظتها عندما توصل اثنان من المقدّرين إلى استنتاجات مختلفة من الحالات قُدّرت بنسبة 40%.

(٣) ينبغي أن نأخذ في الحسبان عدم وجود رقم سحري يصبح بعده الاتفاق مقبولا بالعموم. يحتاج المحلل دائمًا إلى تقييم طبيعة الخلاف وقوة الترميز بحسب سياق دراسته التحليلية.

إذن، المواءمة الخام هي تقريب أولي مفيد لموثوقية التصنيف، وعلينا أن نراعي إمكانية تحقيق بعض الموافقة بين المُقدِّرَيْن حتى لو أجريا الترميز عشوائياً. لذلك فإن مقاييس ملائمة المقدر البيني الأشد تعقيداً مثل كوهن كآبَا Cohen's K أو Gwet's AC₁ تأخذ هذه الموافقة بالصدفة في الحسبان وتطرحها من المواءمة الخام. يُستخدم كوهن كآبَا عادةً للمتغيرات الاسمية إلا أن الدراسات الحديثة (Gwet 2002) تُوصي باستخدام إحصاء Gwet's AC₁. يستند كل من كوهن كآبَا (K) أو Gwet's AC₁ إلى المعادلة نفسها لكنهما يختلفان في تقدير الموافقة التي تحدث بالصدفة (انظر: المعادلة ١٠، ٣).

$$\text{كوهن كآبَا } (K) = \frac{\text{الاتفاق الخام} - \text{الاتفاق بالصدفة}}{1 - \text{الاتفاق بالصدفة}} = \text{Gwet's AC}_1$$

وتُحسب الموافقة بالصدفة للمقياسين على النحو الآتي:

كآبَا K الموافقة بالصدفة = فرصة التصنيف من كلا المُقدِّرَيْن (س) + فرصة التصنيف من كلا المُقدِّرَيْن (ص)

AC₁ الموافقة بالصدفة = 2 × فرصة أن يصنفها كلا المُقدِّرَيْن بأنه (س) × فرصة ألا تُصنّف بـ (س)

في المثال أعلاه، يُحسب كوهن كآبَا على النحو الآتي:

الموافقة بالصدفة =	(٣، ١١)	لاحظ: 0.3×0.7 هي تعبير رياضي عن فرصة
$0.42 = 0.7 \times 0.3 + 0.3 \times 0.7$		التصنيف الإيجابي من خلال المُقدِّر الأول (7)
$0.31 = \frac{0.6 - 0.42}{1 - 0.42} = K$	(٣، ١٢)	من (10) ومن خلال المقدر الثاني (3 من 10).
		0.7×0.3 تعبير رياضي عن فرصة التصنيف
		السلبى من خلال كلا المقدرين.

وتابعاً للمثال أعلاه يُحسب AC₁ على النحو الآتي:

الموافقة بالصدفة =	(٣، ١٣)	لاحظ: في $\frac{10}{2 \times 10}$ القيمة 10 وهي عدد التصنيفات
$\left(\frac{10}{2 \times 10} - 1\right) \times \frac{10}{2 \times 10} \times 2$		الإيجابية من خلال كلا المقدرين (7 + 3)،
		و 2×10 هي لعدد تقييحات كلا المقدرين.

$$0.2 = \frac{0.6-0.5}{1-0.5} = AC_1 \quad (3, 14) \quad \left(\frac{10}{2 \times 10} - 1 \right) \text{ هي الاحتمالية المكتملة لعدم التصنيف الإيجابي بالصدفة.}$$

جدول (١٢، ٣). نظرة عامة على مقاييس ملائمة المقدّر البيني.

نوع المتغير التقديري	رقم القيم	عدد المقيمين	الإحصائيات المستخدمة
اسمي (تصنيفات)	2 فأكثر	2	Gwet's AC ₁ وكوهن كابا
	2 فأكثر	3 فأكثر	Gwet's AC ₁ وفليسّ كابا K Fleiss's
ترتيبي (رُتب)	2 فأكثر	2 فأكثر	Gwet's AC ₁
المجال/ النسبة (المقياس)	المدى	2 فأكثر	الارتباط بين الفئات interclass correlation ICC

أنتج كلا المقياسين (كابا K) و (AC₁) أرقامًا منخفضة: 0.31 و 0.2 على التوالي، وهما أقرب إلى (0) الذي يُعدّ المنطلق للموافقة العشوائية من (1) الذي يُعدّ الموافقة المطلقة والتامة.

يمكن أن يساعد المقياس أدناه على تفسير نتائج K^(٤) و AC₁. تستند القيمتان 0.67 و 0.8 بوصفهما النقاط الفاصلة للموافقة والموافقة الشديدة على التوالي إلى توصيات كريبنندورف Krippendorff (٢٠١٢ [١٩٨٠]) لتحليل محتوى النصوص. يتحدث كريبنندورف عن استنتاج مبدئي للقيم بين 0.67 و 0.8 واستنتاج نهائي للقيم أعلى من 0.8. تُفسّر هذه الاستنتاجات حسب حاجة أي نقاط فاصلة لحجم التأثير على أنها توصيات تستند إلى الخبرة بسلوك المقدّر في مجال معين وليس بوصفها مسلّمات.

يمكنُ استكمال مقاييس حجم التأثير K و AC₁ بقيمة الاحتمال p-value التي ستأخذ في الحسبان حجم العينة (عدد الحالات التي قُيِّمت) والتي ستخبرنا عما إذا كانت هناك موافقة ذات

(٤) لاحظ أن k ليس له قيمة دنيا مباشرة إذ يُستخدم لتقييم موافقة المقدّر وليس لتقييم اعتراضه. إذن، حاجة التدقيق والنظر إلى القيم الأكبر من الصفر مهمة.

دلالة إحصائية بين المقدّرين أم لا. الفرضية الصفريّة التي نختبرها هي على النحو الآتي: (هناك موافقة بالصدفة وحدها). إذا كانت قيمة الاحتمال أصغر من 0.05 فمن البديهي أن نرفض الفرضية الصفريّة ونقول إن الموافقة ليست بسبب الصدفة.

نرى أن المثال السابق عن المقدّرين وعن الفئتين (الإيجابية والسلبية) يسيرٌ جدًّا. إن كان لدينا عدد أكبر من المقدّرين أو عدد فئات تصنيفية أكثر من اثنين أو أُجري التصنيف باستخدام رُتب أو قيم رقمية يمكن وضعها على مقياس فسكون بحاجة إلى اختيار مقاييس مختلفة لموافقة المقدّر. يوضح الجدول (١٢، ٣) الإحصائيات التي يمكن استخدامها في مثل هذه الحالات. تُطبق جميع هذه الإحصاءات في أداة الموافقة في موقع أدوات إحصائيات لانكستر Lancaster Stats.

التقرير الإحصائي

١- ما يجب الإفصاح عنه

ينبغي توفير المعلومات التالية للقارئ عند تقييم موثوقية ترميز المتغير التقديري. أولاً: عدد المقدّرين، وثانياً: كمية البيانات المرُمّزة (مجموعة البيانات بأكملها أو عينة عشوائية منها)، وثالثاً: ملاءمة المقدّر البيني، ورابعاً: قيمة الاحتمال، وخامساً: تفسير النتيجة. ينبغي أن يشمل التقرير على هذه المعلومات في مبحث الطريقة في تقرير البحث.

٢- مثال الإفصاح في التقرير

وفقاً لمخطط الترميز الموصوف أعلاه، أجرى اثنان من المقدّرين المستقلين ترميز عينة عشوائية من 100 سطر من الكشف السياقي من إجمالي 1053 سطرًا يحتوي على كلمة religion في المدونة. أظهر مقياس Gwet's AC_1 وجود موافقة بين المقدّرين ($AC_1 = 0.7$) ومستوى دلالة الاختبار: $P < 0.001$. أُجريت مراجعة للاختلافات بين المقدّرين فلم يُلاحظ أي نمط منهجي للخلاف، ونظرًا لطبيعة المتغير التقديري، فقد عُدَّ مقدار الموافقة كافيًا.

(٦، ٣) التطبيق وأمثلة أخرى: ما رأي قراء الصحف البريطانية في الهجرة؟

يوضح هذا المبحث الإجراءات الإحصائية المقدمة في هذا الفصل في سياق دراسة تحليل الخطاب القصير. تركز الدراسة في تصور قارئين لصحف الأخبار البريطانية الجارديان (صحيفة

شهيرة لها ميول سياسية يسارية) وصحيفة ديلي ميل (صحيفة مسيطرة يمينية) عن موضوع (المهاجرين من أوروبا الشرقية). باستخدام مدونتين اثنتين تتضمنان تعليقات القراء تحت مقالات الهجرة على المواقع الإلكترونية للجريدة فإن الافتراض هو أن التعليقات ستكشف عن وجهات نظر مختلفة.

قد يكون من المفيد أن نُجري بحثًا تاريخيًا لتوفير ما قد يفيد التحليل من السياق التاريخي. فتحت بريطانيا في يناير ٢٠١٤ سوق العمل للمواطنين القادمين من رومانيا وبلغاريا. في المدة التي سبقت هذا الحدث ناقشت الصحافة البريطانية مرارًا التأثير المحتمل لهذا القرار في الاقتصاد البريطاني ونوعية الحياة في بريطانيا. في وسائل الإعلام أُجريت مقارنات مع حدث سابق لعام ٢٠١٤م بعشر سنوات عندما فُتح سوق العمل لمواطني بولندا والمجر وجمهورية التشيك وسلوفاكيا.

جدول (١٣، ٣). الكلمات المفتاحية.

مدونة الجارديان	مدونة ديلي ميل
جارديان Guardian	حزب الاستقلال البريطاني UKIP
كرات Balls	(أل) التعريف THE
روسيا Russia	ديلي ميل DM
دافي (مغنية) Duffy	مشرّد homeless
اقتصادي economic	فوائد benefits
جدال argument	شرطة police
مناظرة debate	الآن NOW
أبيض white	إلى TO
روسيا Russian	محتلون squatters
ملصق، بريد post	لا NO

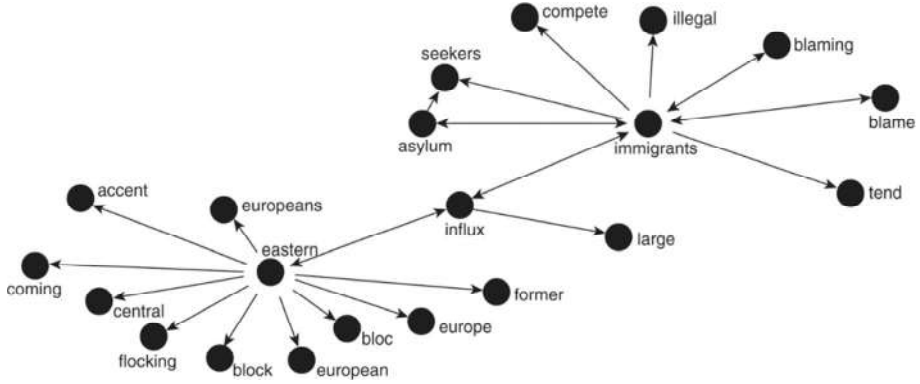
بالعودة إلى بيانات التعليقات عن مقالات موضوع الهجرة في الصحيفتين، فهي تتناول ما بين ٢٠١٠-٢٠١٣م. حُددت جميع المقالات في الجارديان والديلي ميل التي تحتوي على كلمات البحث east europeans أو eastern europeans واستُخرجت تعليقات القراء على هذه المقالات. تشير عبارة East(ern) Europeans إلى المفهوم الجمعي الذي تستخدمه الصحافة البريطانية استخدامًا

متكرراً للإشارة إلى أشخاص من دول الاتحاد الأوروبي الجديدة (مثل رومانيا أو بلغاريا أو جمهورية التشيك أو بولندا)، واستُخرجت 942,232 كلمة فعلية من الجارديان (مدونة الجارديان) و2,149,493 من ديلي ميل (مدونة الديلي ميل).

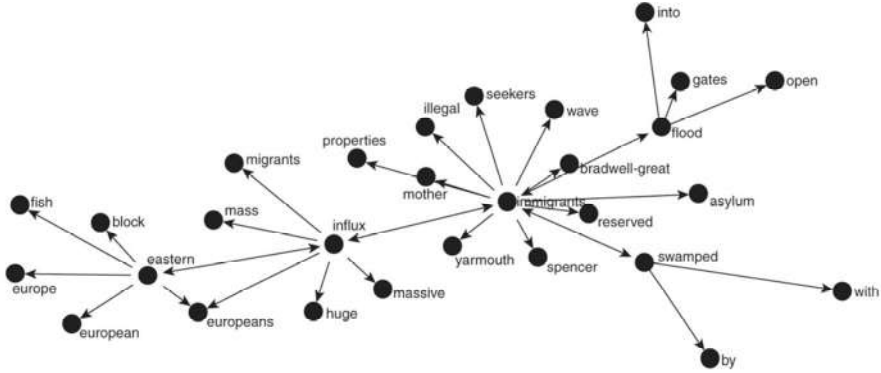
لإظهار الفرق الكلي بين المدونتين حُدِّدت أبرز 10 كلمات مفتاحية إيجابية لكل من مدونة GU ومدونة DM. عند استخراج الكلمات المفتاحية لإحدى الصحيفتين كانت تعليقات قراء الصحيفة الأخرى بمنزلة مدونة مرجعية لإبراز كلمات خاصة بقراء الجارديان أو ديلي ميل. لتحديد الكلمات المفتاحية، استخدم البارامتر الرياضي البسيط لكيلجاريف (Kilgariff 2009) مع الثابت constant: (100)، ولم تُحدّد نقاط فاصلة للتكرار. عُرِضَت الكلمات الأساسية في الجدول (١٣، ٣).

بصرف النظر عن الكلمتين المفتاحيتين المميزتين اللتين ظهرت في النتائج، وهما - بطبيعة الحال - اسم كل صحيفة: Guardian و Daily Mail يمكننا أن نرى نمطاً مثيراً للاهتمام من الاختلافات بين المدونتين. الكلمات المفتاحية في مدونة الجارديان تبدو أكثر حيادية، وتتصل بالجوانب النظرية عن موضوعات الهجرة (اقتصادي economic، وجدلي argument، ونقاشي debate)، أما الكلمات المفتاحية الموجودة في مدونة ديلي ميل فتشير في الأغلب إلى الجوانب السلبية للهجرة {بلا مأوى homeless، والفوائد benefits، والشرطة police، والمستوطنون squatters}. يمكن أيضاً ملاحظة الشدة العاطفية في خطاب قراء الديلي ميل في الاستخدام المتكرر للأحرف الكبيرة^(٥) (NO، TO، NOW، THE). بُنيت شبكة تصاحب للعقدة الأولية (immigrants) لكل مدونة. يمكن رؤية النتائج في الشكلين (٦، ٣) و(٧، ٣).

(٥) لاحظ استخراج الكلمات المفتاحية باعتبارها صغيرة لحالة الأحرف. يُمكن في النصوص التي عُولجت للتحليل تجاهل حالة الأحرف في الإنجليزية سواء كانت كبيرة أم صغيرة؛ لأنها لا تنقل أي معنى لغوي أو اجتماعي، ولذلك فإنه عادةً ما تُستخرج الكلمات المفتاحية بوصفها صغيرة لحالة الأحرف. في منتديات النقاش العام على الإنترنت فإن الكتابة بالأحرف الكبيرة هي إحدى وسائل التأكيد (بطبيعة الاستعمال اللغوي التفاعلي).



شكل (٦، ٣). شبكة التصاحب لكلمة *immigrants* (المهاجرين) في الجارديان (قيمة المعلومات المتبادلة الدنيا ٦، ٥- يس، 5، تكرار التصاحب ١٠ وتكرار التصاحب ١٠. لم تُحدد أي نقاط قيم فاصلة)



شكل (٧، ٣). شبكة التصاحب لكلمة *immigrants* (المهاجرين) في ديلي ميل (قيمة المعلومات المتبادلة الدنيا ٦، ٥- يس، 5، تكرار التصاحب ٢٠ وتكرار التصاحب ٢٠. لم تُحدد أي نقاط قيم فاصلة).

يعرض الشكل (٦، ٣) كلمة (*immigrants* أي: المهاجرون) من تعليقات قُرّاء الجارديان. تشير المتصاحبات المباشرة إلى تساوق هذه الكلمة مع كلمة (*illegal* أي: غير شرعي) وعبارة (*asylum seekers* أي: اللاجئون غير الشرعيين). على الرغم من أن توارد الكلمات يقتضي عادةً معنىً واحدًا نرى من الأمثلة الخمسة أدناه أن قُرّاء الجارديان قد ميّزوا تمييزًا واضحًا بين المهاجرين الشرعيين وطالبي اللجوء من ناحية أولى وبينها وبين المهاجرين غير الشرعيين من ناحية أخرى. إذن، فالاستعمالات حصريّة بدقة.

(1) NO NO NO NO NO WAY. ENOUGH.

[الترجمة]: مستحيل مستحيل مستحيل مستحيل. يكفي (ديلي ميل ٢٥ أبريل ٢٠١١).

(2) TOTAL DESTRUCTION OF THE UK

[الترجمة]: تدمير كامل للمملكة المتحدة (الجارديان ٣ مارس ٢٠١١)

(3) Rarely is the distinction made between asylum seekers, immigrants and illegal immigrants. Personally, I have no time for people who easily take a swipe at hard working low-paid legal migrants who often take jobs that unemployed UK citizens sometimes find unpalatable.

[الترجمة]: نادرًا ما يُميز بين طالبي اللجوء، والمهاجرين، والمهاجرين غير الشرعيين.

إنني شخصيًا، ليس لدي وقت للناس الذين يستهينون بانتقاد المهاجرين الشرعيين الكادحين ذوي الأجور المتدنية، الذين يعملون عادةً في وظائف لا يرتضيها مواطنو المملكة المتحدة العاطلون عن العمل أحيانًا (الجارديان ٢٩ أبريل ٢٠١٠).

(4) There you go again. Is "immigrants and asylum seekers" some kind of single entity to you?

[الترجمة]: كالمعتاد. هل "المهاجرون وطالبو اللجوء" يمثلون كيانًا واحدًا بالنسبة لك؟ (الجارديان ٢٩ مارس ٢٠١٠).

برزت المتصاحبتان (blame أي: يلوم) و (blaming أي: اللوم) في نقاش الجارديان، وعلى الرغم أن الكلمتين سلبيتان فإنهما تميلان ميلًا حاسمًا إلى جعل المهاجرين كباش فداء للقضايا الاجتماعية المختلفة التي يراها كثير من قراء الجارديان غير مبررة (انظر المثال الآتي).

(5) Sure there are issues, but blaming immigrants for everything isn't going to address the real issues is it?

[الترجمة]: من المؤكد أن هناك مشكلات، لكن إلقاء اللوم على المهاجرين في كل شيء لن يعالج القضايا الحقيقية؟ (الجارديان ٦ يونيو ٢٠١٠).

عندما نقارن الشكليين (٣, ٦) و (٣, ٧) نلاحظ الاختلاف الأكثر وضوحًا في التسميات المستخدمة للدلالة على كمية المهاجرين القادمين إلى بريطانيا. استعمل قراء الجارديان في تعليقاتهم كلمة (influx أي: التدفق) إلى حد كبير، واستعمل قراء ديلى ميل في تعليقاتهم الكلمات (flood أي: فيضان)، و (wave أي: أمواج)، و (swamped أي: مغمور).

للتحقق فيما إذا كان افتراض استعمال كلمة (immigrant(s) إيجابيًا أم سلبيا في تعليقات القراء، أُستخرجت عينة عشوائية مكونة من 100 سطر من كل مدونة، ثم قُزمت يدويًا على أيدي اثنين من

المقدرين على مقياس ليكرت Likert الخمسي، والقيمة (1) إيجابية للغاية، أما القيمة (5) فهي سلبية للغاية. عُدَّت ملاءمة المقدّر البيني مرتفعة إذ بلغت 0.8 للجاردان و0.93 للدليل ميل: (DM: AC2 = 0.8; GU: AC2 = 0.93)، وقد وُضِع ملخص تقيّسات القراء في الجدول (١٤، ٣).

يمكننا ملاحظة أن قُرّاء الجاردان يستخدمون كلمة immigrant(s) في السياقات الإيجابية أكثر من استخدامهم إيّاها في السياقات السلبية، أما قراء ديلي ميل فيستخدمون تقيّسات إيجابية قليلة جداً، وقد هيمنت السلبية لديهم بنسبة تتجاوز ٥٠٪.

جدول (١٤، ٣). تقيّسات immigrant(s) في مدونتي الجاردان وديلي ميل.

الصحيفة	1 (إيجابي للغاية)	2 (إيجابي)	3 (محايد)	4 (سلبى)	5 (سلبى للغاية)
الجاردان	4%	31%	39%	26%	0%
ديلي ميل	2%	6%	40%	45%	7%

(٣، ٧) تمارين

التصاحبات اللفظية

١- ما مقاييس التصاحب التي ستستخدمها في سيناريوهات البحث التالية؟ لك أن تحجب بأكثر من إجابة في كل حالة. فكّر في خط الأساس المنطقي لإجابتك.

(أ) تحتاج إلى تحديد المصطلحات الفنية المرتبطة بكلمة (process أي: معالجة) في مدونة مقالات بحثية عن الكيمياء العضوية. على سبيل المثال (petrochemical process أي: المعالجة

البتروكيماوية). لاحظ أن المصطلحات الفنية هي مجموعات كلمات حصرية ذات معنى محدد.

(ب) تريد دراسة مُتصاحبات العقدة enemy في خطاب الصحيفة، ومهتم برؤية كلمات محتوى عن هذه العقدة بدلاً من الكلمات الوظيفية المتكررة.

(ج) تريد أن تؤلف قاموس متصاحبات لتعليمي اللغة الإنجليزية يشتمل على مجموعة واسعة من التعبيرات الثابتة مثل find out وtake responsibility وdire consequences وما إلى ذلك. ينبغي أن تكون التصاحبات التي تُدرجها وحدات محددة ذات معنى، كما ينبغي أن ترد في مجموعات متكررة.

٢- انظر المعلومات في الجدول (٣, ١٥) حول توارد كلمة issue في نافذة التصاحب بامتداد ثلاث كلمات قبلية (يس3) وثلاث كلمات بعدية (يم3) في مدونة BE06، وهي مدونة من مليون كلمة باللغة الإنجليزية المكتوبة. إستخدم حاسبة التصاحب Collocation Calculator في أدوات لانكستر الشبكية لحساب أربعة مقاييس للتصاحب: المعلومات المتبادلة MI، والأرجحية اللوغارتمية LL، ودلتا B Delta P، واللوج دايس log Dice.

- عدد الكلمات الفعلية في المدونة بالكامل (n): 1,001,514
- تكرار العقدة في المدونة كلها (R_1)
- حجم نافذة التصاحب: ست كلمات قبلية وبعديّة (3يس، 3يم)

جدول (٣, ١٥). مُتصاحبات كلمة issue في مدونة BE06.

المُصاحبة	ع ^١	مل ^{١١}	قيمة المعلومات المتبادلة	قيمة الأرجحية اللوغارتمية	قيمة دلتا ب	قيمة اللوج دايس
the التعريف	58,591	101				
this هذا	4,815	38				
important مهم	322	7				
address عنوان	88	6				
قناة) بي بي سي bbc	98	5				
HUPO-PSI اختصار لـ مبادرة معايير البروتيوميّات التابعة لمنظمة البروتيوم الإنسانية لتنظيم قواعد البيانات	1	1				

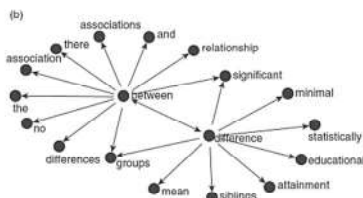
٣- ناقش ترتيب rank مقاييس المتصاحبات الست بالتمرين الثاني. أي مقاييس التصاحب ستختار؟
شبكات التصاحب

٤- قارن بين أزواج شبكات التصاحب في الشكل (٣, ٨) بين مدونة BE06 الفرعية المكونة من 840,000 كلمة بالإنجليزية البريطانية المكتوبة من الصحف والشر العام والرواية من جهة أولى، وبين المدونة الفرعية الأكاديمية BE06 التي تتكون من 160,000 كلمة إنجليزية أكاديمية. لاحظ أن المدونة

مذونة BE06 عبر الأكتينية (S40 ألف)

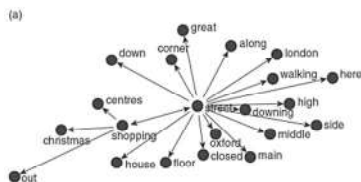


التكرار المطلق (العقدة "between"): 641



النكرار المطلق (العقدة "between"): 482

10 أسلوج نابهر (10)، بهر 5-بهس 5، C5-NC5، لم يطبق أي فقرة

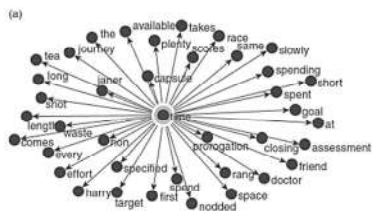


72 الشُّرَارُ الْمُطْلَقُ (العَمَلُ "shopping")

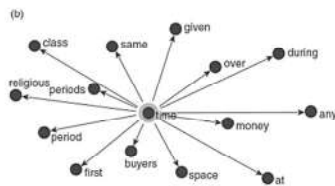


1. التكرار المطلق (العقد "shopping") :

3. ألو ج المعلومات المتبادلة (10)، يم-5، نس-5، CS-NCS، لم يطبق أي فلترة



التكرار المطلق (العقد "shopping": 1444



210 التكرار الزمني (العدد "time")

التكرار المطلق (العقدة "shopping") : 1444

شكل (٨, ٣). شبكات التصاحب المختارة.

٥- استخدم أدوات لانكستر #LancsBox، الذي يمكن تحميله من <http://corpora.lcnas.ac.uk/> لبناء شبكات التصاحب على أساس مدونة LOB (متاحة أيضًا على الموقع) المكونة من مليون كلمة تمثل الإنجليزية البريطانية المكتوبة في الستينيات، ثم ابحث عن الكلمات الآتية: (university أي: جامعة) و(time أي: وقت).

قارن بين شبكات تصاحب time و university على أساس مدونة LOB مع شبكات التصاحب التي بنيت باستخدام مدونة BE06 التي تمثل اللغة الإنجليزية البريطانية في عام 2006. كما هو موضح في المبحث (٣, ٣)، هل هناك أي اختلاف يدل على تطور اللغة؟

الكلمات المفتاحية

٦- راجع الحالات (أ)، و(ب)، و(ج) واختر نوعاً مناسباً من المدونة المرجعية (مثل مدونة اللغة العامة أو مدونة متخصصة إلخ). علّل إجابتك.

(أ) نحن مهتمون بدراسة أدبية أسلوبية. جمعنا مدونة من جميع أعمال مؤلف معين ونريد تحديد الكلمات المفتاحية المعتادة لهذا المؤلف.

(ب) نحن مهتمون بالكلمات المفتاحية النموذجية لمجال الكتابة الأكاديمية. جمعنا مدونة من المقالات البحثية والكتب في تخصصات متعددة تمثل جميع المجالات الأكاديمية الرئيسية.

(ج) نحن مهتمون بالكلمات المفتاحية النموذجية للغة المنطوقة. مدونتنا الرئيسية هي الجزء المنطوق من المدونة الوطنية البريطانية.

٧- احسب إحصائيات البارامتر الرياضي البسيط SMP للكلمات الواردة في الجدول (١٦, ٣). حدّد أي الكلمات تنتمي إلى الكلمات المفتاحية الإيجابية (+) أو إلى الكلمات المفتاحية السلبية (-) أو إلى كلمات الإيقاف (0).

جدول (١٦, ٣). الكلمات المفتاحية.

الكلمة	المدونة الرئيسية (الكلمات الفعلية: 1,007,532)	المدونة الفرعية (الكلمات الفعلية: 1,017,879)	البارامتر الرياضي البسيط	القرار (+/-) (0 /
بي بي سي BBC	106	3		
قبل before	970	854		
لندن London	471	119		
أمة nation	51	195		
هي she	4,162	4,494		
بطء slowly	83	94		

تابع جدول (١٦، ٣).

الكلمة	المدونة الرئيسية (الكلمات الفعلية: 1,007,532)	المدونة الفرعية (الكلمات الفعلية: 1,017,879)	البارامتر الرياضي البيسط	القرار (+/-) (0 /)
اليوم today	270	278		
غداً tomorrow	47	48		
واشنطن Washington	27	222		
(اسم موصول) which	2,680	2,056		

ملاءمة المقدّر البيني

٨- حُصل على التصنيفات التالية في ثلاثة مواقف تنطوي على متغير تقديري. احسب ملاءمة المقدّر البيني في كل حالة.

• الحالة الأولى: في دراسة تحليل الخطاب، أُجري ترميز متغير تقديري ذو ثلاث قيم ممكنة (1 و2 و3) على أيدي ثلاثة مقدرين مستقلين. كان المتغير الرئيسي هو المتغير الاسمي الذي يصوّر فئة الخطاب.

المقدّر أ: 1، 2، 1، 1، 2، 1، 3، 3، 2، 3، 1

المقدّر ب: 1، 2، 1، 2، 2، 1، 1، 2، 2، 2، 3، 1

المقدّر ج: 1، 2، 1، 1، 2، 1، 2، 2، 2، 2، 3، 1

• الحالة الثانية: في دراسة لغوية تطبيقية استخدمت نصوص من متحدثي اللغة الثانية. بالاستناد إلى هذه النصوص أُجري ترميز لمستوى إتقان متحدثي اللغة الثانية باستخدام فئات مرتبة ترتيباً هرمياً (متغير ترتيبي) يُراوح من (1) لأدنى كفاءة إلى (6) لأعلى كفاءة. أُجري ترميز عينته عشوائية من 20 في المئة من النصوص لتقييم متانة الترميز.

المقدّر أ: 4، 4، 4، 3، 4، 3، 3، 4، 2، 4، 4، 3، 4، 4، 4

المقدّر ب: 4، 4، 4، 3، 4، 3، 3، 4، 3، 4، 5، 5، 4، 4، 5

• الحالة الثالثة: أُعطي اثنان التسجيل الصوتي نفسه لتفريغه كتابياً. يحتوي التسجيل الصوتي على تفاعل بين ستة متحدثين. وبالنظر إلى صعوبة ترميز كل متحدث في الحوار فقد أُعتمدت موثوقية رموز المتحدثين بالأرقام (من 1 إلى 6) في أول كل دور حوار من أجل قياس ملاءمة المقدّر البيني.

الناسخ أ: 1، 4، 5، 4، 3، 4، 2، 1، 6، 2، 1، 4، 1، 2، 1، 6، 1، 4، 6، 1

الناسخ ب: 1، 4، 5، 4، 3، 4، 2، 1، 6، 2، 1، 4، 2، 6، 4، 2، 6، 4، 2، 6، 4

٩- انظر الأمثلة المأخوذة من مدونة Trinity Lancaster في الجدول (١٧، ٣). توضح هذه الأمثلة كيف يعبر متحدثو اللغة الإنجليزية كلغة أجنبية عن الرفض. حدّد كيف يكون هؤلاء المتحدثون مؤيدين أو غير مؤيدين عندما يعبرون عن الرفض. استخدم التصنيف التالي على مقياس ليكرت Likert الخمسي:



جدول (١٧، ٣). أمثلة للتصنيف.

التقييم	المثال
(أ)	I completely disagree with this because er I repeat as I said ... أختلفُ مع هذا تمامًا لأنه تكرر إيمم 11 مرة كما قلت ...
(ب)	I agree with this point but don't you think maybe the ti= fact that times are changing is a good thing? أوافق على هذه النقطة، لكن ألا ترى أن تغيّر الأوقات ربما يكون شيئًا جيدًا؟
(ج)	but I personally would disagree that that money would necessarily be spent on that (d) erm no no it's not so لكنني شخصيًا لا أوافق على أن هذه الأموال ستنفق بالضرورة على ذلك إيمم لا لا ليس كذلك
(د)	well I'm not totally convinced but er you know I live in a really traditional family حسنًا، لست مقتنعًا تمامًا ولكنني إيمم كما تعلم أعيش في أسرة تقليدية جدًا
(هـ)	mm I can understand your opinion erm but I was still wondering ... إيمم أستطيع أن أفهم رأيك ولكن ما زلت أتساءل ...
(و)	I can't agree with you لا أستطيع أن أتفق معك
(ز)	er er I I think erm I I think they I I think they are wrong إيمم إنني إنني إيمم إنني أعتقد، أعتقد أنهم أنهم على خطأ
(ح)	I think they're completely wrong أعتقد أنهم مخطئون تمامًا.
(ط)	no way إطلاقًا.
(ي)	I think he's stupid أعتقد أنه غبي.
(ك)	I I I can understand what you're saying but I'm not I don't agree with that أنا أنا أنا يمكنني أن أفهم ما تقوله لكنني لستُ لا أوافق على ذلك

بعد التقييم، أجب على الأسئلة التالية:

- ما مدى ثقتك في التصنيفات التي قدمتها؟
 - هل يُعدّ مستوى التأدّب متغيرًا تقديريًا قويًا؟
 - ما مدى أهمية وجود مقدّر آخر لهذا المتغير التقديري؟
١٠. قارن بين ترميزك في التمرين التاسع وبين ترميز مجموعة البيانات نفسها الذي أجراه مقدّر آخر (على سبيل المثال، اطلب من صديق مساعدتك على هذا التمرين). احسب مقياس المواءمة المناسب باستخدام حاسبة المواءمة:

المقياس المحسوب: _____

القيمة: _____

- استمر - إن أمكنك ذلك - في إضافة المزيد من المقدرين وحساب ملاءمة المقدّر البيئي.
- ١١ - تخيل أنك بحاجة إلى إعداد تقرير بحثي يستند إلى مجموعة البيانات التي نوقشت في التمرينين التاسع والعاشر. صمّم في التقرير نتائج مقياس ملاءمة المقدّر البيئي من التمرين العاشر. عُدّ إلى مربع (التقرير الإحصائي).

تذكر الآتي

- هناك العديد من مقياس التصاحب التي يركز كل منها على جوانب مختلفة من العلاقة التصاحبية (مثل التكرار أو الحصرية). لا يوجد مقياس للتصاحب أفضل من الآخر.
- يمكن تقديم التصاحبات في جدول أو شكل بياني.
- تُظهر شبكات التصاحب ارتباطات تقاطعية معقدة في النصوص والخطابات.
- معالجة الكلمة المفتاحية هي في جوهرها مقارنة تعتمد على عدد من البارامترات. لا وجود لشيء اسمه مجموعة واحدة من الكلمات المفتاحية.
- بالنسبة للمتغيرات التقديرية ينبغي أن يشمل التقرير على إحصاء ملاءمة المقدّر البيئي. يمكن استخدام ارتباطات AC₁ و AC₂ وكوهن وفليس وكذلك الارتباط الداخلي حسب الحالة.

قراءات إضافية

- Brezina, V., McEncry, T., & Wattam, S. (2015). Collocations in context. *International Journal of Corpus Linguistics*, 20(2), 139-73.
- Evert, S. (2008). Corpora and collocations, in A. Lüdeling & M. Kytö (eds.), *Corpus linguistics: an international handbook*, vol. 2, pp. 223-33. Berlin: Walter de Gruyter.
- Evert, S. & Krenn, B. (2001). Methods for the qualitative evaluation of lexical association measures. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, Toulouse, pp. 188-95.
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), 378.
- Gablasova, D., Brezina, V. & McEncry, A. M. (2017). Exploring learner language through corpora: comparing and interpreting corpus frequency information. *Language Learning*, 67(S1), 130-54.
- Gries, S. Th. (2013). 50-something years of work on collocations: what is or should be next.... *International Journal of Corpus Linguistics*, 18(1), 137-66.
- Gwet, K. L. (2014). *Handbook of inter-rater reliability: the definitive guide to measuring the extent of agreement among raters*. Gaithersburg, MD: Advanced Analytics.
- Kilgariff, A. (2012). Getting to know your corpus. In *Text, Speech and Dialogue*, pp. 3-15. Berlin: Springer.
- Scott, M. (1997). PC analysis of key words - and key key words. *System*, 25(2), 233-45.
- Sinclair, J., Jones, S. & Daley, R. (2004). *English collocation studies: the OSTI report*. London: Continuum.

أدوات لانكستر الإحصائية

١- يمكن استنساخ جميع التحليلات المبينة في هذا الفصل باستخدام أدوات لانكستر الإحصائية.

تشمل الأدوات المتاحة على وجه الخصوص:

• حاسبة التصاحب.

• LancsBox.

• حاسبة الكلمات المفتاحية.

• حاسبة المواءمة.

٢- يقدم الموقع أيضًا موادًا إضافية للطلاب والمعلمين.

النحو المعجمي: من الإحصاءات البسيطة إلى النماذج المعقدة

(١, ٤) ما الذي يتناوله هذا الفصل؟

يُركّز هذا الفصل في التحليل الإحصائي للخصائص النحوية المعجمية في اللغة كالأدوات أو الجمل المبنية للمجهول أو الأساليب الشرطية. نبدأ بمناقشة منهجين لدراسة النحو المعجمي^(١) في المدونات. يُستخدم المنهج الأول تصميم المدونة، ويقارن تكرارات المتغير اللغوي في المدونات الفرعية فيها، ويُستخدم المنهج الثاني تصميم (الخاصية اللغوية)، ويحدد بعناية السياقات التي يمكن أن يرد فيها متغير محدد في الاستعمال المعجمي النحوي، ويحلل العوامل التي تُسهم في ورود متغير معين من المتغيرات دون غيره. يُبيّن هذا الفصل وفق المنهج الثاني كيفية إمكان تلخيص التنوع النحوي المعجمي باستخدام الجدولة المتقاطعة cross-tabulation، وتحديد أي المقاييس الإحصائية يمكن استخدامه استناداً إلى الجدولة المتقاطعة. تُراوح هذه المقاييس من النسب المئوية البسيطة إلى اختبار كاي تربيع χ^2 والانحدار اللوجستي logistic regression. ولأنّ الانحدار اللوجستي يمثل إجراءً إحصائياً متقدماً، فإن أجزاء كبيرة من الفصل مخصصة لشرحه وتفسير نتائجه. سنجيب في هذا الفصل عن الأسئلة الأربعة الآتية:

١- كيف يمكن وصف التنوع المعجمي النحوي على أفضل وجه؟ ما أنواع التصميم البحثي التي يمكن استخدامها؟ (المبحث ٢, ٤).

(١) في هذا الفصل يدل مصطلح (النحو المعجمي) على انتظامات في اللغة تنطوي على عدد كبير من الخصائص على طول السلسلة بين كلمات المعجم lexis والقواعد النحوية التي استُكشفت باستخدام تقنيات المدونة. يمكن على وجه التحديد الاطلاع على الخصائص التي عولجت في هذا الفصل في إطار معجمي نحوي معين (انظر المناقشة في

- ٢- كيف يمكن تلخيص الاختلاف المعجمي النحوي؟ وما المقاييس الإحصائية البسيطة التي يمكن حسابها؟ (المبحث ٣، ٤)..
 ٣- كيف يمكننا بناء نماذج معقدة تراعي متغيرات متعددة تتنبأ بالتنوع المعجمي النحوي؟ (المبحث ٤، ٤).
 ٤- كيف يمكن استخدام التقنيات الإحصائية التي نوقشت في هذا الفصل في تحليل النحو المعجمي؟ (المبحث ٥، ٤).

(٢، ٤) تحليل خصائص البيانات المعجمية النحوية

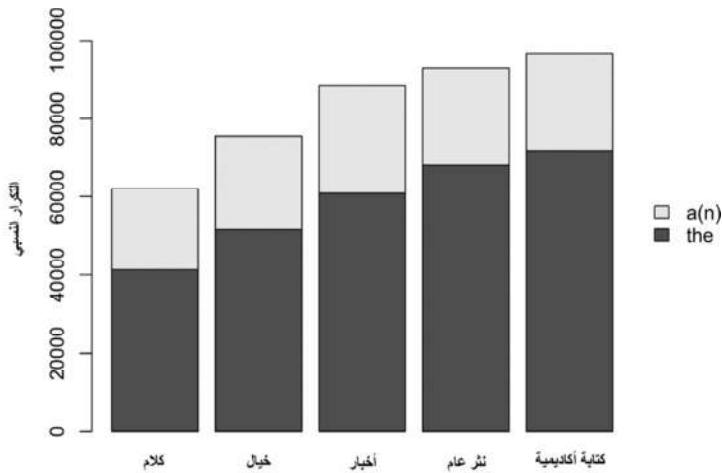
فكّر...

قبل قراءة هذا المبحث، فكر في الموقف التالي:

يُريك صديق لك يتعلم الإنجليزية جملةً في مقال صحفي بعنوان (جوجل تكشف عن شعار جديد بوصفه نقطة تحول في تاريخ الشركة).^(١) هذه الجملة هي: (The logo has undergone many, mainly small, changes in its history لقد مر الشعار في تاريخه بكثير من التغيرات الصغيرة في الأغلب). قد يسألك صديقك: لماذا استخدمت أداة التعريف في هذه الجملة؟ فكّر في كيفية الإجابة عن هذا السؤال. ما القواعد النحوية التي تفسر استخدام أداة التعريف في هذه الحالة؟

عند إتمام النظر إلى النحو المعجمي يمكننا أن نرى قدرًا كبيرًا من التنوعات المتعلقة بالمواقف التي نستخدم فيها اللغة. على سبيل المثال: قد تُظهر الكلمات النحوية/ الوظيفية البحتة كالأدوات في اللغة الإنجليزية التي قد نتوقع أن تكون مستقرة في اللغة تباينًا كبيرًا في توزيعها في أوعية أو سجلات لغوية مختلفة في اللغة الإنجليزية المنطوقة والمكتوبة (الكلام، والرواية، والصحف، والكتابة العامة، والكتابة الأكاديمية). تُوضح خطاطة الأعمدة المتكدسة في الشكل (١، ٤) تكرارات a/an and the في سجلات مختلفة من المدونة الوطنية البريطانية، وهذه الخطاطة ما هي إلا تمثيل بياني للتكرارات (النسبية) لمتغير لغوي ذي متغيرات متعددة في أجزاء مختلفة من مدونة (مدونة فرعية). هذا النوع من التمثيل مفيد لمقارنات توزيع خاصية لغوية. تُستخدم عادةً هذه التمثيلات، على سبيل المثال، في القواعد النحوية (Biber et al. 1999) التي تناقش عددًا من الأنماط المعجمية النحوية في سجلات لغوية مختلفة.

يمكننا في هذا الشكل أن نرى أن أداة التعريف أكثر تكرارًا من أداة التنكير في جميع أجزاء المدونة الوطنية البريطانية الفرعية. مع أن التنوع في استخدام أداة التنكير في المدونة الوطنية البريطانية الفرعية ليس لافتًا للنظر، هناك اختلافات واضحة في استخدام أداة التعريف. يحتوي الكلام والرواية على أدوات تعريف أقل بكثير من النثر العام والكتابة الأكاديمية. تحتل الصحف مكانًا وسطًا بين هاتين المجموعتين، إذ تشكل أداة التعريف في المتوسط ٦٪ من المدونة الفرعية الصحفية. بالعودة إلى مهمة (فكر) والتفكير في السؤال باستخدام الأدلة الممثلة في الشكل (١، ٤) يمكننا التعليق على توزيع الأدوات في سجلات لغوية مختلفة ويمكننا حتى أن نقول إنه عند الشك يُفضل استخدام أداة التعريف بدلًا من أداة التنكير بغض النظر عن نوع النص أو سجله اللغوي. لكن هل ستساعدنا هذه الملاحظات حقًا على الإجابة عن السؤال المطروح في مهمة (فكر)؟ لنأخذ في الاعتبار ما يأتي: على الرغم من أننا اكتشفنا التوزيع العام للأدوات في المدونة الوطنية البريطاني، لم نتطرق حتى الآن إلى السياقات اللغوية المباشرة (النصوص المصاحبة^(٣)) التي ترد بها الأدوات في اللغة الإنجليزية.



شكل (١، ٤). أداة التعريف وأداة التنكير في المدونة الوطنية البريطانية الفرعية.

(٣) يشير النص المصاحب إلى الكلمات الفعلية التي تحيط بالخاصية اللغوية المعنية. يستخدم هذا المصطلح للإشارة إلى السياق اللغوي المباشر للخاصية اللغوية المعنية التي يمكن أن تساعدنا على تحديد خصائص مختلفة للخاصية اللغوية مثل موقعها النحوي أو وظيفتها النحوية. يمكن ملاحظة النصوص المصاحبة في الكشف السياقي.

وتلك السياقات ضرورية لفهم أعمق للخصائص النحوية المعجمية والمبادئ اللغوية الداخلية لاستخدامها؛ لأنه سيساعدنا على الإجابة عن سؤال عن سبب استخدام خاصية نحوية معجمية معينة في سياق محدد كحال استخدام أَل التعريف. إذا اعتمدنا فقط على المعلومات الواردة في الشكل (١، ٤) فقد نخلص إلى أن القواعد التي تحكم استخدام أداة التعريف تختلف بين السجلات اللغوية المختلفة. لماذا ينبغي أن تكون أداة التعريف مفضلة بدرجة كبيرة في الكتابة الأكاديمية على سبيل المثال؟ ارتفاع نسبة استخدام الأدوات يرتبط مباشرة بالاستخدام المتكرر للأسماء، إذ كلما زاد عدد الأسماء الموجودة في نص أو سجل لغوي زاد احتمال ظهور الأدوات معها. في هذا الجواب تسهيل كبير؛ لأنه يمكن استخدام الأسماء بوصفها أحد ركني الإضافة الاسمية أو بوصفها محوَّرات لأسماء أخرى. وعليه فالأدوات في هذه الحالة مؤشرات إلى التنوع الوظيفي الأوسع بين السجلات اللغوية (في الفصل الخامس مزيد من المعلومات عن تباين السجل اللغوي). إذن: كيف يمكننا أن نستكشف بطريقة منهجية السياقات اللغوية التي ترد فيها الأدوات؟ تبدأ الخطوة الأولى باستخدام تصميم بحثي مختلف (Biber & Jones 2009).

File	Article_type	Context_type	Noun_type	Left	KWIC	Right
A0T	definite	non-determined	proper	ly Grey Walter , who called it	<<< the >>>	Contingent Negative
A0Y	definite	determined	count_pl	upported by a pillow , and all	<<< the >>>	natural outlets of th
A1D	definite	determined	uncount	th his obsessive recording of	<<< the >>>	weather . Matteo Fa
A6L	definite	determined	count_pl	he first bank that comes into	<<< the >>>	ratings is NatWest o
A7D	definite	determined	uncount	ence , she was in England for	<<< the >>>	publication of her n
A7L	definite	determined	count_sg	British filmmaking , nor was	<<< the >>>	film industry the on
ABD	indefinite	non-determined	count_sg	debate in a journal , Nature ,	<<< a >>>	week before the cor
AJX	definite	determined	proper	lens promised by Labour and	<<< the >>>	Liberal Democrats .
ALB	indefinite	non-determined	count_sg	n mind when they state : -- [	<<< A >>>	framework for estab
ALU	definite	determined	count_sg	of a beautiful red colour and	<<< the >>>	other half of a deep
ASA	definite	determined	count_sg	could n't believe my luck . In	<<< the >>>	end none of us won
ASL	definite	determined	count_sg	t where it starts will become	<<< the >>>	anus . The embryo h

شكل (٢، ٤). مقتطف من مجموعة بيانات the مقابل a(n): تصميم الخاصية اللغوية.

اعتمدنا على المدونة الوطنية البريطانية في تصميم المدونة بأكملها. نحتاج إلى تصميم الخاصية اللغوية من أجل لفت الانتباه إلى السياقات اللغوية التي ترد فيها الأدوات (انظر: المبحث ١، ٤ عن تصاميم البحث). سنبحث في المدونة عن كل تكرارات أداة التعريف وأداتي التنكير في المدونة الوطنية البريطانية، ونُرَمِّز كل مثال وفق خصائص سياقية معينة مثل نوع السياق ونوع الاسم. تظهر

مجموعة البيانات الناتجة في الشكل (٢، ٤)، ونظرًا لكثرة الأمثلة في المدونة الوطنية البريطانية التي يلزم ترميزها يدويًا^(٤) (أكثر من 8.5 مليون) يمكننا أخذ عينة عشوائية من، على سبيل المثال، 100 أو 500 أو 1000 حالة وبناء تحليلاتنا عليها.

مجموعة البيانات في الشكل (٢، ٤) هي مقتطفة من عينة عشوائية من 100 سطر من الكشف السياقي^(٥) من المدونة الوطنية البريطانية. رُمزت ثلاثة متغيرات اسمية يدويًا تعبر عن فئات غير مرتبة في هذه المجموعة (المبحث ٣، ١)، وهي: نوع الأداة Article_type، ونوع السياق Context_type، ونوع الاسم Noun_type. تتضمن مجموعة البيانات اسم الملف File وسطر الكشف السياقي نفسه KWIC الذي أُستخدم في الترميز اليدوي. من بين المتغيرات الثلاثة، يُعد نوع الأداة هو المتغير اللغوي linguistic variable المعني، ويطلق عليه أحيانًا المتغير الناتج outcome variable لأن قيمته تعتمد على المتغيرين الآخرين (نوع السياق ونوع الاسم) المُسميين المتغيرات المستقلة أو المتنبئات explanatory variables or predictors. يشير مصطلح (المتنبئ) إلى أنه يمكننا استخدام هذه المتغيرات للتنبؤ بالنتيجة لقيمة المتغير الناتج لأداة التعريف. سيتضح لنا كيف يعمل هذا التنبؤ في المبحث (٤، ٤) باستخدام الانحدار اللوجستي.

لنلق نظرة على بعض أمثلة التنوع في مجموعة البيانات في الشكل (٢، ٤). فيما يتعلق بنوع السياق الذي ترد فيه الأدوات فإن الانقسام الأساسي هو بين الاستخدامات المحددة في السياق وبين الاستخدامات غير المحددة في السياق كما هو مرمز في مجموعة البيانات. يُمثل الأول الحالات التي تشير فيها الأداة إلى شيء أو كائن أو مجرد سبق ذكره تصريحًا أو تلميحًا (المثال الأول) أو مُعَيّن

(٤) يمكن أن تساعد أدوات المدونة على ترميز الأمثلة بطرق مختلفة. على الرغم من أنه يمكن ترميز بعض المتغيرات مثل نوع الأداة في الشكل (٢، ٤) تلقائيًا فإن الفئات الوظيفية الأكثر تعقيدًا مثل نوع السياق الذي ترد فيه الأداة تتطلب عادةً مرمزًا بشريًا يمكنه تقييم كل مثال تقييمًا كاملاً. يمكن معالجة فئات مثل نوع الاسم (مفرد، جمع، علم) الذي يرد آليًا بعد الأداة باستخدام وسم أجزاء الكلام (POS) المناسب للاسم، ويلزم التحقق يدويًا من هذه العملية الآلية لأن وسم أجزاء الكلام ليس دقيقًا بنسبة ١٠٠٪ على الإطلاق. ينبغي أيضًا التحقق من قابلية جمع الأسماء يدويًا.

(٥) اختير هذا العدد من أسطر الكشف السياقي لتسهيل الشرح. سنحتاج إلى مجموعة بيانات أكبر في حال قمنا بالتحليل علميًا.

في السياق اللاحق مباشرة (المثال الثاني). أما الاستخدامات غير المحددة في السياق فيشمل الإشارة إلى شخص أو شيء أو كيان مذكور لأول مرة أو غير مُعيّن (المثال الثالث والرابع).

- (1) Lands were granted to a group of men known as feoffees, who became the legal owners of the land, while the grantor enjoyed the use of the lands - in other words, all the rights and profits arising from them. But because the feoffees were the legal owners, the lands could not be taken into wardship if the grantor died leaving an heir under age ... (BNC, file: E9V)
- (2) In September, a month after the RSPCA conference, she was in England for the publication of her new book. (BNC, file: A7D)
- (3) The kit includes a fine brass pendulum and chain along with a detailed book to point you in the right direction. (BNC, file: CBC)
- (4) This effect, which is strongest over the frontal lobes, was first observed in 1964 by Grey Walter, who called it the Contingent Negative Variation. (BNC, file: AOT)

(١) مُنحت الأراضي لمجموعة من الرجال يُعرفون بالمُقَطَّعين الذين أصبحوا أصحاب الأراضي الشرعيين، في حين كان المانح يتمتع باستخدام الأراضي وجميع الحقوق والأرباح. وعلى الرغم من أن المُقَطَّعين كانوا هم الملاك الشرعيين، لا يُعطى الوريث الشرعي تحت سن (...) حق الإرث (المدونة الوطنية البريطانية، ملف: E9V).

(٢) في سبتمبر، بعد شهر من مؤتمر الجمعية الملكية لمنع القسوة على الحيوانات، كانت في إنجلترا لنشر كتابها الجديد. (المدونة الوطنية البريطانية، ملف: A7D).

(٣) تحتوي المجموعة على بندول ساعة من النحاس النقي وعلى سلسلة، إلى جانب كتاب مفصل لتوجيهك في الاتجاه الصحيح. (المدونة الوطنية البريطانية، ملف: CBC).

(٤) لوحظ هذا التأثير الذي هو الأقوى على الفص الجبهي، لأول مرة في عام ١٩٦٤ من قبل غراي والتر الذي أطلق عليه اسم التباين السلبي المحتمل. (المدونة الوطنية البريطانية، الملف: AOT).

قبل النظر في أنواع التحليلات التي يمكننا إجراؤها باستخدام تصميم الخاصية اللغوية (المبحث ٣، ٤، و ٤، ٤) من المهم أن نلاحظ بعض الخصائص المهمة للمتغيرات المعجمية النحوية. هذه المتغيرات هي متغيرات يمكن التعبير عنها بوصفها خيارات بين متغيرين أو أكثر كالتغيرين: أداة التعريف وأداة التنكير. يعتمد إدراك أحد المتغيرين بدلاً من الآخر على السياق اللغوي، فمن السياقات ما يفضل أحدهما، ومنها من يفضل الآخر. لاستخدام تصميم الخاصية اللغوية لا بُد أن

نكون قادرين على تحديد السياقات التي يشغلها المتغير اللغوي (الناتج)، وهذا ما يطلق عليه الإطار المعجمي النحوي *lexico-grammatical frame* أو ظرف التنوع *envelope of variation* أو سياق المتغير *variable context* (Tagliamonte 2006: 70ff; Grieve-Smith 2007). إذا كنا نريد الاختيار بين أداة التعريف وأداة التنكير فنحن بحاجة إلى تحديد جميع السياقات التي تظهر فيها هذه الأدوات بالبحث عن جميع الأدوات في المجموعة. تشكل هذه السياقات الإطار النحوي المعني. إذا كنا نريد استخدامًا أكثر تحديدًا للمحددات التي تسبق الأسماء الإنجليزية عندما لا تُستخدم الأدوات فحسب، بل تُستخدم أيضًا محدّدات أخرى مثل *this, that, my, your* أو قد لا تُستخدم أصلاً فإن الإطار النحوي سيكون مختلفاً. في هذه الحالة سيتعين علينا البحث عن جميع الأسماء والتحقق مما إذا كان يسبق بعضها محدّد، وما هذا المحدّد إن وُجد.

يقدم الجدول (١، ٤) أمثلة على الخصائص المعجمية النحوية التي يمكن اختبارها باستخدام تصميم بحث الخاصية اللغوية. يُظهر الجدول نوع أسئلة البحث التي يمكن طرحها، إضافة إلى خيارات متغير الناتج (الخيارات التي لدى المتحدثين) والإطار المعجمي النحوي.

جدول (١، ٤). أمثلة على المتغيرات المعجمية النحوية ذات الإطار النحوي.

سؤال البحث	خيارات متغير الناتج	الإطار المعجمي النحوي
متى نستخدم البناء للمجهول؟	مبني للمعلوم، مبني للمجهول	جميع صيغ الفعل التي يمكن استخدامها في المبني للمجهول، مثل الأفعال المتعدية
في أي سياقات نستخدم <i>which</i> وفي أيها نستخدم <i>that</i> في العبارات الموصولة؟	<i>which, that</i>	جميع العبارات الموصولة
متى يحذف المتحدثون <i>that</i> ؟ مثل: <i>I think Ø is good.</i>	[حذف ضمير الوصل] <i>that, Ø</i>	جميع الجمل التي يثبت فيها <i>that</i> أو يُحذف
ما الفرق بين مختلف التعبيرات الناقصة الدالة على الإلزام الجازم؟	<i>must, have to, need to</i>	جميع السياقات التي ترد فيها الأفعال الناقصة الإلزامية الجازمة

تنويه هام: ينبغي أن ندرك أن المتغيرات اللغوية ليست جميعها مناسبة لنوع تصميم البحث الموضح أعلاه. لا يمكن التعامل مع المتغيرات اللغوية المحيطة *ambient linguistic variables* مثل واسمات الخطاب أو التردد *hesitations* أو الكلمات النابية التي يمكن أن تظهر في أي سياق محتمل وبدون إطار معجمي نحوي محدد بوضوح بطريقة التعامل نفسها مع المتغيرات الواردة بالجدول (١، ٤). بالنسبة لهذه المتغيرات يكون تصميم بحث للنصوص المكتوبة أو المنطوقة هو المناسب (المبحث ٣، ٦). وُضِّح الفرق بين المتغير اللغوي ذي الإطار المعجمي النحوي والمتغير اللغوي المحيط في الأسفل. مع أن للمبني للمجهول الوارد في المثال (٥) إطارًا معجميًا نحويًا واضح المعالم (جميع صيغ الفعل مع الفعل المتعدي) يمكن أن تظهر واسمات الخطاب كما هو ظاهر في المثال (٦) في أي مكان تقريبًا من الأقوال. للمتغير ذي الإطار المعجمي النحوي الواضح أيضًا عدد محدد من المتغيرات. يمكن أن تكون جميع صيغ الفعل المتعدي إما مبنية للمعلوم وإما مبنية للمجهول على الرغم من أن هناك عددًا كبيرًا من واسمات الخطاب التي ليس بينها منافسة مباشرة ويمكن أن تظهر مرارًا وتكرارًا، ويتنافس بعضها على الحيز التركيبي (النحوي) نفسه كما يوضح المثال (٦).

(5) It's about time that was done. (BNC, file: KBB)

(6) Well, you know, if you see, time were, I don't know I suppose, I don't know but I never seemed to be afraid ... (BNC, file: HDK)

(٥) لقد انتهينا متأخرين. (المدونة الوطنية البريطانية، ملف: KBB).

(٦) حسنًا، كما تعرف، وكما ترى، كان الوقت، لا أعرف ولكنني أفترض، لا أعرف

لكنني لم أبدأ أبدًا خائفًا... (المدونة الوطنية البريطانية، ملف: HDK).

لتلخيص ما ذكر، تناولنا طريقتين لتحليل الاختلاف المعجمي النحوي. الأولى في التوزيعات العامة للخصائص المعجمية النحوية في المدونات الفرعية المعروفة على نطاق واسع، والثانية عن التركيز في السياقات اللغوية الفردية التي تستخدم فيها خصائص معجمية نحوية. الطريقة التي يمكن القول إنها واعدة للإجابة عن السؤال الوارد في مهمة (فكر) هي استكشاف السياقات اللغوية التي تستخدم فيها الأدوات في اللغة الإنجليزية بعد تصميم بحث الخاصة اللغوية. أما التقنيات الإحصائية المستخدمة في مثل هذا الاستكشاف فهي موضوع الجزء المتبقي من هذا الفصل.

(٣, ٤) الجدولة المتقاطعة والنسب المئوية واختبار كاي تربيع

فكّر...

قبل قراءة هذا المبحث فكّر في الأسئلة التالية.

١ - ما أكثر التعبيرات التي تترد على لسانك؟

- يجب أن أذهب. I must go.
- علي أن أذهب. I have to go.
- أريد أن أذهب. I need to go.

٢ - هل يمكنك التفكير في سياقات يُستخدم فيها تعبيرٌ من هذه التعبيرات؟

بوصفها بداية جيدة لاستكشاف البيانات يمكن استخدام جدول تلخيصي يسير. في الفصل الثالث في المبحثين (٢, ٣ و ٤, ٣) رأينا جداول تصادُفُ أنشئت لتحليل التصاحب اللغوي والكلمات المفتاحية من أجل إظهار كل الاحتمالات (التصادفات) لورود أو توارد كلمة. الجدولة المتقاطعة (cross-tab) التي نستكشفها في هذا المبحث هي تقنية مماثلة (Hill et al. 2006: 32-8) فهي تبحث في العلاقة بين المتغيرات الفئوية categorical variables (المتغيرات الاسمية nominal والترتيبية ordinal) المستخدمة لتصنيف الملاحظات (المبحث ١, ١). تُنشأ جداول النتائج المتقاطعة^(٦) من خلال المخطط العرضي والجدولة المتقاطعة لمتغير لغوي واحد ومتغير مستقل واحد أو أكثر. أيسر أشكال الجدولة المتقاطعة هو جدول 2×2 (مكون من صفين وعمودين) بمتغير لغوي واحد مستقل^(٧)، لكل منهما فئتان أو مستويان كما هو موضح في الجدول (٢, ٤).

يعرض هذا الجدول استخدام الأداة في أنواع مختلفة من السياق. يتخذ المتغير اللغوي (الأداة) قيمتين محتملتين تُرمزان بالفئتين: (التنكير) و (التعريف). يشتمل نوع السياق متغيراً مستقلاً

(٦) هناك مصطلحات مختلفة تُستخدم للإشارة إلى هذه الجداول: (جدول النتائج المتقاطعة cross-tabulation table)، أو

(جدول التصاحب contingency table)، أو (الجدول المحوري pivot table). الأخير هو مصطلح يستخدم في

برامج جداول البيانات مثل Excel أو Calc.

(٧) في المصطلحات الإحصائية، يُسمى هذا جدولاً ثنائي الاتجاه؛ لأنه يشتمل على متغيرين. يمكننا أيضاً إنتاج جداول

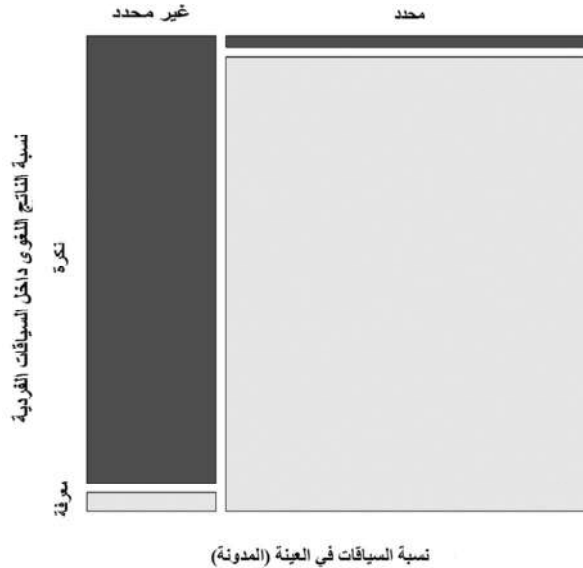
ثلاثية أو رباعية الاتجاه (انظر الجدول ٤, ٤) وما إلى ذلك بمتغير لغوي أو متغيرين أو ثلاثة متغيرات مستقلة؛ لكن

كلما كانت الجداول أكثر تعقيداً ازدادت صعوبة تفسيرها.

على فئتين أيضاً، وهما: (غير محدد) و(محدد)، وتُسرَد فئات المتغير اللغوي في أعمدة وتُسرَد فئات المتغيرات المستقلة في صفوف. يتضمن الجدول مجاميع الأعمدة والصفوف column/row totals (تسمى أحياناً التكرارات الهامشية، ومظلة بلون غامق في الجدول) بالإضافة إلى المجموع الكلي grand total (مظلل بلون أغمق في نفس الجدول).

جدول (٤, ٢). الجدولة المتقاطعة لنوع الأداة حسب التحديد السياقي.

المجموع	تعريف	تذكير	نوع السياق نوع الأداة
26	1	25	غير محدد
74	72	2	محدد
100	73	27	المجموع



شكل (٤, ٣). مخطط لنوع الأداة حسب التحديد السياقي.

تتضمن الجدولة المتقاطعة جميع المجموعات الممكنة لفئتي المتغير وعدد التكرار في كل فئة فرعية. في الجدول (٢، ٤) المجموعات الآتية، الأولى: أداة تنكير غير محددة في السياق، والثانية: أداة تعريف غير محددة في السياق، والثالثة: أداة تنكير محددة في السياق، والرابعة: أداة تعريف محددة في السياق.

يمكن تمثيل المعلومات بمخطط فسيفسائي mosaic plot كما في الشكل (٣، ٤). يحول المخطط الفسيفسائي معلومات التكرار إلى أحجام ذات مساحات. يعرض الشكل أيضًا نسب فئات متغير المتنبئ في المدونة، ونسب قيم المتغير اللغوي (النتائج) في كل فئة متنبئ. يمكننا أن نرى أيضًا بوضوح أن أكبر مساحة تمثل مجموعة من أداة تعريف محددة في السياق مع 72 ملاحظة. أما المساحة الأصغر فتُمثل أداة تعريف غير محددة في السياق مع ملاحظة واحدة فقط. تُظهر عروض المستطيلات نسبة كل نوع سياق في العينة أو المدونة (غير محددة: محددة = 26: 74)، وتُظهر أطوالها نسبة نتائج أدوات التعريف والتنكير من ضمن فئة نوع السياق (المتنبئ). لاحظ أن عرض البيانات في المخطط هو عكس الجدولة المتقاطعة، ففي المخطط الفسيفسائي تُسرد فئات المتغير المستقل أفقيًا، أما فئات المتغير اللغوي فتُسرد رأسيًا. لمزيد من المعلومات راجع (Theus and Urbanelec 2008: 50ff; Friendly 2002).

تتضمن جداول النتائج المتقاطعة وعدد مرات التكرار فيها النسب المئوية في الغالب من أجل المقارنة الأسهل عن طريق الفئات. يمكن حساب النسب المئوية من مجاميع الصفوف أو مجاميع الأعمدة أو المجموع الكلي وفقًا للمعادلة التالية:

$$(١، ٤) \quad \text{النسبة المئوية في الجدولة المتقاطعة} = \frac{\text{قيمة الخلية}}{\text{المجموع النسبي}} \times 100$$

من المهم ملاحظة أن لكل خيار من خيارات النسب المئوية الثلاث (النسب المئوية لمجموع الصف، أو مجموع العمود، أو المجموع الكلي) تفسيرًا مختلفًا تمامًا ومفيدًا لنوع مختلف من المقارنة. على سبيل المثال، تُحسب النسبة المئوية لمجموع صف الخلية الأولى في الجدول (٢، ٤) لأداة تنكير غير محددة في السياق على النحو الآتي:

$$(٢، ٤) \quad \text{النسبة المئوية للسياقات غير المحددة لأداة التنكير} = 96.2\% = 100 \times \frac{25}{26}$$

بناء على هذه النتيجة فإن الأغلبية العظمى قد فضّلت استعمال أداة التنكير على أداة التعريف، بمعنى أن هناك احتمالاً بنسبة 96.2 (أو 0.962) لورود أداة التنكير. يمكننا تفسير هذه النسبة المئوية بناءً على مجموع الصف بوصفه الاحتمال لهذا المتغير اللغوي مع تفضيل ما تزيد نسبته على 50 أو عدم تفضيل ما تقل نسبته عن 50. بدلاً من ذلك، يمكننا حساب النسبة المئوية استناداً إلى مجاميع الأعمدة، وذلك باستخدام نفس المثال الوارد في الجدول (٤, ٢) كما يأتي:

$$\text{النسبة المئوية لأدوات التنكير الواردة في السياق غير المحدد} = 100 \times \frac{25}{27} = 92.6\% \quad (٤, ٣)$$

وبناء على هذه المعادلة فإن ما نسبته 92.6 من أدوات التنكير يرد في سياقات غير محددة، وعلى الرغم من تقارب هذه النسبة مع النسبة السابقة فمنطق هذه العبارة مختلف؛ لأننا قارنا بين السياقين (المحدد وغير المحدد)، ولم نقارن بين الشكليين المختلفين للمتغير اللغوي.

جدول (٤, ٣). النسبة المئوية للخيارات في الجدولة المتقاطعة.

المجموع ذو الصلة	التفسير	الاستخدام
الصف	احتمالية شكل مختلف واحد من المتغير اللغوي في سياق معين	مقارنة استخدام الأشكال اللغوية المختلفة في سياق معين
العمود	احتمالية سياق واحد بنوع معين من الأشكال اللغوية المختلفة	مقارنة بين السياقات المختلفة
المجموع الكلي	تمثيل مختلف الفئات الفرعية في المدونة	وصف المدونة

يمكننا أخيراً حساب النسب المئوية من المجموع الكلي كما يأتي:

$$\text{النسبة المئوية لأدوات التنكير غير المحددة في السياق بالمدونة} = 100 \times \frac{25}{100} = 25\% \quad (٤, ٤)$$

ونظراً لأن المجموع الكلي في الجدول (٤, ٢) هو 100 فإن النسب المئوية من المجموع الكلي ستكون القيم الفعلية نفسها في الخلايا. تعكس هذه النسب المئوية فئات فرعية مختلفة مثل أداة التنكير غير المحددة في السياق في المدونة. هذه النسب المئوية هي نتيجة من عينات المدونة، وليست نتيجة مبدأ نحوي/ لغوي متأصل. يعرض الجدول (٤, ٣) ملخصاً لخيارات حساب النسب المئوية في الجدولة المتقاطعة مع تفسيرها، وقد كانت النسب المئوية القائمة على مجاميع الصف مناسبة.

لم نَرِ حتى الآن سوى شكل يسير جداً من الجدولة المتقاطعة (2×2). عند وجود عدد أكبر من المتغيرات المستقلة وفئاتها يمكننا إنشاء جداول نتائج تقاطعية أكثر تعقيداً. في مهمة (فكر)، طُلب منك التفكير في سياقات استخدام *must* و *have to* و *need to*. كل هذه تعبيرات تُعدّ أفعالاً مساعدة ناقصة وذات معنى متماثل يشير إلى الإلزام الجازم، وتشير إلى شيء ضروري أو يلزم القيام به. يعرض الجدول (٤، ٤) سياقات مختلفة من جهة التنوع والوعاء والفاعل. يشير كلُّ صفٍّ من الجدول إلى استخدام *must* و *have to* و *need to* في سياق محدد مثل الكتابة الأكاديمية الأمريكية مع ضمير المتكلم (I) كما في الصف الأول. حُسبت النسب المئوية في هذا الجدول من مجاميع الصفوف. يمكن تفسيرها على أنها احتمالات لكل تعبير من تلك التعبيرات المساعدة الثلاثة في سياق معين تحدده مجموعة المتنبئات predictors: (التنوع والوعاء والفاعل).

سنناقش في نهاية هذا البحث الاختبار المناسب للدلالة الإحصائية من الجدولة المتقاطعة، وُيَقِّم اختبار الدلالة الإحصائية وفق نتيجة مستوى دلالة الاختبار للفرضية الصفرية (عُدْ إلى المبحث ٣، ١).

جدول (٤، ٤). الأفعال (شبه) الناقصة الإلزامية الجازمة في مختلف الأوعية في الإنجليزية البريطانية والأمريكية.

التنوع	النوع	الفاعل	الفعل الناقص			المجموع
			<i>must</i>	<i>have to</i>	<i>need to</i>	
أكاديمي	I	0 (0.0%)	1 (100.0%)	0 (0.0%)	1	
	you	3 (33.3%)	1 (11.1%)	5 (55.6%)	9	
	other	63 (64.3%)	18 (18.4%)	17 (17.3%)	98	
	I	8 (12.7%)	34 (54.0%)	21 (33.3%)	63	
	you	16 (24.6%)	35 (53.8%)	14(21.5%)	65	
	other	51 (28.5%)	87 (48.6%)	41 (22.9%)	179	
	I	8 (19.5%)	27 (65.9%)	6(14.6%)	41	
	you	4 (10.0%)	24 (60.0%)	12(30.0%)	40	
	other	152 (56.5%)	72 (26.8%)	45 (16.7%)	269	
صحفي	I	1 (50.0%)	1 (50.0%)	0 (0.0%)	2	
	you	2 (20.0%)	7 (70.0%)	1 (10.0%)	10	
	other	44 (33.6%)	48 (36.6%)	39 (29.8%)	131	

تابع جدول (٤، ٤).

النوع	النوع	الفاعل	الفعل الناقص		المجموع
أكاديمي	I	0 (0.0%)	0 (0.0%)	1 (100.0%)	1
	you	1 (100.0%)	0 (0.0%)	0 (0.0%)	1
	other	37 (39.8%)	23 (24.7%)	33 (35.5%)	93
خيال	I	19 (28.8%)	34 (51.5%)	13 (19.7%)	66
	you	14 (27.5%)	29 (56.9%)	8 (15.7%)	51
	other	88 (41.3%)	107 (50.2%)	18 (8.5%)	213
عام	I	9 (23.1%)	27 (69.2%)	3 (7.7%)	39
	you	21 (32.8%)	21 (32.8%)	22 (34.4%)	64
	other	193 (53.0%)	105 (28.8%)	66 (18.1%)	364
صحفي	I	4 (25.0%)	12 (75.0%)	0 (0.0%)	16
	you	2 (18.2%)	7 (63.6%)	2 (18.2%)	11
	other	60 (48.4%)	40 (32.3%)	24 (19.4%)	124
المجموع		800	760	391	1951

يُمكن استخدام اختبار كاي تربيع ما دام الجدول المتقاطع يعمل مع البيانات الفئوية (Azen & Walker 2011: 58-9; Sheskin 2007: 493-561 Balakrishnan et al. 2013;). يُعدُّ كاي تربيع مناسباً للجدول البسيط ذات المتغير اللغوي الواحد والمتغير المستقل الواحد كما في الجدول (٤، ٢). ينبغي التحقق من افتراضات الاستخدام المناسب وإجراءاته قبل استخدام هذا الاختبار. لكاي تربيع افتراضان:

١- الافتراض الأول: استقلال الملاحظات (المتغيرات الملحوظة/المشاهدة كما هي) *independence of observations*. نحن نفترض أن كل ملاحظة مثل استخدام أداة في اللغة الإنجليزية كما في الجدول (٤، ٢) مستقلة عن الأخرى. نحتاج إلى تقليص هذا الافتراض في المدونات، وقد ندفع ثمن هذا التقليص على أي حال. ولكن، عندما ننظر إلى النصوص في المدونة نرى أن الخصائص اللغوية في كل نص مترابطة، ولذلك فإن المدونة بغض النظر عن جودتها ليست عينة (عشوائية) من الخصائص اللغوية (أو الكلمات)، بل عينة من النصوص التي تجمع بين عدد من الخصائص اللغوية المترابطة (Kilgariff 2005). يدفعنا هذا القول إلى نقض افتراض استقلال الملاحظات إلى حد ما، وقد يؤدي هذا النقض إلى زيادة عدد النتائج ذات الدلالة الزائفة والخطئة.

٢- الافتراض الثاني: تكرار متوقع أكبر من 5 expected frequencies greater than 5، أي: زيادة لا تقل عن ٨٠٪ من التكرارات المتوقعة عن خمسة في الجدولة المتقاطعة التي تتضمن صفين وعمودين (2x2). التكرارات المتوقعة هي التكرارات الأساسية التي نحسبها عند إجراء كاي تربيع (انظر الفقرة التالية). تطبيق هذا الافتراض على مجموعة بيانات المدونة التي تكون عادة كبيرة ليس صعباً. عند انتقاض هذا الافتراض يكون اختبار الاحتمالية اللوغارتمية (المعروف أيضاً باسم اختبار نسبة الاحتمالية Likelihood ratio test (G test)، أو اختبار الدقة ليفشر Fisher exact test أكثر ملاءمة لتقدير الدلالة الإحصائية من قيم الاحتمال (Upton 1992; Sprent 2011 Rayson et al. 2004).

يُحسب كاي تربيع وفقاً للمعادلة التالية^(٨):

$$\text{كاي تربيع} = \text{مجموع الخلايا كلها للنتيجة الآتي} \frac{2(\text{التكرار الملحوظ} - \text{التكرار المتوقع})}{\text{التكرار المتوقع}} \quad (٥, ٤)$$

نعلم من المبحث (٢, ٣) أن التكرارات المتوقعة هي تكرارات نتوقع أن نراها ما لم تكن هناك علاقة بين المتغيرات في البيانات، وكانت الفرضية الصفرية صحيحة. نستخدم التكرارات المتوقعة أساساً لتحديد ما إذا كانت هناك علاقة حقيقية بين المتغيرات. في حالتنا هنا، نختبر تأثير المتغير المستقل في المتغير اللغوي، وعليه تُحسب التكرارات المتوقعة على النحو الآتي:

$$\text{التكرار المتوقع} = \frac{\text{مجموع الصف} \times \text{مجموع العمود}}{\text{المجموع الكلي}} \quad (٦, ٤)$$

لنأخذ على سبيل المثال استخدام أدوات التعريف في اللغة الإنجليزية. وردت في الجدول (٢, ٤) بيانات التكرارات الملحوظة. باستخدام المعادلة (٦, ٤) مع البيانات الواردة في الجدول (٢, ٤) حَسَبْنَا التكرارات المتوقعة كما في الجدول (٥, ٤).

(٨) أحياناً قد تظهر قيمة كاي تربيع مختلفة قليلاً عن التي تنتجها المعادلة (٥, ٤) بسبب المعالجة البرمجية. يرجع ذلك إلى أن بعض الحزم تستخدم كاي تربيع مع ما يسمى بتصحيح Yates. على أي حال، يكون الفرق ضئيلاً من الناحية العملية.

$$(\xi, \gamma) \quad 85.25 = \frac{2(54.02-72)}{54.02} + \frac{2(19.98-2)}{19.98} + \frac{2(18.98-1)}{18.98} + \frac{2(7.02-25)}{7.02} = \text{کای تربیع}$$

المجموع	تعريف	تنكير	نوع السياق
			نوع الأداة
26	$\frac{26 \times 73}{100} = 18.98$	$\frac{26 \times 27}{100} = 7.02$	غير محدد
74	$\frac{26 \times 27}{100} = 54.02$	$\frac{26 \times 27}{100} = 19.98$	محدد
100	73	27	المجموع

بالنسبة لجدول العمودين والصفّين من المتغيّرات (2×2)، تبلغ قيمة 0.05 للدلالة الحرجة (أي: النقطة الفاصلة): 3.84 كما تبلغ قيمة 0.01 للدلالة الحرجة 6.63^(١)، وهذا يعني أن أي قيمة لاختبار كاي تربيع تكون أكبر من 3.84 و 6.63 تُعدّ ذات دلالة عند مستوى 0.05 ومستوى 0.01. في سياق تحليلنا هنا تكون قيمة الاحتمال المرتبطة بقيمة الاختبار 85.25 صغيرة جدًا $p < 0.0000000000000001$ ، وتدون في التقرير على النحو: < 0.0001 . تذكّر أن الاختبارات الإحصائية ليست منافسةً لإظهار أصغر قيمة احتمال (المبحث ٤ ، ١). نقيّم مقدار الأدلة التي لدينا في البيانات لرفض الفرضية الصفرية التي تنص على عدم وجود علاقة بين المتغيرات في البيانات (استخدام الأدوات ونوع السياق في الجدول ٢ ، ٤). إضافةً إلى الدلالة الإحصائية، نعبر أيضًا باصطلاحات موحدة عن حجم الفرق بين الفئات في الجدولة المتقاطعة. يجري ذلك بالإفصاح عن قياس حجم التأثير. بالنسبة لاختبار كاي تربيع فلدينا العديد من الخيارات، فالعملية الإحصائية التي تُحسب

(٩) من الناحية الإحصائية، نشير هنا إلى درجة الحرية (df) في توزيعات كاي تربيع. يمكن العثور على النقاط المهمة للدلالة في الجداول الإحصائية تحت درجة الحرية المناسبة. تكون درجة الحرية (1) للجدولة المتقاطعة: 2×2 ، وتكون (2) للجدولة المتقاطعة: 2×3 ، وتكون (4) للجدولة المتقاطعة: 3×3 ، أي أن نتيجة درجة الحرية تساوي دائماً (عدد الصفوف - 1) $\times$ (عدد الأعمدة - 1). لا تقلق إذا كان هذا التعليق تقنياً. تحسب الأداة الإحصائية في أداة لانكستر قيم الاحتمال تلقائياً؛ حتى لا تضطر أبداً إلى البحث عن درجة الحرية وقيم الاحتمال المقابلة لها.

لحجم التأثير العام الكلي في الجدولة المتقاطعة هي معامل كريمر ف Cramer's V. هذه المعامل^(١٠) هي قيمة موحدة لكاي تربيع تضبط إجمالي عدد الملاحظات (المجموع الكلي في الجدولة المتقاطعة) في حال ظلت قيم إحصائيات كاي تربيع تزداد بازدياد حجم العينة. هذه الزيادة هي التأثير المقصود لاختبار الفرضية الصفرية؛ لأن اختبار كاي تربيع يقيّم مقدار الأدلة مقابل الفرضية الصفرية. ومع ذلك، ينبغي أن يكون حجم التأثير (معامل كريمر ف) قابلاً للمقارنة عبر أحجام عينات مختلفة من المدونات، ومن ثمّ يستخدم التوحيد التالي:

$$\text{معامل كريمر ف} = \frac{\sqrt{\text{كاي تربيع}}}{\sqrt{\text{مجموع الملاحظات} \times (\text{عدد الصفوف أو الأعمدة} - 1)}} \quad (٨, ٤)$$

جدول (٦, ٤). تفسير معامل كريمر ف.

درجات الحرية	حجم التأثير		
	صغير	متوسط	كبير
1 (2 × 2)	.10	.30	.50
2 (2 × 3 أو 3 × 2)	.07	.21	.35
3 (2 × 4 أو 4 × 2)	.06	.17	.29

بال تطبيق على النتيجة السابقة في المعادلة (٧, ٤) يجري حساب معامل كريمر ف وفق الآتي:

$$\text{معامل كريمر ف} = \sqrt{\frac{85.25}{(1-2) \times 100}} = 0.923 \quad (٩, ٤)$$

يمكن تفسير القيمة الناتجة 0.923 على أنها ذات حجم تأثير كبير جداً. تُراوح معامل كريمر ف بين (0) و(1) ويُوضّح الجدول (٦, ٤) التفسير الموصى به الذي اقترحه كوهن (Cohen 1988) للقيم.

(١٠) الإحصائية المشابهة لمعامل كريمر هي فاي Phi وتُكتب أحياناً Φ . تُستخدم فاي فقط لجداول 2x2 ولها نفس قيم معامل كريمر ف. الاختلاف في التسميات اختلاف تاريخي فقط.

هناك خيار آخر يُفضل في كثير من الحالات بدلاً من حجم التأثير العام، وهو نسبة الاحتمال **Probability ratio (PR)**، المعروفة أيضًا باسم **الخطر النسبي relative risk** أو **نسبة الخطر risk ratio**، وهي من معناها تعني نسبة احتمالين من الجدولة المتقاطعة، تُقارن احتمال ظهور نتيجة لغوية معينة (مثل أداة التعريف) في نوع سياق واحد مع النتيجة التي تظهر في نوع السياق الآخر. نظرًا لأن حجم التأثير هذا هو نسبة من قيمتين فهي تتناسب مع جداول النتائج المتقاطعة البسيطة (2 × 2) إذ لا يوجد سوى فئتين (أو مستويين) لكل متغير. تُحسب هذه النسبة على النحو الآتي:

$$\text{نسبة الاحتمال} = \frac{\text{احتمال ناتج الاكتراث في السياق 1}}{\text{احتمال ناتج الاكتراث في السياق 2}} \quad (١٠, ٤)$$

تُحسب احتمالية أداة التعريف وأداة التنكير في نوعين من السياق كما يظهر في الجدول (٧، ٤)^(١١)، وذلك بقسمة القيم في الخلايا على مجاميع الصف.

على عكس البحوث التجريبية في الدراسات الطبية^(١٢) التي يعتني فيها الباحثون بنوع واحد من التأثير (نتيجة تحليل عينة الدراسة في مقابل المجموعة المرجعية: العينة كلها)، يُمكننا في لغويات المدونة الحاسوبية حساب أكثر من نسبة احتمال واحدة حسب محور التحقيق لدينا. لنسب الاحتمالات أربعة خيارات بناءً على الجدول (٧، ٤):

- نسبة احتمال أداة التنكير في سياق غير محدد مقابل أداة التنكير في سياق محدد = $\frac{0.962}{0.027} = 36$
- نسبة احتمال أداة التنكير في سياق محدد مقابل أداة التنكير في سياق غير محدد = $\frac{0.027}{0.962} = 0.03$
- نسبة احتمال أداة التعريف في سياق غير محدد مقابل أداة التعريف في سياق محدد = $\frac{0.038}{0.973} = 0.04$
- نسبة احتمال أداة التعريف في سياق محدد مقابل أداة التعريف في سياق غير محدد = $\frac{0.973}{0.038} = 25.3$

(١١) يجري الإقرار عادة عن احتمالات مقياس الأرقام العشرية بين (0) و(1)، وهي الأرقام المثوية نفسها بناءً على مجموع الصفوف المضروبة في ١٠٠.

(١٢) مصطلح (الخطر النسبي) أو (نسبة الخطر) هو تسمية بديلة لـ (نسبة الاحتمالية) وتُستخدم في الدراسات الطبية والوبائية التي تُجرى تحليلاتها على نسبة خطورة الأمراض.

جدول (٤, ٧). الاحتمالات لنوع الأداة حسب التحديد السياقي.

نوع الأداة	نوع السياق	تنكير	تعريف	المجموع
غير محدد		$0.962 = 26 \div 25$	$0.038 = 26 \div 1$	26
محدد		$0.027 = 74 \div 2$	$0.937 = 74 \div 72$	74
المجموع		27	73	100

ويساوي احتمال ورود أداة التنكير في سياق غير محدد ستة وثلاثين ضعفاً من ورودها في سياق محدد، وبلغ احتمال ورودها في سياق محدد (0.03) أي ٣٪ من نسبة احتمالية ورودها في سياق غير محدد، وهو احتمال صغير جداً عند المقارنة. المقياس الذي تعمل عليه نسبة الاحتمال من صفر إلى ما لا نهاية infinity، والتفسير على النحو الآتي:

- نسبة الاحتمال (1) تعني أنه لا يوجد فرق بين السياقين.
- نسبة الاحتمال الأصغر من (1) تعني أن ظهور النتيجة اللغوية المعنوية في السياق الأول أقل احتمالاً منه في السياق الثاني.
- نسبة الاحتمال الأكبر من (1) تعني أن ظهور النتيجة اللغوية المعنوية في السياق الأول أكبر احتمالاً منه في السياق الثاني.

يُستعمل أحياناً مقياس بديل يسمى نسبة الأرجحية **odds ratio** بدلاً من نسبة الاحتمال (لمزيد من المناقشة عن هذه النسبة انظر البحث ٤, ٤). لا يتأثر تفسير نسبة الأرجحية بنسبة الاحتمال لأننا نميل إلى فهم الاحتمالات أفضل من الأرجحيات (Cohen 2000; Davies et al. 1998). لتحويل نسبة الاحتمال إلى نسبة الأرجحية، يمكننا استخدام المعادلة التالية:

$$\text{نسبة الأرجحية} = \text{نسبة الاحتمال} \times \frac{\text{احتمال ناتج الاكتراث في السياق 1}}{\text{احتمال ناتج الاكتراث في السياق 2}} \quad (٤, ١١)$$

تجدر الإشارة إلى أهمية حساب مجالات الثقة CIs (٩٥٪) لحجم التأثير حتى تتمكن من تقدير مدى احتمال التأثير في المجموع اللغوي كله.

التقرير الإحصائي: الجدولة المتقاطعة وكاي تربيع

١- ما يجب الإفصاح عنه

نستخلص النتائج التقاطعية ذات النسب المئوية (نتائج اختبار كاي تربيع) في حالات المواقف البسيطة ذات المتغير اللغوي الواحد والمتغير المستقل الواحد، وينبغي الإقرار عما يأتي. أولاً: درجات الحرية (الحاشية رقم ٣٥)، وثانياً: قيمة الاختبار، وثالثاً: قيمة الاحتمال، ورابعاً: حجم التأثير (نسبة الاحتمال أو معامل كريمر ف أو كلاهما) وخامساً: مجال الثقة (٩٥٪) لحجم التأثير.

في المواقف الأكثر تعقيداً بازدياد المتغيرات المستقلة تكون جداول النتائج المتقاطعة وحدها كافية مع وصف تفصيلي للاختلافات المثيرة للاهتمام. إذا أردنا الإقرار عن الإحصاءات الاستدلالية بجداول معقدة فينبغي تطبيق الانحدار اللوجستي (المبحث ٤، ٤).

٢- مثال الإفصاح في التقرير

• كان بين نوع السياق ونوع الأداة ارتباط كبير (كاي تربيع $(1) = 85.25$ ، قيمة الاحتمال $0.001 <$).
ظهر التأثير الكلي كبيراً حيث: معامل كريمر ف $= 0.923$: (مجال الثقة ٩٥٪ $[0.727, 1]$ ، وأداة التعريف أكثر احتمالاً للظهور في السياق المحدد بنسبة 25.3 مرة مما في السياق غير المحدد: (مجال الثقة ٩٥٪ $[3.7, 172.95]$).

عُد إلى الجدول (٤، ٢) عن الأمثلة البسيطة لجدول التقاطعات البسيطة، وإلى الجدول (٤، ٤) عن الأمثلة المعقدة لجدول التقاطعات المعقدة.

(٤، ٤) الانحدار اللوجستي

فكّر...

قبل قراءة هذا المبحث، راجع معنى المصطلحات التالية المستخدمة سابقاً في هذا الفصل. هل

يمكنك تعريفها جميعاً؟

فئة متغير category of a variable

متغير تأويلي (مستقل) explanatory variable

مستوى المتغير level of a variable

المتغير اللغوي	linguistic variable
النتيجة	outcome
المتغير الناتج	outcome variable
المتنبئ	predictor
متغير المتنبئ	predictor variable
قيمة متغير فئوي	value of a categorical variable
المتباين (للمتغير اللغوي)	variant (of a linguistic variable)

سنناقش في هذا المبحث تقنية إحصائية فعّالة تسمى الانحدار اللوجستي. استكشفنا طبيعة المتغيرات المعجمية النحوية في المبحث (٤, ٢) وكذلك بعض التقنيات السهلة نسبياً للتعامل مع التنوع المعجمي النحوي في المبحث (٤, ٣). سننظر هنا في تحليل التنوع المعجمي النحوي مع الانحدار اللوجستي، وهو أسلوب يستخدم متغيرات مستقلة (تسمى أحياناً بالمتنبئ) يمكن أن تكون فئوية ومقياسية على حد سواء من أجل تقدير تأثيرها في المتغير اللغوي (الناتج) الذي ينبغي أن يكون متغيراً فئوياً^(١٣). هذا يعني أننا سنبحث عن أي خصائص سياقية (متنبئات) لها تأثير في ورود أشكال مختلفة لمتغير لغوي كأداة التعريف أو أداة التنكير مثلاً. عند تحليل البيانات نبنى نموذجاً خوارزمياً يُمثل علاقات معقدة بين المتغيرات في البيانات، ثم نستخدم هذا النموذج لتفسير تأثير متغيرات المتنبئ (ومجموعاتها) على متغير الناتج. نظراً لأن الانحدار اللوجستي أسلوب إحصائي معقد إلى حد ما، يقدم هذا المبحث عدداً من المصطلحات والتفاصيل الفنية الجديدة، وعليه يُرجى عدم القلق إذا لم تستوعب كل تفاصيل هذا الإجراء؛ لأنّ المهم أن تعرف المبادئ الأساسية لهذا الأسلوب ولتفسير المخرجات، والانحدار اللوجستي يُحسب آلياً ويمكن إجراؤه باستخدام أداة الانحدار اللوجستي من أدوات إحصاءات لانكستر في الموقع الشبكي للمصاحب للكتاب. فيما يأتي أحد الأشكال الممكنة لمعادلة الانحدار اللوجستي للراغبين في التفاصيل:

(١٣) إن كان المتغير الناتج متغيراً مقياسياً ينبغي استخدام انحدار المربعات الصغرى العادية ordinary least squares

(OLS). انظر (Gries (2013a: 261-282) للاطلاع على شرح مفصل لهذا الإجراء.

(١٢, ٤)
$$\text{probability of outcome of interest} = \frac{e^{(\text{intercept} + b_1 X_1 + b_2 X_2 + \dots + \text{unaccounted variation})}}{1 + e^{(\text{intercept} + b_1 X_1 + b_2 X_2 + \dots + \text{unaccounted variation})}}$$
 احتمالية ناتج
= الاكتراث

سيفيد شرح هذه المعادلة ببعض المصطلحات الجديدة الآتية: (e) ثابت رياضي يساوي تقريباً 2.71828^(١٤). والحصص (الكمية الثابتة) intercept وهو قيمة المنطلق (الخط الأساس) baseline، b_1 و b_2 إلخ تسمى تقديرات أو معامل coefficients. و X_1 و X_2 ... إلخ هي متنبئات مُدخلة في النموذج. سنتناول كل هذه المصطلحات بالشرح في هذا المبحث لاحقاً. ولكن، بهذه المصطلحات الأساسية يُقدَّر الانحدار اللوجستي احتمال نتيجة معينة على أساس متنبئات فردية تُسهم في هذه النتيجة بدرجات مختلفة يكون بعضها أهم من الآخر، والانحدار اللوجستي في جوهره تصنيفٌ معقد نجمع فيه بين متنبئات ذات علاقة مثل الأنماط اللغوية والقواعد المعجمية النحوية التي تُنبئ عن الفئة التي تنتمي إليها خاصية لغوية معينة. والسؤال الذي يتطلب طرحه هنا هو الآتي: هل يُفضَّل سياقٌ ما استخدام المتغير أ (على سبيل المثال أداة التعريف) أم المتغير ب (وليكن أداة التنكير)؟ مثلنا لهذه الفكرة تمثيلاً بيانياً في الشكل (١, ٤).

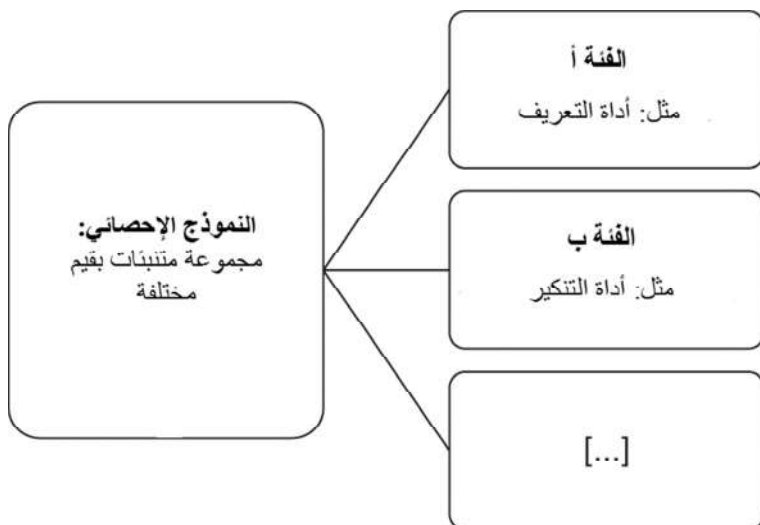
لنراجع المصطلحات الأساسية قبل الشروع في شرح تفاصيل تقنية الانحدار اللوجستي. في مهمة (فكر) استعرضنا المصطلحات المستخدمة استخداماً مترادفاً في سياق الانحدار اللوجستي المطبق على دراسة النحو المعجمي. يمكن تقسيم هذه المصطلحات إلى أربع مجموعات:

- خصائص النحو المعجمي التي هي محور البحث: المتغير اللغوي = المتغير الناتج.
- المصطلحات المتعلقة على وجه التحديد ببنية متغيرات الناتج: المتباين/ الأشكال المختلفة (لمتغير لغوي) = النتيجة.
- المتغيرات السياقية التي تساعدنا على شرح استخدام الخصائص المعجمية النحوية: المتغير المستقل = متغير المتنبئ = المتنبئ.

(١٤) يُمثّل الرمز e أعداداً غير نسبية، ولذا لا يمكن أن تحلّ محله الأعداد الكسرية أو النسبية أو الجذرية. قد تُعرف أيضاً

بوصفها أساس اللوغاريتم الطبيعي (ln).

- المصطلحات المتعلقة ببنية المتغيرات اللغوية التي تُستخدم متغيرات للناتج ومتغيرات للمتنبئ على حد سواء: ^(١٥) فئة المتغير = مستوى المتغير = قيمة المتغير الفئوي.
- يمر إجراء الانحدار اللوجستي بثلاث مراحل، الأولى: فحص البيانات (المتطلبات والافتراضات)، والثانية: بناء نموذج، والثالثة: تفسير النموذج. لتوضيح تقنية الانحدار اللوجستي سنستخدم مثال أداة التعريف التي نوقشت سابقاً (الشكل ٢، ٤).



شكل (٤، ٤). المخطط الأساسي للانحدار اللوجستي.

المرحلة الأولى: فحص البيانات (المتطلبات والافتراضات)

يلزم قبل تحليل البيانات (أي: قبل بناء نموذج) التحقق من أن مجموعة البيانات مناسبة للانحدار اللوجستي وأن افتراضات الاختبار الإحصائي مستوفاة.

أولاً: ينبغي التأكد من انتظام مجموعة البيانات وفقاً لمبادئ تصميم البحث للخاصية اللغوية (المبحث ٤، ١). هذا يعني أن ورود الخاصية اللغوية المعنوية مثل أداة التعريف وأداة التنكير يكون في سطر منفصل في كل مرة، وتُحشَى تحشيةً صحيحة بالمتغيرات المستقلة كما هو موضح في الشكل

(١٥) ملحوظة: يمكن أن تكون المتنبئات متغيرات مقياس أيضاً.

(٢, ٤). ينبغي أيضًا التحقق من أن الخاصية اللغوية تُلائم هذا النوع من البحث، وأن الإطار المعجمي النحوي قد عُرِف تعريفًا صحيحًا (المبحث ٢, ٤).

ثانيًا: ينبغي قياس المتغيرات وترميزها بدقة وثبات كما هو مشروط في كل أنواع التحليل الكمي. تُعدّ التحليلات متعددة المتغيرات (مثل الانحدار اللوجستي) حساسة وقد تُؤلّد أخطاء قياسية، وهو ما يمكن أن يكون له تأثير مضاعف بسبب المتغيرات المتعددة المستخدمة في بناء النموذج. يُرمّز العديد من أوجه المتغيرات المعجمية النحوية يدويًا، مع أهمية التأكد من اتساق الترميز اتساقًا دقيقًا، وإذا كان هناك متغير تقييمي *judgement variable* فيوصى بترميز مزدوج لجزء معين من البيانات كأن يتناول عشوائيًا ٢٠٪ من البيانات على سبيل المثال (المبحث ٥, ٣). لمزيد من النقاش عن إعداد البيانات وتهيئتها اقرأ (Osborne 2012: 195ff).

ثالثًا: نحتاج إلى الحصول على بيانات كافية، وكلما زادت المتغيرات المستقلة التي نستخدمها زاد عدد أسطر البيانات. ناقش هوسمر وآخرون (Hosmer et al. 2013: 407-408) قانون ١٠ الذي ينص على أن الحد الأقصى لعدد المعامل التي يمكن إدراجها في نموذج b_1 و b_2 إلخ في المعادلة (٤, ١٢) يُقدّر من خلال أخذ تكرار النتيجة الأقل تكرارًا وقسمته على (١٠). ولحساب عدد المعامل ننظر في مخرجات الانحدار اللوجستي لحساب المعامل وحساب عدد الصفوف في الجدول، مع ترك الحصر *intercept* كما في مجموعة البيانات الواردة في الجدول (٢, ٤). وعلى الرغم من أن النتيجة الأقل تكرارًا هي أداة التنكير بواقع ٢٧ حالة، فيمكننا إدراج معامل بحد أقصى (٢٧/١٠) في النموذج. راعى هوسمر وآخرون (2013: 408) أيضًا القيود المفروضة على هذا الحساب التقريبي بالإشارة إلى أن هناك كثيرًا من العوامل المختلفة التي تجعل من البيانات كافية في الانحدار اللوجستي. وتكون البيانات في المدونات عادة كافية لهذا النوع من التحقيق على الرغم من أن ترميز الحالات يعتمد في الأغلب قدر معين من العمل اليدوي، ويمكن أن يكون مستهلكًا للوقت. من المستحسن أيضًا تحديد متغيرات التنبيه تحديًا متأنياً بالاعتماد على الدراسات اللغوية بدلًا من إيراد أكبر عدد ممكن من المتغيرات المستقلة.

رابعًا: نحتاج إلى التحقق مما إذا كان من المنطقي بناء نموذج. إذا فسّر أو صنّف أحد المتنبئات جميع الحالات على نحو تام فسيكون بناء نموذج يعمل لتقدير تأثير المتنبئات المختلفة أمرًا غير

ضروري. هناك أيضًا مبررات رياضية لعدم إمكان إنشاء هذا النموذج (Hosmer et al. 2013: 147-8)، ففي المصطلحات الإحصائية، تسمى هذه المشكلة الفصل (شبه) الكامل (quasi) complete separation. على سبيل المثال: يتنبأ نوع الاسم (الاسم المعدود، والجمع المعدود، والأسماء غير المعدودة، والعلم) بدقة شديدة بظهور أداة التنكير، إذ يُمكن أن تظهر a/an فقط أمام الاسم المعدود. يوضح الجدول (٨، ٤) هذا الموقف من مجموعة البيانات الخاصة بنا.

خامسًا: نحتاج إلى التحقق من افتراضات الاختبار نفسه (Osborne 2015: 85-130)، فهناك ثلاثة افتراضات رئيسية: استقلال الملاحظات، وعدم ارتباط الخطية (المتعددة)، والخطية.

جدول (٨، ٤). الاحتمالات لنوع الأداة حسب نوع الاسم.

تعريف	تنكير	نوع السياق	نوع الأداة
34	27		معدود مفرد
19	0		معدود جمع
10	0		غير معدود
10	0		علم

في استقلال الملاحظات independence of observations، كما هو الحال مع اختبار كاي تربيع الموضح في المبحث (٣، ٤)، نفترض أن كل ملاحظة مستقلة عن الملاحظات الأخرى، ففي المدونات ينتقض عادةً هذا الافتراض نوعًا ما بسبب طبيعة اللغة إذ تترابط الخصائص اللغوية بسبب عينات المدونة التي تؤخذ على مستوى النصوص دون مستوى خصائص النصوص الفردية. نفترض عند دراسة التنوعات المعجمية النحوية نوعًا من التوحيد بين النصوص من النوع نفسه، وعليه فلا داعي لأن نقلق بشأن هذه المشكلة كثيرًا. يمكن إجراء ذلك بأخذ عينة فرعية عشوائية من الخصائص اللغوية من المدونة، وبالنسبة للدراسات التي تحتاج إلى مراعاة عينة النصوص أو اللغة المنطوقة تُستخدم طريقة تُسمى نمذجة التأثير الممتزج mixed-effect modelling (المبحث ٥، ٦).

في عدم ارتباط الخطية (المتعددة) بين المتنبئات (multi)collinearity between predictors، يتميز ارتباط الخطية بقوة الارتباط ($r = 0.8$ فصاعداً) بين متغيرات المتنبئ (المبحث ٢، ٥). ينبغي التحقق من ذلك لكل من المتغير المقياسي scale والفئوي categorical^(١٦). يعتمد الارتباط الخطي collinearity على قياس التشابه بين المتغيرات المرتبطة ومن ثم لا نحتاج فيه إلى الأنواع الأخرى من المتغيرات. إذا واجهنا ارتباطاً خطياً (متعدداً) فيمكننا استبعاد المتغيرات أو دمجها لتجنب مشكلة التشابه. ينتمي الانحدار اللوجستي في الخطية إلى مجموعة من الأساليب الإحصائية التي تستخدم النمذجة الخطية. هذا يعني أن العلاقة بين المتغيرات منحدره خطياً (المبحث ٢، ١). وليست كل العلاقات بين المتغيرات خطية، فعلى سبيل المثال تمثل الانحنائية الخطية تغير اللغة بمرور الوقت، ويمكن التعبير عن ذلك على أنه انحدار خطي غير مستقيم curvilinear (راجع المبحث ٥، ٧ لمزيد من مناقشة النماذج غير الخطية). نفترض في الانحدار اللوجستي أن هناك علاقة خطية بين متنبئات المقياس والاحتمالية اللوغارتمية لمتغير الناتج (سنشرح الاحتمالية اللوغارتمية لاحقاً). ولمزيد من التفاصيل عن اختبار هذا الافتراض انظر: (Hosmer et al. 2013: 94ff) و (Filed et al. 2012:344-5).

المرحلة الثانية: بناء نموذج

بعد التحقق من البيانات والافتراضات يمكننا التركيز في تحليل البيانات (بناء النموذج). الخطوة الأولى هي تحديد قيم خط الأساس أو المنطلق للمتغيرات الفئوية بما في ذلك متغير الناتج. قيمة المنطلق لمتغيرات المقياس هي دائماً (0)، لذلك لا نحتاج إلى تحديدها أصلاً. قيم خط الأساس أو المنطلق baseline values (وتسمى أحياناً مستويات المرجع reference levels) هي قيم يُقارن على أساسها النموذج آثار المتنبئ/المتنبئات على النتيجة، وتعد هذه القيم غير أساسية في التحليل اللغوي. إذا كان سؤال بحثنا هو (في أي السياقات تُستخدم أداة التعريف في اللغة الإنجليزية؟) فإن أداة التعريف ستكون ناتج الاكتراث outcome of interest أو الناتج الذي يهم، ولذلك عيّنا أداة التنكير بوصفها قيمةً منطلق. مثال آخر: إذا افترضنا استناداً إلى الدراسات اللغوية أن السياقات المحددة تفضل أدوات التعريف فسنختار السياقات غير المحددة كقيمة منطلق لمتغير المتنبئ.

(١٦) لإجراء اختبار الارتباط مع المتغيرات الفئوية، نستبدل تسميات الفئات بالأرقام، على سبيل المثال غير محدد = ٠؛

محدد = ١.

نصيحة عملية: من المفيد التمييز بين قيم المنطلق والقيم الأخرى للمتغيرات الفئوية باستخدام أحرف السوابق (أ، ب، ج)، تُرمز قيمة المنطلق بالسابقة أ (على سبيل المثال: أ_نكرة)، أما القيم الأخرى فتُرمز بأحرف سوابق أخرى (مثل: ب_محدد). يجري ذلك تلقائيًا بحزم البرامج بما فيها أداة الانحدار اللوجستي في Lancaster Stats Tools وتُدخل البيانات في النموذج أبعديًا. يظهر في الشكل (٥، ٤) مثال على مجموعة بيانات جاهزة للتحليل باستخدام تقنية الانحدار اللوجستي، ولدينا فيها ثلاثة متنبئات (نوع_السياق، ونوع_الاسم، وطول_المركب.الاسمي) ومتغير ناتج واحد (نوع_الأداة). ومع استثناء (طول_المركب.الاسمي) الذي يقيس عدد الأحرف فجميع المتغيرات فئوية (أي: اسمية) ويُوضع قبلها رمزًا من حرف واحد.

م	نوع_السياق	نوع_الاسم	طول_المركب.الاسمي	نوع_الأداة
1	أ_غير_محدد	د_عَلَم	29	ب_تعريف
2	ب_محدد	ب_معدود_جمع	15	ب_تعريف
3	ب_محدد	ج_غير_معدود	7	ب_تعريف
4	ب_محدد	ب_معدود_جمع	7	ب_تعريف
5	ب_محدد	ج_غير_معدود	27	ب_تعريف
6	ب_محدد	أ_معدود_مفرد	14	ب_تعريف
7	أ_غير_محدد	أ_معدود_مفرد	4	أ_تنكير
8	ب_محدد	د_عَلَم	17	ب_تعريف
9	أ_غير_محدد	أ_معدود_مفرد	51	أ_تنكير
10	ب_محدد	أ_معدود_مفرد	26	ب_تعريف
11	ب_محدد	أ_معدود_مفرد	3	ب_تعريف
12	ب_محدد	أ_معدود_مفرد	4	ب_تعريف
13	ب_محدد	أ_معدود_مفرد	6	ب_تعريف

شكل (٥، ٤). عينة من استخدام الأداة في اللغة الإنجليزية من مجموعة بيانات.

تُقرّر في الخطوة التالية المتغيرات التي يتعين إدخالها في النموذج، ويكمن هدف ذلك في إيجاد نموذج يفسر أكبر قدر ممكن من التنوع في البيانات وبأقل عدد ممكن من المتغيرات. يُسمى هذا أحياناً النموذج المقتر *parsimonious model*. لذلك نريد أن نقتصر في النموذج على المتغيرات ذات الصلة فقط. يمكن تحقيق ذلك بطريقتين مختلفتين. يمكننا بناءً على أبحاث ونظريات لغوية تخمين المتغيرات التي ينبغي إدخالها وتحليلها ثم الاقتصار على المتغيرات التي لها تأثير دالّ إحصائي، ويُسمى هذا المدخل الثابت *block entry*. يمكن أيضاً اللجوء إلى خيار المدخل التدريجي *stepwise entry* فنسمح للبرنامج الإحصائي بإضافة المتغيرات أو إزالتها خطوة خطوة حتى يحصل التحسين التام لمعيار المعلومات المسمى معيار معلومات أكايكي (AIC) *Akaike information criterion* للنموذج. من ضمن هذا الإجراء التدريجي يمكننا أن نتخيّر من بين الإجراء الأمامي والخلفي والمختلط. يبدأ الإجراء الأمامي دون وجود متغيرات متنبئة في النموذج، ويظل يضيف المتغيرات المتاحة واحداً تلو الآخر، ويبدأ الإجراء العكسي بنموذج يتضمن جميع المتغيرات المتاحة ويحذفها واحداً تلو الآخر. يجمع الإجراء المختلط بين الإجراءين الأمامي والخلفي إضافةً وحذفاً للمتنبئات بناءً على إعادة تقييم النموذج في كل مرحلة. تشمل مزايا المدخل الثابت التحكم الكامل للمحلّل في العملية مع اتخاذ القرار اللازم بشأن المتغيرات التي ينبغي تضمينها بناءً على أسس نظرية بدلاً من إجراء آلي غير معتمد على نظرية لغوية. يُطبّق الإجراء التدريجي في التحليلات التي تُكشّف اللغة في حال عدم الاعتماد على نظرية لغوية فلسفية (Osborne 2015: 251-3; Hosmer et al. 2013: 93-4). يكون إجراء (المدخل الثابت) عادة هو الأفضل في بحث المدونة المعجمي النحوي.

باستخدام مجموعة البيانات في الشكل (٥، ٤) عن استخدام أداة التعريف والتنكير، لدينا العديد من الخيارات لبناء نموذج. تُسرد الخيارات الأهم في الجدول (٩، ٤)، ويُسمّى النموذج (0) نموذج المنطلق (أو الصفر) *baseline (or null) model* الذي لا يتضمن أي متنبئات، ويستخدم بوصفه مرجعاً لنماذج أكثر تعقيداً. يُمكن تحسين هذا النموذج الذي لا يتضمن أي معلومات عن السياقات التي ترد فيها المتغيرات اللغوية بإضافة متنبئات ذات صلة، ويحتوي النموذج الأول على متنبئ واحد فقط (نوع_السياق) في حين يشتمل النموذج الثاني والثالث على اثنين من المتنبئات (نوع_السياق إضافة إلى متنبئ آخر). إذن فأَي النماذج نختار؟

توفر الحزم الإحصائية آلية المقارنة بين أي نموذج نبيه وبين نموذج المنطلق. إذا لم يكن أداء نموذجنا أفضل من أداء نموذج المنطلق فسيكون عديم الجدوى إلى حد كبير؛ لأن المتنبات المحددة لن يكون لها تأثير يُذكر وعليه نستبعد النموذج. يُمكن في الجدول (٩، ٤) أن نرى أفضلية النموذج الأول والثالث على نموذج المنطلق. تُثبت الدلالة الإحصائية باختبار الاحتمالية اللوغارتمية (المعروف أيضاً باسم اختبار نسبة الاحتمالية) الذي يُعدّ مقياساً معروفاً في لغويات المدونة الحاسوبية (المبحث ٣، ٤ و ٤، ٣). أما النموذج الثاني فيُظهر أخطاء معيارية كبيرة للغاية لنموذج الانحدار.

جدول (٩، ٤). نظرة عامة عن النماذج.

اسم النموذج	الناتج	المتنبات المتضمنة	النتيجة بناءً على ناتج البرنامج الإحصائي (انظر: المرحلة الثالثة: تفسير النموذج)
الصفر أو المنطلق	نوع_الأداة	[لا شيء]	نموذج المنطلق مع الحصر فقط
الأول	نوع_الأداة	نوع_السياق	ذو دلالة إحصائية، أي: أفضل بكثير من نموذج المنطلق، ومعيار معلومات أكايكي = 30.86
الثاني	نوع_الأداة	نوع_السياق ونوع_الاسم	أخطاء المعيار الكبيرة - حدث خطأ ما، ويُسبب نوع_الاسم مشكلة مع الفصل الكامل
الثالث	نوع_الأداة	نوع_السياق، وطول_المركب.الاسمي	ذو دلالة إحصائية، أي: أفضل بكثير من نموذج المنطلق؛ ومع ذلك، ليس أفضل بكثير من النموذج الأول، ومعيار معلومات أكايكي = 31.91 (أكبر من معيار معلومات أكايكي للنموذج الأول)

جدول (١٠، ٤). أخطاء المعيار الكبيرة لجزء من ناتج الانحدار اللوجستي.

خطأ المعيار	التقدير (الأرجحية اللوغارتمية)	
4530.376	-21.056	(الحصر intercept)
4530.376	23.889	نوع_السياق.ب_محدد
6706.381	18.733	نوع_السياق.ب_معدود_جمع
9244.108	18.733	نوع_الاسم.ج_غير.معدود
8958.692	39.961	نوع_الاسم.ج_غير.معدود

يُظهر الجدول (١٠, ٤) جزءًا من ناتج الانحدار اللوجستي مع الأخطاء المعيارية، وكما نرى فإن الأخطاء القياسية أكبر بعدة مرات من التقديرات، وهذه إشارة إلى أن هناك خطأ ما في النموذج. سنشرح معنى كل من الأخطاء والتقديرات المعيارية (المعامل) أدناه عندما نناقش تفسير النموذج. في حالة النموذج الثاني تتمثل المشكلة في الفصل الكامل للمتنبئ: (نوع_الاسم)، وعليه يمكننا استبعاده عن التحليل كي نركز في مقارنة النموذج الأول مع النموذج الثالث.

يبدو أن النموذج الذي يحتوي على المزيد من المتنبئات أفضل، لكن ذلك غير صحيح. يمكننا مقارنة النموذجين بالطريقة نفسها التي قارنا بها نماذجنا بنموذج المنطلق لأجل معرفة ما إذا كان أحدهما أفضل بكثير من الآخر. إضافة إلى الدلالة الإحصائية التي تُقاس من خلال اختبار الاحتمالية اللوغارتمية نستخدم معيار معلومات أكايكي AIC لتحديد النموذج الأكثر فعالية من خلال الوصول إلى الدلالة بأقل عدد ممكن من المتغيرات. يُحسب معيار معلومات أكايكي على النحو الآتي:

$$\text{معيار معلومات أكايكي} = \text{التنوع غير المفسر من خلال النموذج} + 2 \times (\text{عدد فئات}) \quad (١٣, ٤)$$

النتائج - 1 + عدد المتنبئات

سيُفسَّر معيار معلومات أكايكي بمقارنة نموذجين يعتمدان على مجموعة بيانات لغوية واحدة، بمعنى: كلما صغر حجم معيار معلومات أكايكي كان النموذج أفضل. المنطلق واضح نوعًا ما فنحن نأخذ التنوع في البيانات التي لا يفسرها النموذج، ثم ننتقد النموذج على استخدام العديد من المتنبئات (كلما زادت المتنبئات زاد قصور النموذج). إذن فالنموذج الناجح هو نموذج مقتر ومختصر، ويفسر أكبر قدر ممكن من التنوع مع وجود أقل عدد ممكن من المتنبئات. لاحظ أن معيار معلومات أكايكي ليس مقياسًا موحدًا لحجم التأثير، إذ لا يمكن استخدامه لمقارنة النماذج القائمة على مجموعات البيانات المختلفة. في الجدول (٩, ٤) فالنموذج الثالث ليس أفضل من النموذج الأول، وقيمة معيار معلومات أكايكي التي تساوي (30.86) للنموذج الأول أصغر من النموذج الثالث الذي تساوي قيمته (31.91)، وبناء على ذلك سنختار النموذج الأول ذي المتغير الواحد.

نحتاج في هذه المرحلة الأخيرة إلى أن نقرر ما إذا كنا نريد أن ندرج فقط ما يسمى التأثيرات الرئيسية main effects للمتنبئات المحددة أو التفاعلات interactions بين المتنبئات، والتفاعلات

هي مجموعة سياقات لمتنبئ محدد يتضمن المتغير اللغوي، ويمكن أن يكون لها في بعض الحالات تأثير كبير في النتيجة بالرغم من أن المتنبئات الفردية ذات السياق الواحد ليست كذلك. إذا لم تُختبر التفاعلات، فقد لا يُكتشف هذا التأثير، وعند اختيار المتغير ينبغي أن يستند إدراج التفاعلات إلى الدارسات اللغوية ذات العلاقة.

المرحلة الثالثة: تفسير النموذج

يُنتج الانحدار اللوجستي ناتجاً يتكون من جزأين رئيسيين، الأول: ملخص النموذج، والثاني: المعامل (التقديرات). على سبيل المثال: يظهر ناتج أساسي للنموذج الأول وهو نموذجنا المفضل، لاحظ أن هذا الشكل المحدد من الناتج يصدر من أداة الانحدار اللوجستي في أدوات إحصائيات لانكستر، وقد يختلف إلى حد ما عن نواتج الحزم الإحصائية الأخرى.

النموذج الأول

ملخص النموذج:

اختبار نسبة الاحتمالية (الاحتمالية اللوغارتمية): دالٌّ > (p < 0.0001)

دليل الكشف C-index: 0.96 - > بارز

مقياس ناجيلكيركي R^2 : 0.86

معيّار معلومات أكايكي: 30.87

المعامل coefficients

التقدير (الأرجحية اللوغارتمية)	خطأ المعيّار	قيمة ز (اختبار زاي والد)	قيمة الاحتمال	التقدير (الأرجحية)	مجال الثقة الأقل	مجال الثقة الأعلى
(الخصر intercept)	-3.219	-3.156	0.002	0.04	0.002	0.189
نوع_السياق_ب_محدد	6.802	5.457	0.000	900	116.878	21421.229

بالنظر إلى ملخص النموذج يمكننا أن نرى أن للنموذج دلالة إحصائية، إذ بلغت قيمة اختبار نسبة الاحتمال $89.79^{(١٧)}$ وبقية احتمالية منخفضة جداً ($p < 0.0001$). هذا يعني أن النموذج الأول بالمتنبئ الواحد (نوع_السياق) أفضل بكثير من نموذج المنطلق بدون متنبئات. يمكن استخدام مقياس أخرى مثل دليل الكشف والنسخ المختلفة من إحصائية R^2 الزائفة لتحسين تقييم أداء النموذج. دليل الكشف السياقي^(١٨) هو مقياس لنجاح تصنيف النموذج وتحديد مدى إمكانه بالتنبؤ بالنتيجة [the $a(n)$ أو مثالنا]. نحن نبحت عن قيمة دليل كشف أعلى من $0.7^{(١٩)}$ (Hosmer et al. 1977: 177). في بعض الأحيان تُدرج في التقرير أيضًا قيم R^2 الزائفة مثل ناجيلكيركي Nagelkerke. تسعى هذه المقاييس إلى تصوير مقدار التنوع في البيانات التي أوضحها النموذج. يُراوح نطاق إحصائية R^2 الزائفة بين 0 (لم يُوضح أي تنوع) و 1 (وُضحت جميع التنوعات) لسوء الحظ، وعليه فإن مقياس R^2 الزائفة ليست دائمًا موثوقة (Osborne 2015: 51).

لنركز الآن في تأثير المتنبئات الفردية بالنظر في الجزء الثاني من الناتج المسمى (المعامل). تُعرض المعامل في جدول. يعرض الصف الأول دائمًا ما يسمى الحصر (أو الثابت) intercept، وهو قيمة منطلق في النموذج تُقدر الموقف الذي تكون فيه جميع المتنبئات عند قيم المنطلق. تذكر أنه بالنسبة للمتنبئات الفئوية مثل (نوع_السياق) فهي القيم التي حددناها بوصفها منطلقًا. بالنسبة لمتنبئات المقياس تكون القيمة دائمًا 0 (راجع المرحلة الثانية: بناء نموذج المشروحة سابقًا). في

(١٧) لاحظ أنه سيُحصل على القيمة نفسها البالغة (89.79) بإجراء الاحتمالية اللوغارتمية وفقًا للجدول (٢، ٤) لأن أبسط حالات الانحدار اللوجستي مع أحد المتنبئات مماثلة لإجراء اختبار الاحتمالية اللوغارتمية. لاحظ أيضًا أنه قد حصلنا على قيمة مشابهة جدًا تبلغ (85.25) في البحث (٣، ٤) عند إجراء اختبار كاي تربيع في الجدول (٢، ٤). (١٨) تشمل التسميات الأخرى للمقياس نفسه مؤشّر الدقة ومؤشّر جيني ومؤشّر منطقة ما تحط المنحنى area under the curve AUC.

(١٩) إذا زاد المؤشر على 0.8 فإن النموذج ممتاز في التصنيف. أما إن زاد على 0.9 (وهو أمر نادر الحدوث) فإن نجاح التصنيف فائق جدًا.

(٢٠) على الرغم من استخدام دليل الكشف C-index على نطاق واسع فقد انتقد مؤخرًا، واقترح إجراء آخر وهو مقياس H-index (Hand 2010). النقاش متخصص إلى حد ما، وما زلت أوصي بالإفصاح عن دليل الكشف مع التحذير، بحيث لا يوجد مقياس واحد يمكنه تمثيل النجاح التنبئي للنموذج في جميع الحالات.

النموذج الأول يقدّر الحصر أرجحيات أداة التعريف التي ترد في السياق الأساسي المنطلق (غير المحدد). الأرجحيات ضئيلة جداً: 0.04 وبنسبة مجال ثقة من ٩٥٪ تقدر بـ: 0.002 و 0.189).

دعونا نتوقف للحظة ونشرح الوحدات المستخدمة لقياس تأثير المتنبئات، وهي الأرجحيات odds وأرجحيات اللوج (log odds). تُحسب الأخيرة بأخذ اللوغاريتم الطبيعي للأرجحية. يعمل الانحدار اللوجستي بأرجحيات اللوج. نظراً لصعوبة تفسير أرجحيات اللوج نسبياً نُحوّلها عادةً إلى مجرد أرجحية odds. تُعرّف الأرجحية على النحو الآتي:

$$\text{الأرجحية} = \frac{\text{احتمالية الناتج } 0}{\text{عدم احتمالية الناتج } 0} = \frac{\text{احتمالية ناتج الاكترات في سياق المنطلق}}{\text{احتمالية ناتج المنطلق في سياق المنطلق}} \quad (١٤, ٤)$$

تُستخدم الأرجحية في التطبيق العملي عادةً في المراهنات الرياضية، ففيها نسأل أسئلة من قبيل: ما أرجحية فوز فريقنا بالمباراة؟ إذا كانت الأرجحية من 2 إلى 1، فإن احتمال فوز فريقنا أكبر بمقدار ضعف (أي 66.7%) من احتمال خسارته (33.3%). في مثالنا السابق في الجدول (٧، ٤) نُحسب أرجحية قيمة المنطلق (الحصر)، أي: أداة تعريف ترد في سياق غير محدد، على النحو الآتي:

$$\text{الأرجحية (الحصر)} = \frac{\text{احتمالية ورود the في السياق غير المحدد}}{\text{احتمالية ورود a(n) في السياق غير المحدد}} = \frac{0.038}{0.962} = 0.04 \quad (١٥, ٤)$$

لاحظ أن هذه القيمة تطابق قيمة أرجحية الحصر في جدول الناتج في المعامل في العمود السادس (تقدير الأرجحية). يُركّز في بقية الصفوف في جدول المعامل في الناتج، لأنها توجب على سؤال البحث حول تأثير المتنبئات المختلفة، أي: متغيرات متنبئ المقياس القائمة التي تُسمى المتغيرات الصورية dummy variables (مجموعة المتنبئات الفئوية وقيمها غير منطلقية). يُظهر الصف (نوع السياق المحدد ب) تأثير السياق المحدد على استخدام أداة التعريف. يُقاس هذا التأثير في نسب أرجحيات اللوج التي لها فائدة في إجراءات الانحدار المنطقي ونسب الأرجحية التي لها فائدة في تفسير التأثير بإجراء مقارنة مع قيمة المنطلق. تُحسب نسبة الأرجحية على النحو الآتي:

$$\text{نسبة الأرجحية} = \frac{\text{احتمالية ناتج الاكترات في السياق المعني}}{\text{احتمالية ناتج المنطلق في السياق المعني}} = \frac{\text{الأرجحية (قيمة المتنبئ المعني)}}{\text{الأرجحية (الحصر)}} \quad (١٦, ٤)$$

تشير نسبة الأرجحية الأكبر من (1) إلى أن أرجحية ناتج الاكتراث الذي يرد في السياق المعني أكبر من أرجحية الناتج نفسه الذي يرد في سياق المنطلق. أما نسبة الأرجحية الأصغر من (1) فتُظهر استبعاد السياق المعني عن ناتج الاكتراث. إذا كانت نسبة الأرجحية هي (0) فلن يكون هناك أي تأثير. في مثالنا في الجدول (٧، ٤) تحسب نسبة الأرجحية التي توضح تأثير نوع السياق المحدد في استخدام أداة التعريف على النحو الآتي:

$$(١٧، ٤) \quad \text{نسبة الأرجحية} = \frac{\text{احتمالية a(n) في السياق المحدد}}{\text{احتمالية a(n) في السياق غير المحدد}} = \frac{\frac{0.973}{0.027}}{\frac{0.038}{0.962}} = \frac{36}{0.04} = 900$$

وتظهر هذه القيمة من صف جدول الناتج الثاني: (نوع_السياق.المحدد_ب) في العمود السادس (تقدير الأرجحية)، وعليه فقيمة أرجحية ورود أداة التعريف في السياق المحدد هي 900 مرة (مع مجال ثقة ٩٥٪ [117, 21.421] وهي أكبر من أرجحية ورود أداة التعريف في سياق غير محدد. بالنسبة لكل صف، يسرد جدول الناتج أيضًا أخطاء المعيار **standard errors** واختبار زاي والد **Wald's z** للدلالة الإحصائية وقيمة الاحتمال المقابلة التي نخبرنا عما إذا كان التقدير المحدد ذا دلالة إحصائية. تعبّر أخطاء المعيار عن مدى دقة التقديرات، وتكون أخطاء المعيار الأصغر هي الأفضل. يُحسب إحصاء زاي والد بقسمة التقدير مع الخطأ المعياري. حجم التأثير لكل بارامتر هو نسبة الأرجحية التي نوقشت سابقاً، والتي تستكمل مجالات الثقة ٩٥٪ التي تبين لنا حجم ثقة نسبة الأرجحية في عينة المجموع اللغوي.

في النماذج الأكثر تعقيداً التي تتضمن متنبئات عديدة نقيم التقديرات واحداً تلو الآخر. يوضح الجدول الآتي الإضافة إلى معامل الناتج عند تضمين مقياس المتنبئ: (طول_المركب.الاسمي) في النموذج الذي يقيس طول المركب الاسمي.

التقدير (الأرجحية اللوغارتمية)	خطأ المعيار	قيمة ز (اختبار زاي والد)	قيمة الاحتمال	التقدير (الأرجحية)	مجال الثقة الأعلى	مجال الثقة الأقل	طول_المركب.الاسمي
0.037	0.039	0.939	0.348	1.037	0.966	1.138	

نلاحظ أن (طول_المركب.الاسمي) ليس مؤشراً مهماً ($p > 0.05$) ونسبة الأرجحية فيه صغيرة جداً، والقيمة (1.037) قريبة من (1)، وهو ما يعني عدم وجود تأثير. تؤكد هذه الملاحظة من خلال مجال الثقة الذي يتضمن بالفعل (1). هذه علامة على أن النتيجة غير دالة إحصائية؛ لأن التأثير في المجموع اللغوي قد يكون معدماً (نسبة الأرجحية 1). السبب في إظهار هذا الناتج هو مناقشة معنى نسبة الأرجحية في حالة متنبئ المقياس. تشير نسبة الأرجحية الخاصة بمتنبئ المقياس إلى حجم زيادة أرجحية ناتج الاكتراث (إذا كانت > 1) أو حجم نقصانها (إذا كانت < 1) مقارنةً بأرجحية ناتج المنطلق مع تغيير وحدة واحدة من متنبئ المقياس. في الجدول السابق كانت وحدة طول المركب الاسمي حرفاً واحداً، لذلك يشير التقدير (1.037) إلى أن أرجحية أداة التعريف تزيد بهذا المقدار مرة واحدة عند كل إضافة حرف إلى المركب الاسمي. هذا يعني أنه كلما طال المركب الاسمي زاد احتمال ظهور أداة التعريف. لوحظ أن هذا التأثير ليس ذا دلالة إحصائية، وعادة يُستبعد من النموذج لأنه لا توجد أدلة كافية على ذلك في البيانات.

يعد الانحدار اللوجستي طريقة قوية تتيح لنا نظرة تفصيلية عن تأثيرات السياقات المختلفة على الناتج اللغوي. تجدر الإشارة إلى أننا في هذا المبحث اكتشفنا نوعاً من الانحدار الخطي يسمى الانحدار اللوجستي ثنائي الحد **binomial logistic regression** وهو انحدار منطقي يكون متغير الناتج فيه ذا فئتين، مثل: a/an/the. تُطبق هذه الحالة النموذجية في البحوث النحوية المعجمية للتحقق من قوة أثر خاصيتين لغويتين داخل الإطار المعجمي النحوي. إذا كان لمتغير الناتج أكثر من فئتين تُستخدم تقنية ماثلة تسمى الانحدار اللوجستي متعدد الحدود **multinomial logistic regression**. يتبع هذا النوع من الانحدار مبادئ الانحدار ثنائي الحد لكن المقارنات أكثر تعقيداً إلى حد كبير (انظر: Field et al. 2012: 8.9.1-8.9.2 و Arppe 2008 للاطلاع على مثال لمثل هذا الإجراء).

التقرير الإحصائي: الانحدار اللوجستي

١- ما يجب الإفصاح عنه

نظراً لأن الانحدار اللوجستي إجراء معقد فإن النجاح في استخدامه يعتمد على عدد من الخطوات. ينبغي تحديد هذه الخطوات في مدة وجيزة في تقريرنا البحثي للسماح بالتناجحية (بتماثل النتائج). أولاً:

ينبغي أن نوضح كيف عُرِفَ الإطار المعجمي-النحوي وما المتغيرات المستخدمة ولماذا. ثانياً: ينبغي أن نسمح للقراء بمعرفة كيفية الحصول على البيانات (على سبيل المثال: عينة فرعية عشوائية لجميع تكرارات الخصائص اللغوية من مدونة ما) وكيفية ترميزها، وكذلك ما إذا كان أي جزء من البيانات مزدوج الترميز، مع ضرورة الإفصاح عن إحصاء ملاءمة المقدّر البيني. ثالثاً: نحتاج إلى توفير تفاصيل عن الإحصائيات الإجمالية للنموذج (الاحتمالية اللوغارتمية، وقيمة الاحتمال، ودليل الكشف)، إضافةً إلى جدول بالمعامل الفردية بما في ذلك معلومات الدلالة الإحصائية ونسب الأرجحية ومجال الثقة ٩٥٪ لنسب الأرجحية.

٢- مثال الإفصاح في التقرير

• نظرًا لأن محور البحث قد كان عن التنوع بين أداة التعريف وأداة التنكير، فقد عُثِرَ على جميع مرات ورودهما في المدونة، ثم اختيرت مئة حالة اختياريًا عشوائيًا، وُرْمِزَت من حيث وجود أداة تعريف (متغير الناتج) أو عدم وجودها. بعد ذلك جرى قياس متغيرين (متنبئين) سياقيين ورد في الأدبيات أن لهما تأثيرًا في استخدام الأدوات في اللغة الإنجليزية، وهما: نوع السياق وطول المركب الاسمي.

• كان نوع السياق مؤثرًا ذا دلالة لنوع الأداة المستخدمة، وقد حَسِّنَ إدخال السياق المحدد في النموذج كثيرًا من ملاءمة النموذج {الاحتمالية اللوغارتمية = 89.79، $p < 0.0001$ }. يحتوي النموذج أيضًا على خصائص تصنيف بارزة (دليل الكشف = ٩٥٪). لم يكن لطول المركب الاسمي تأثير كبير. ويتضح من جدول المعامل أدناه أنَّ استخدام أداة التعريف يكون أكثر أرجحيةً في السياق المحدد (نسبة الأرجحية = 900، ٩٥٪ مجال الثقة [21، 117، 421]) منه في السياق غير المحدد.

التقدير (الأرجحية اللوغارتمية)	خطأ المعيار	قيمة ز (اختبار زاي والد)	قيمة الاحتمال	التقدير (الأرجحية)	٩٥٪ مجال الثقة أقل	٩٥٪ مجال الثقة أعلى
(الخصر intercept)	-3.219	1.020	-3.156	0.002	0.04	0.189
نوع_السياق.ب_محدد	6.802	1.247	5.457	0.000	900	21421.229

(٤, ٥) التطبيق: Which أم That؟

Lexico-grammatical variables as discussed in this chapter are variables which can be expressed as choices between two or more variants, such as the definite and the indefinite article.
 [الترجمة: المتغيرات النحوية-المعجمية كما نوقشت في هذا الفصل متغيرات يمكن التعبير عنها بوصفها خيارات بين اثنين أو أكثر من المتغيرات، مثل أداة التعريف وأداة التنكير.

شكل (٤, ٦). صُححت جملة من هذا الكتاب نحويًا.

واجهتُ في أثناء كتابة هذا الفصل الخطأ المتذبذب تحت مركب تضمن ضمير الوصل which، وهو ما يشير إلى وجود خطأ نحوي محتمل أو عدم دقة (الشكل ٤, ٦). كان التصحيح المعروض هو إما إضافة فاصلة أمام which وإما استخدام ضمير الوصل that بدلاً من ذلك على أن يكون المنطق على النحو الآتي: إذا كانت هذه الكلمات غير ضرورية لمعنى جملتك استخدم (which) وافصل الكلمات بفاصلة (Microsoft 2010).

بعد التغلب على الاستياء الذي حصل من جراء تصحيح جهاز الحاسوب القواعد النحوية، بدأت أفكر في كيفية اختبار ما إذا كانت القاعدة التي طبقها المدقق النحوي تعكس بالفعل كيفية استخدام اللغة. أجريْتُ دراسة باستخدام المدونتين الفرعيتين BE06 و AmE06، تضمنت كلٌّ منهما مليون كلمة إنجليزية مكتوبة.

تستند هذه الدراسة إلى مدونة BE06 وهي مدونة متوازنة من اللغة الإنجليزية البريطانية المعاصرة، ومدونة AmE06 وهي مدونة متوازنة من اللغة الإنجليزية الأمريكية المعاصرة. تتكون كل مدونة من أربعة أوعية رئيسية مكتوبة (النثر العام والرواية والصحف والكتابة الأكاديمية).

في مدونتي BE06 و AmE06 هناك 4,736 حالة لورود which و 22,749 حالة لورود that. في حين أن which تُستخدم في الأغلب ضمير وصل relativizer في سياقات مثل تلك الموجودة في الشكل (٤, ٦) فإن that لها وظائف عدة أخرى (محدد، أو اسم إشارة، أو توكيدية، إلخ). في هذه الدراسة حُددت الاستخدامات المعجمية النحوية ذات الصلة لكل من which و that (والتبادل المحتمل بينهما) باستخدام وسوم أجزاء الكلام المناسبة (DDQ لـ which و CST لـ that من مجموعة وسوم CLAWS7^(١)) وذلك

(٢١) وسوم أجزاء الكلام هي تسميات نحوية متصلة تلقائيًا بالكلمات باستخدام واسم أجزاء الكلام. لمزيد من

حسب الحالة النحوية للكلمات التالية لهما في الجملة. اقتصّر على النظر في الحالات التي يكون فيها which و that قابلين للتبادل من جهة المبدأ^(٢٢).

إذا لم تكن الكلمات ضرورية لمعنى الجملة فافصل بينها وبين which بفاصلة. لمعالجة أسلوب الفصل طريقتان، الأولى شكلية، والثانية وظيفية. تتطلب الطريقة الشكلية فصل which بفاصلة، أما الوظيفية فتتطلب استخدام which في الحالات التي تكون فيها العبارة المُصدّرة بـ which غير ضرورية للمعنى العام للجملة، ومن ثمّ يمكن الاستغناء عنها، وهو ما يُصطلح عليها نحوياً بـ: العبارة غير المقيّدة (أو الفصلة).

جدول (١١، ٤). الجدولة المتقاطعة لاستخدام الفاصلة مع ضميري الوصل which و that.

المجموع	بلا فاصلة	بفاصلة (سواء علامة فاصلة أو شرطة)	ضمير الوصل
			ظهور الفاصلة
2,200	804 (37%)	1,396 (63%)	Which
7,472	7,281 (97%)	191 (3%)	That
9,672	8,085	1,587	المجموع

هذا يعني ضمناً أن ضمير الوصل الآخر that يتعين استخدامه في التوزيع التكميلي للضمير which الذي يكون في العبارات التقيدية (العُمدية) غير مفصول بفاصلة. لاختبار كل من الجانب الشكلي والوظيفي سنضع سؤالين بحثيين:

- السؤال البحثي الأول (الشكلي): هل which مسبوق بفاصلة أو شرطة (-) وهل that غير مسبوق بفاصلة أو شرطة (-)؟
- السؤال البحثي الثاني (الوظيفي): ما العوامل التي تؤثر في استخدام which و that؟

(٢٢) استُبعدت الحالات التي يلي فيها that الاسم الذي يشير إلى شخص أو مجموعة من الأشخاص For a guy that did

(well he dressed down (BE06_K16) لكون محل that هو اسم الموصول who وليس which.

تُعرض نتائج البحث في مدونتي BE06 و AmE06 التي تجيب عن السؤال الأول في الجدول (١١، ٤). لاحظ أن الجدول يعرض نتائج عمليات البحث التلقائية في المدونة دون التحقق اليدوي (انظر السؤال البحثي الثاني).

تُستخدم that (7,472) أكثر بكثير من which (2,200) في العبارات المحورة (الوصفة) للاسم. يُفصل ما يقرب من ثلثي العبارات المستخدم فيها which بفاصلة أو شرطة، أما that فتسبقها فاصلة أو شرطة فقط في ٣٪ من الحالات. عند النظر عن كثب في هذه النسبة نلاحظ أن الأغلبية العظمى من هذه الحالات تتضمن مركباً اعتراضياً قبل عبارة that، مثل:

The communications circuit proposed by Robert Darnton rightly identifies many factors, besides authorship, that govern any reading experience (BE06_J02)

المركب الاعتراضي مركبٌ يقطع الاتصال الرئيسي للجملة وعادةً ما يكون مفصلاً بفاصلتين من كلا الجانبين، وعليه فإن الفاصلة التي نلاحظها تنتمي إلى المركب الاعتراضي، وليس العبارة الموصولة نفسها.

أكد اختبار كاي تربيع أن هناك ارتباطاً كبيراً بين ضمير الوصل (which أو that) ووجود فاصلة أو شرطة ($X^2(1) = 4595.47, p < .001$). التأثير الكلي هنا كبير، ومعامل كيرمر ف = 0.689، ومجال الثقة ٩٥٪ [0.669, 7.09]. الفاصلة أو الشرطة قبل which أكثر عرضة للظهور وبواقع 24.8 مرة (مجال الثقة ٩٥٪ [21.5, 28.7]) من that. وعلى عكس ذلك فعدم استخدام الفاصلة قبل that أكثر عرضة للظهور وبواقع 2.7 مرة (مجال الثقة ٩٥٪ [2.5, 2.8]) من which. هذه العلاقة المعقدة مصوّرة في الشكل (٧، ٤).

يمكننا باختصار ملاحظة تفضيل واضح لعدم استخدام فاصلة قبل that، وقد تُستخدم أولاً تستخدم الفاصلة قبل which، على الرغم من وجود تفضيل لاستخدام فاصلة أو شرطة قبلها. لا يمكن تضمين كل حالات which بوصفها حالات فتوية جاءت مفصولة بفاصلة أو شرطة؛ لأن عدد الحالات التي لم تكن ذلك قد بلغ ٨٠٤ حالات كما في الجدول (١١، ٤).

شكل (٧, ٤). تمثيل العلاقة بين *which* و *that* والفاصلة.

للإجابة عن السؤال البحثي الثاني، رُمّزت مجموعة فرعية مكونة من 360 كشافاً سياقياً عشوائياً مأخوذاً من 9672 حالة لكل من *which* و *that* (الجدول ١١, ٤). رُمّز يدوياً عددٌ من المتغيرات السياقية مثل مستوى تنويع اللغة الإنجليزية والفاصلة ونوع العبارة ونوع التركيب، ثم قيس ورُصد المتغير (طول الجملة الموصولة) الذي جاء بعد *which* و *that*. على سبيل المثال، سُرّمز الجملة التالية على النحو الآتي: التنويع: بريطاني، والفاصلة: لا توجد، ونوع العبارة: مقيدة، والموقع النحوي: فاعل^(٢٣)، والطول: ٥ كلمات.

Apple will just veto and refuse to distribute any application which does not meet its terms. (BE06_E33)

[الترجمة]: ستستخدم أبل حق النقض وترفض توزيع أي تطبيق لا يفي بشروطها

يسرد الجدول (١٢, ٤) جميع متغيرات الناتج الفئوية المدرجة وتكرارات *which* أو *that* في السياقات المحددة بمتغيرات الناتج. لا يمكن إجراء جدولة متقاطعة لمتغير (الطول) الذي يجري التعامل معه على أنه مقياس.

(٢٣) هو هنا فاعل العبارة الموصولة. انظر *Still, it was his only answer, which he repeatedly struck*

(AmE06_K15)، حيث *which* هي المفعول به.

من الجدول (١٢، ٤) يمكننا ملاحظة أن الصورة معقدة للغاية ولا يسهل فهم جميع العوامل بمجرد النظر. لذا استُخدم تحليل الانحدار اللوجستي لتحديد العوامل ذات الصلة التي لها أثرٌ في اختيار which أو that.

يعد النموذج الذي يتضمن جميع متغيرات المتنبئ (النوع، والفاصلة، ونوع العبارة، والموقع النحوي، والطول) دالاً إحصائياً، فالاحتمالية اللوغارتمية: 222.31؛ $p < 0.0001$ ، وللنموذج أيضاً خصائص تصنيف واضحة (دليل الكشف: 0.91).

جدول (١٢، ٤). الجدولة المتقاطعة لـ which و that في مواقف سياقية مختلفة.

التنوع	الفاصلة (، أو-)	العبارة	الموقع النحوي	ضمير الوصل		المجموع
				which	that	
أمريكي	لا	غير تقييدية	مفعول	3	1	4
			فاعل	11	2	13
		تقييدية	مفعول	18	2	20
			فاعل	126	5	131
	نعم	غير تقييدية	مفعول	0	6	6
			فاعل	0	20	20
		تقييدية	مفعول	0	0	0
			فاعل	1	0	1
بريطاني	لا	غير تقييدية	مفعول	3	2	5
			فاعل	3	8	11
		تقييدية	مفعول	14	8	22
			فاعل	76	15	91
	نعم	غير تقييدية	مفعول	0	4	4
			فاعل	0	31	31
		تقييدية	مفعول	1	0	1
			فاعل	0	0	0
المجموع				256	104	360

يعرض الجدول (١٣، ٤) المعامل الفردية في النموذج. يمكننا ملاحظة أنه باستثناء المتبني: (الموقع.النحوي.ب_فاعل)، فإن جميع التقديرات ذات دلالة ($p < 0.05$)، ومن ثمّ فالمتغيرات: (التنوع البريطاني، ووجود فاصلة، والورود في عبارة غير مقيدة، وطول العبارة الأكبر) قد جاءت كلها لمصلحة which من حيث الدلالة^(٩). على سبيل المثال، أرجحية استخدام which في اللغة الإنجليزية البريطانية هي 5.3 (بمجال الثقة 95٪ [2.5, 12.1]) وهي ضعف أرجحية استخدامها في اللغة الإنجليزية الأمريكية. وبالمثل، فأرجحية استخدام which بعد الفاصلة هي 53.8 (بمجال الثقة 95٪ [12.9, 376.4]) وهي ضعف استخدامها بدون فاصلة. أيضاً، كلما طالت الجملة زاد احتمال استخدام which، وتزداد الأرجحية بنسبة 1.1 مرة (بمجال الثقة 95٪ [1.02, 1.15]) مع كل كلمة إضافية.

نكتفي بهذا القدر من التحليل، ويبقى سؤال واحد ملح: هل كانت المعالجة الحاسوبية دقيقة؟ يعتمد الجواب على التعامل مع طبيعة فئوية الحالة لكل من which و that. أوضح التحليل أن هناك مجموعة من العوامل المتعددة التي تفضل أو التي تستبعد استخدام which و that، وينبغي تفسير هذه العوامل كلها مهما كانت على أنها احتمالات وأرجحيات على وجه الدقة، وليست يقينية.

جدول (١٣، ٤). تقديرات الانحدار اللوجستي لكل من which و that.

	التقدير (الأرجحية اللوغارتمية)	خطأ المعيار	قيمة ز (اختبار زاي والد)	قيمة الاحتمال	التقدير (الأرجحية)	مجال الثقة أقل	مجال الثقة أعلى
(الخصر intercept)	-3.354	0.563	-5.958	0.000	0.035	0.011	0.099
التنوع.ب_بريطاني	1.667	0.397	4.195	0.000	5,296	2.511	12.080
الفاصلة.ب_نعم	3.985	0.825	4.832	0.000	53.795	12.876	376.448
العبارة.ب_غير تقييدية	2.046	0.446	4.588	0.000	7.733	3.235	18.812
الموقع.النحوي.ب_فاعل	-0.614	0.421	-1.460	0.144	0.541	0.240	1.260
الطول	0.079	0.029	2.739	0.006	1.083	1.023	1.147

(٢٤) لاحظ أن ضمير الوصل that، والتنوع أمريكي، وبدون فاصلة، والجملة تقييدية وفي موضع المفعولية من حيث النحو، قد وُضعت على أنها قيم المنطلق للمتغيرات المعنية.

(٤, ٦) تمارين

١ - انظر موضوعات البحث وأسئلته في الجدول (٤, ١٤). حدد ما إذا كان تصميم بحث الخاصية اللغوية مناسباً (نعم/ لا). إذا كانت الإجابة بنعم فحدد الإطار المعجمي النحوي المناسب. ملأنا لك الصف الأول بوصفه مثالاً.

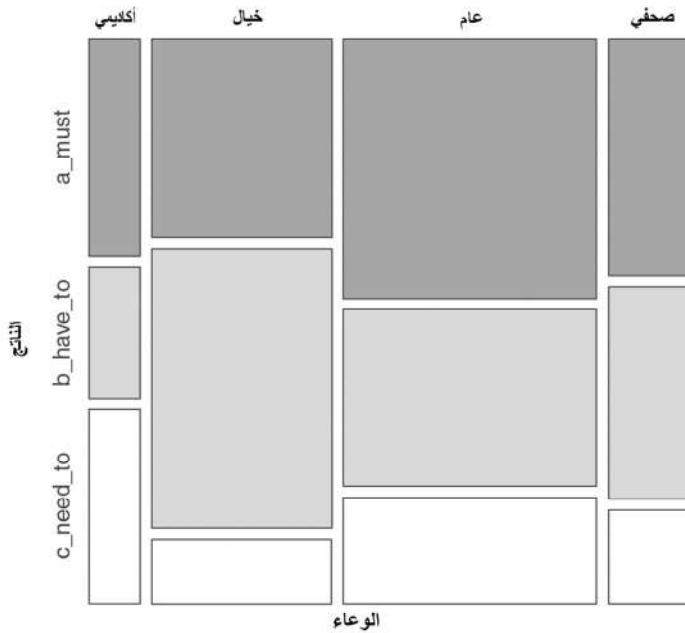
جدول (٤, ١٤). تصميم البحث المناسب.

الموضوع: سؤال البحث	الخصائص اللغوية (أمثلة)	هل صُمم وفق الخاصية اللغوية؟	الإطار المعجمي النحوي
تناوب المفعول به غير المباشر: ما العوامل المؤثرة في تناوب المفعول به غير المباشر في اللغة الإنجليزية؟	She handed <u>the student</u> the book أعطت <u>الطلاب</u> الكتاب She handed the book to <u>the student</u> . أعطت <u>الكتاب</u> للطلاب	نعم	جميع تركيبات المفعول به غير المباشر
تناوب أداة التنكير (a) و (an): متى تُستخدم النسخة غير القياسية من أداة التنكير (a) قبل كلمة تبدأ بصوت متحرك في اللغة الإنجليزية المنطوقة؟	an apple, a apple		
الكلمات النابية: هل يستخدم المتحدثون المزيد من الكلمات النابية القوية أو الضعيفة؟	fuck, cunt, motherfucker etc. vs damn, crap, hell etc.		
التناوب الإضافي: ما العوامل المؤثرة في الاختيار بين s- و of الإضافية؟	president's speech, the speech of the president كلام الرئيس		
العلامات المعرفية: هل تدعم بيانات المدونة الفرضية القائلة: "نحن نقول إننا متأكدون عندما لا نكون كذلك" (هاليداي)؟	This is <u>certainly</u> the case هذه هي القضية بكل تأكيد This is <u>maybe</u> the case هذه هي القضية ربما		
this للحاضر أو للغائب: ما العوامل المؤثرة على وجود اسم بعد this؟	This is an example هذا مثال This sentence is an example هذه الجملة مثال		

٢- حلل الجدولة المتقاطعة الآتية بإضافة مجاميع الصفوف والأعمدة بالإضافة إلى المجموع الكلي. هل هناك فرق في استخدام must و have to و need to بين الإنجليزية البريطانية والأمريكية؟ أضف النسب المئوية التي ستساعدك على الإجابة عن هذا السؤال، واحسب اختبار كاي تربيع بناء على التكرارات المطلقة في الجدول.

الأفعال الناقصة				
المجموع	need to	have to	must	التنوع
	201	355	352	أمريكي
	190	405	448	بريطاني
				المجموع

٣- فسر المخطط في الشكل (٨، ٤). يعرض المخطط استخدام ثلاثة تعبيرات مساعدة (ناقصة) من الإلزام الجازم (must و have to و need to) في مدونة BE06، وهي مدونة تتألف من مليون كلمة بالإنجليزية البريطانية المكتوبة. الأوعية من هذه المدونة هي: الكتابة الأكاديمية، والرواية، والنثر العام، والصحافة.



شكل (٨، ٤). must (يجب) و have to (يلزم أن) و need to (يحتاج إلى) في الإنجليزية البريطانية (BE06).

٤- فسر النماذج التالية لنتائج الانحدار اللوجستي عن دراسة التعبيرات المساعدة (الناقصة) الدالة على الإلزام الجازم. استرشد بالسؤال البحثي: في أي سياق يستخدم المتحدثون must بدلاً من have to و need to؟

تم ترميز المتغيرات التالية في مجموعة البيانات:

- متغير الناتج: must مقابل المنطلق: have to و need to.
 - المتنبئ الأول (التنوع): البريطانية مقابلة الأمريكية (المنطلق).
 - المتنبئ الثاني (الوعاء): الرواية مقابل النشر العام مقابل الصحافة مقابل الكتابة الأكاديمية (المنطلق).
 - المتنبئ 3 (الفاعل): I مقابل فاعل آخر (المنطلق).
- النموذج الأول: التأثير الرئيسي (variety) إحصائيات النموذج الإجمالية:

اختبار نسبة الاحتمال اللوغارثمي LL: 3.52 (p = 0.061) - < غير دال

دليل الكشف: 0.52 - < غير مقبول

مقياس ناجيلكيركي ر²: 0

معيّار معلومات أكايكي: 2641.65

	التقدير (الأرجحية اللوغارتمية)	خطأ المعيّار	قيمة ز اختبار زاي والد)	قيمة الاحتمال	التقدير (الأرجحية)	%٩٥ مجال الثقة أعلى	%٩٥ مجال الثقة أقل
(الخصر intercept)	-0.457	0.068	-6.711	0.000	0.633	0.554	0.723
التنوع. ب_بريطاني	0.173	0.092	1.875	0.061	1.189	0.992	1.426

النموذج الثاني: التأثيرات الرئيسية: (التنوع variety) و(النوع genre) إحصائيات النموذج الإجمالية:

اختبار نسبة الاحتمال اللوغارثمي LL: 54.49 (p = 0.001) - < دال

دليل الكشف: 0.6 - < غير مقبول

مقياس ناجيلكيركي ر²: 0.04

معيار معلومات أكايكي: 2596.67

التقدير (الأرجحية) اللوغارتمية	خطأ المعيار	قيمة ز (اختبار زاي والد)	قيمة الاحتمال	التقدير (الأرجحية)	%٩٥ مجال الثقة أقل	%٩٥ مجال الثقة أعلى
(مقطف)	-0.027	0.147	-0.184	0.854	0.973	0.729
النوع.ب_بريطاني	0.163	0.094	1.738	0.082	1.177	0.980
النوع.ب_خيال	-0.870	0.165	-5.277	0.000	0.419	0.303
النوع.ج_عام	-0.172	0.157	-1.092	0.275	0.842	0.618
النوع.د_صحفي	-0.529	0.185	-2.860	0.004	0.589	0.410

النموذج الثالث: التأثيرات الرئيسية: (التنوع) و(النوع) إضافةً إلى التأثيرات المتداخلة بينهما إحصائيات النموذج الإجمالية:

اختبار نسبة الاحتمال اللوغارتمي LL: 75.54 (p = 0.001) <- دالّ

دليل الكشف: 0.61 <- غير مقبول

مقياس ناجيلكيركي ر²: 0.05

معيار معلومات أكايكي: 2581.63

التقدير (الأرجحية) اللوغارتمية	خطأ المعيار	قيمة ز (اختبار زاي والد)	قيمة الاحتمال	التقدير (الأرجحية)	%٩٥ مجال الثقة أقل	%٩٥ مجال الثقة أعلى
(الحصر intercept)	-0.027	0.147	-0.184	0.854	0.973	0.729
النوع.ب_بريطاني	-0.857	0.288	-2.979	0.003	0.424	0.240
النوع.ب_خيال	-1.581	0.238	-6.646	0.000	0.206	0.128
النوع.ج_عام	-0.578	0.225	-2.573	0.010	0.561	0.359
النوع.د_صحفي	-1.166	0.266	-4.387	0.000	0.312	0.184
النوع.ب_بريطاني النوع.ب_خيال	1.440	0.337	4.275	0.000	4.221	2.190
النوع.ب_بريطاني النوع.ج_عام	0.893	0.321	2.785	0.005	2.443	1.308
النوع.ب_بريطاني النوع.د_صحفي	1.319	0.376	3.506	0.000	3.738	1.796

النموذج الرابع: التأثيرات الرئيسية: (التنوع) و(النوع) و(الفاعل) إضافةً إلى تأثيرات (التنوع) و(النوع) المتداخلة
إحصائيات النموذج الإجمالية:

اختبار نسبة الاحتمال اللوغارثمي LL: 145.56 ($p = 0.0001$) - < دالّ

دليل الكشف: 0.66 - < غير مقبول

مقياس ناجيلكيركي R^2 : 0.1

معيّار معلومات أكايكي: 2515.61

التقدير (الأرجحية) اللوغارثمية	خطأ المعيّار	قيمة ز (اختبار زاي والد)	قيمة الاحتمال	التقدير (الأرجحية)	مجال الثقة أعلى	مجال الثقة أقل	%٩٥
(الخصر intercept)	0.542	0.200	2.714	0.007	1.720	1.168	2.561
التنوع.ب_بريطاني	-0.930	0.290	-3.210	0.001	0.395	0.222	0.693
النوع.ب_خيال	-1.318	0.242	-5.441	0.000	0.268	0.165	0.428
النوع.ج_عام	-0.450	0.228	-1.974	0.048	0.638	0.406	0.993
النوع.د_صحفي	-1.191	0.268	-4.445	0.000	0.304	0.179	0.511
الفاعل.ب_I	-1.084	0.174	-6.232	0.000	0.338	0.239	0.472
الفاعل.ج_you	-0.917	0.158	-5.794	0.000	0.400	0.291	0.542
التنوع.ب_بريطاني النوع.ب_خيال	1.482	0.340	4.353	0.000	4.400	2.267	8.620
التنوع.ب_بريطاني النوع.ج_عام	0.952	0.324	2.941	0.003	2.592	1.379	4.915
التنوع.ب_بريطاني النوع.د_صحفي	1.490	0.379	3.927	0.000	4.438	2.118	9.384

تذكّر الآتي

- نحتاج عند تحليل الاختلافات المعجمية النحوية إلى الانتباه إلى السياقات اللغوية الفردية وتحديد الإطار المعجمي النحوي.

- يمكن استخدام الجدولة المتقاطعة لإجراء تحليل يسير للمتغيرات الفئوية. يُمكن إضافةً إلى التكرارات أن تتضمن جداول النتائج المتقاطعة النسب المئوية استنادًا إلى مجموع الصفوف الأكثر فائدة للتحقيق في قواعد اللغة المعجمية ومجموع العمود والمجموع الكلي.
- يمكن تمثيل البيانات الموجودة في جداول النتائج المتقاطعة تمثيلًا فعالًا باستخدام مخططات الفسيفساء mosaic.
- يمكننا اختبار الدلالة الإحصائية للعلاقة بين المتغيرات، وفي جدولين للنتائج المتقاطعة (أي جدول به متغير لغوي واحد ومتغير مستقل واحد)، وذلك باستخدام اختبار كاي تربيع. أحجام التأثير المبلغ عنها هي معامل كريمر ف (التأثير الكلي) ونسب الاحتمالية أو الأرجحية (التأثيرات الفردية).
- الانحدار اللوجستي طريقة متطورة متعددة المتغيرات لتحليل تأثير المتنبئات المختلفة (سواء الفئوية أم المقياسية) على متغير الناتج الفئوية (عادة ما تكون ثنائية).
- في الانحدار اللوجستي، ننظر إلى كل من الأداء العام للنموذج وكذلك المعامل الفردية التي توضح تأثير متغيرات المتنبئ في نتائج الاكتراث.

قراءات إضافية

- Balakrishnan, N., Voinov, V. & Nilculin, M. S. (2013). *Chi-squared goodness of fit tests with applications*. Waltham, MA: Academic Press.
- Friendly, M. (2002). A brief history of the mosaic display. *Journal of Computational and Graphical Statistics*, 77(1), 89-107.
- Geisler, C. (2008). Statistical reanalysis of corpus data. *ICAME Journal*, 32, 35-46.
- Gries, S. Th. (2013). *Statistics for linguistics with R: a practical introduction*. Berlin: De Gruyter Mouton, pp. 247-336.
- Hosmer, D. W., Lemeshow, S. & Sturdivant, R. X. (2013). *Applied logistic regression*, 3rd edn. Hoboken, NJ: John Wiley & Sons.
- Osborne, J. W. (2015). *Best practices in logistic regression*. Thousand Oaks, CA: Sage.

أدوات لانكستر الإحصائية

- ١- يمكن استنساخ جميع التحليلات المبينة في هذا الفصل باستخدام أدوات لانكستر الإحصائية. تشمل الأدوات المتاحة لهذا الفصل:
 - الجدولة المتقاطعة.
 - مقارنة الفئات.
 - آلة حاسبة للانحدار المنطقي.
- ٢- يقدم الموقع أيضًا موادًا إضافية للطلاب والمعلمين.

تنوع السجل اللغوي: الارتباط والتجميع والعوامل

(١, ٥) ما الذي يتناوله هذا الفصل؟

يناقش هذا الفصل مجموعة من الطرق التي يمكن استخدامها في التحليل المتزامن لعدد كبير من المتغيرات اللغوية في عدد متنوع من النصوص والسجلات اللغوية. سنقف عند طبيعة علاقة الارتباط بين متغيرين لغويين، ثم سنوضح الفرق بين ارتباطات بيرسون وارتباطات سيرمان غير البارامترية، ثم سننعم النظر في تصنيف الكلمات والنصوص والسجلات اللغوية باستخدام تقنية التجميعات المتكتلة الهرمية hierarchical agglomerative clustering، وقد دققنا كثيرًا في الخيارات التي سيجري بها تحديد التجميع وتفسير نتائجه بالكامل. سنتطرق بعد ذلك إلى التحليل المتعدد الأبعاد multidimensional وهو منهج يُطبق (تحليل العوامل) باستخراج أنماط المتغيرات وأبعادها الوظيفية بحسب التنوع اللغوي، وذلك عبر خطوات عدة تبدأ بعملية اختيار المتغيرات، ثم تفسيرها، ثم تمثيل أبعادها تمثيلًا بيانيًا.

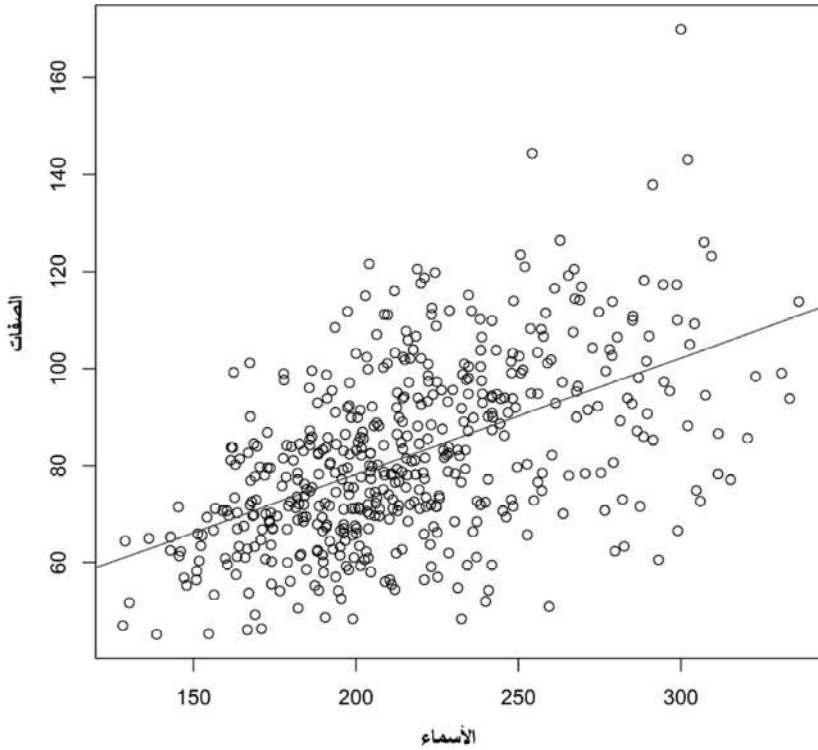
سنجيب في هذا الفصل عن الأسئلة الأربعة الآتية:

١. كيف يمكننا اختبار العلاقة بين متغيرين لغويين؟ (المبحث ٢, ٥).
٢. كيف نصنف الكلمات والنصوص والسجلات اللغوية وما إلى ذلك؟ (المبحث ٣, ٥).
٣. كيف يمكننا استكشاف أبعاد مختلفة من التنوع اللغوي؟ (المبحث ٤, ٥).
٤. كيف يمكن استخدام التقنيات التي نوقشت في هذا الفصل في البحث؟ (المبحث ٥, ٥).

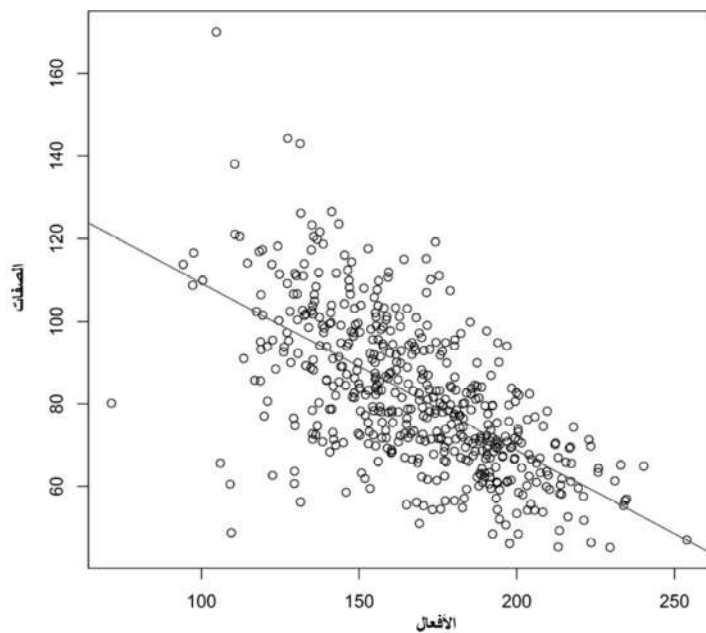
(٥, ٢) العلاقات بين المتغيرات: الارتباطات

فكّر...

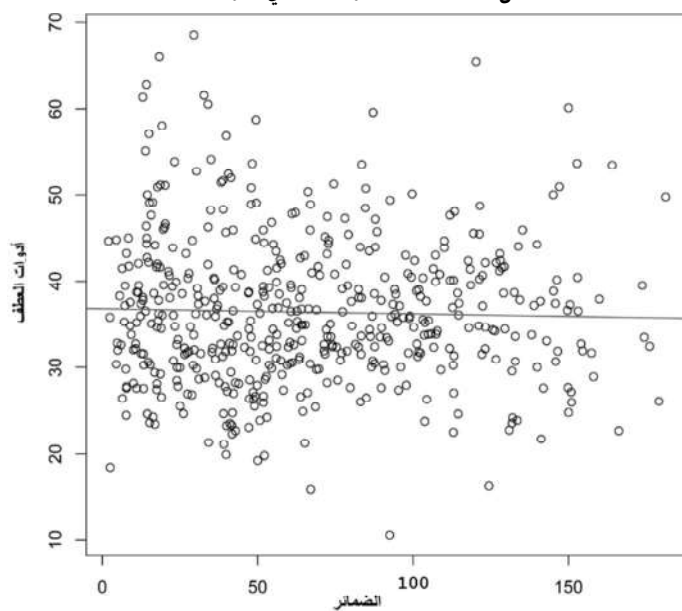
قبل قراءة هذا المبحث، انظر أشكال مخططات التشتت (١, ٥, ٢, ٥, ٣). تُمثّل في كل شكل بياني التكرارات النسبية للمتغيرين لغويين، أحدهما على المحور س والآخر على المحور ص. تمثل كل نقطة نصّاً واحداً من عدة نصوص من مدونة BE06 التي تُعدّ عينة مكونة من مليون كلمة من الإنجليزية البريطانية المكتوبة. هل بين المتغيرين اللغويين أي علاقة واضحة في كل رسم بياني؟



شكل (١, ٥). الأسماء والصفات في مدونة BE06.



شكل (٢, ٥). الأفعال والصفات في مدونة BE06.



شكل (٣, ٥). الضمائر وأدوات العطف في مدونة BE06.

ترتبط المتغيرات اللغوية بطريقة ما بفعل اجتماعها معاً في نص، فقد يرتبط التكرار النسبي للصفات في النصوص الإنجليزية بالتكرار النسبي للأسماء لأن الصفات تحوّر الأسماء ومن ثمّ تتوارد عادةً معها. لا يعني ذلك أن الأسماء لا تردّ دون صفات، إذ يُغلب على النصوص التي تحتوي على عدد كبير من الأسماء أن تحتوي أيضاً على عدد كبير من الصفات كما في الشكل (١، ٥)، ونلاحظ تزايد العدد لكل منهما تزايداً شبه متناظر كما في الخط الصاعد (انظر المبحث ٢، ١). يمكننا قياس العلاقة بين متغيرين لغويين باستخدام تقنية تسمى الارتباط. يقيس الارتباط **correlation** ما إذا كان متغيران ترتيبيّان أو متغيران مقياسيان مرتبطين بين تغايرهما (انظر المبحث ٣، ١). نبحث في حال عدد المتغيرين عما إذا كان المتغير الآخر يزيد أو ينقص أو يبقى كما هو. عندما نرى نمطاً الشكل (١، ٥) نجد أن عدد قيم المتغير الأول (الأسماء) وقيم المتغير الثاني (الصفات) يزيدان معاً، ومن ثمّ فالارتباط هنا ارتباط إيجابي **positive correlation**. أما إذا زاد عدد أحد المتغيرين (الأفعال) وانخفض الآخر (الصفات) كما في الشكل (٢، ٥) فإننا نتحدث عن ارتباط سلبي **negative correlation**. عندما نرى نقاط البيانات شديدة التبعثر بطريقة عشوائية في مجمل مساحة الشكل، كما هو في الشكل (٣، ٥) الذي يوضح التكرارات النسبية للضمائر وأدوات العطف، تُصبح العلاقة عندئذ ضئيلة أو معدومة، ويُجبرنا ذلك بعدم وجود أي ارتباط إيجابي أو سلبي.

هناك نوعان أساسيان من الارتباط: ارتباط بيرسون **Pearson's correlation** وارتباط سيرمان **Spearman's correlation** (Sheskin 2007: 1219-1368). صُمم ارتباط بيرسون للعمل مع متغيرات المقياس **scale variables**، أما ارتباط سيرمان فيتعامل مع الرتب (المتغيرات الترتيبية **ordinal variables**). لاحظ أنه يمكن تحويل متغيرات المقياس إلى رتب عندما تكون بعض المعلومات مفقودة، ولا يمكن القيام بذلك عكسياً، فعلى سبيل المثال: لا يمكن بمعلومات الرتب فقط استنتاج التكرارات الملحوظة الأصلية لمجموعة من الخصائص، ولكن يمكن استخدام ارتباط سيرمان مع المتغيرات اللغوية المقياسية كتلك التي في الأشكال (١، ٥ و ٢، ٥ و ٣، ٥) التي حوّلت سابقاً إلى رتب. عندما تكون قيم المتغيرات أحياناً منحرفة إلى حد كبير، أي: يكون المتوسط نموذجاً غير جيد للبيانات، تكون قيم الاحتمال المرتبطة بارتباط بيرسون غير موثوق بها، ويبنّ إيجل ونون (Edgell & Noon 1984) أن استعمال ارتباط بيرسون مع القيم المنحرفة انحرافاً كبيراً ممكن مع المتغيرات اللغوية المقياسية إن كان الاستعمال افتراضياً غير أساسي.

كيف يُحسب الارتباط؟ يمكن التعبير عن ارتباط بيرسون (ر) بالنحو الآتي:

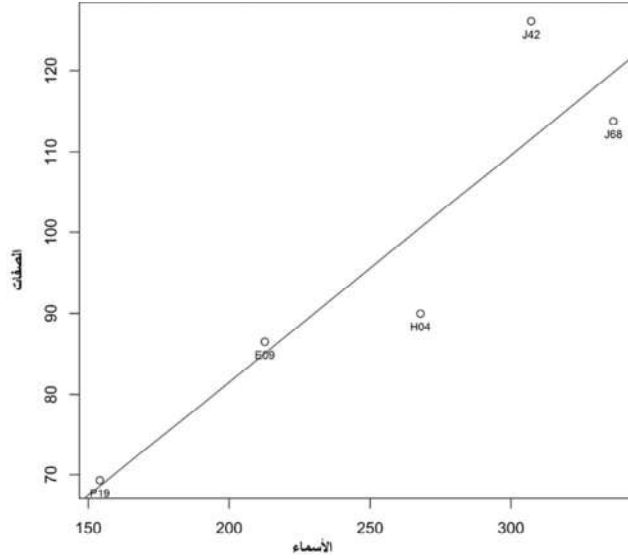
$$r = \frac{\text{التغاير}}{\sqrt{\text{الانحراف المعياري 1} \times \text{الانحراف المعياري 2}}} \quad (٥, ١)$$

يعبر هذا الارتباط عن مقدار التغاير (التنوع الذي تشترك فيه المتغيرات) في البيانات من جهة الانحرافين المعياريين (الانحراف المعياري ١ والانحراف المعياري ٢) للمتغيرين قيد التحليل. يُمنج الانحرافين المعياريين هنا بوصفهما وحدة قياس موحدة لـ (المسافة الإحصائية)، ويُستخرج التغاير بحساب متوسط المتغيرين (المتوسط ١ والمتوسط ٢) مع أخذ كل قيمة مفردة للمتغير الأول وحساب المسافة الإحصائية من المتوسط ١ وضربها في المسافة بين المتغير ٢ والمتوسط ٢. يُعبر عن هذا بواسطة المعادلة الآتية:

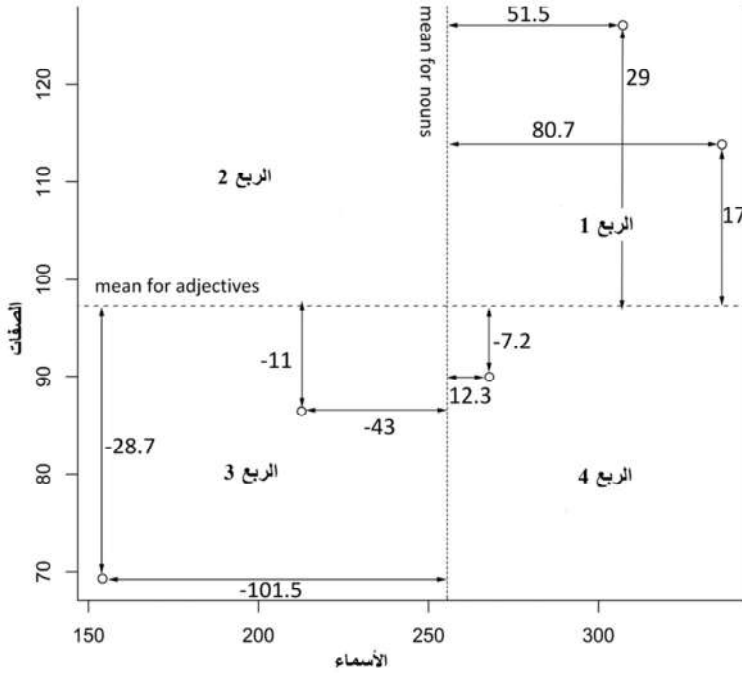
$$\text{التغاير} = \text{covariance} = \frac{\text{مجموع المسافات المضروبة من المتوسط 1 والمتوسط 2}}{\text{إجمالي الحالات} - 1} \quad (٥, ٢)$$

قد تبدو العملية معقدة لكن الفكرة سهلة للغاية. لتوضيحها أكثر، نستخلص خمسة نصوص من المستخدمة في إنتاج الشكل (٥, ١) ونركز على العلاقة بين الأسماء والصفات فيها. يوضح الشكل (٥, ٤) وجود ارتباط إيجابي واضح بين الأسماء والصفات في هذه النصوص الخمسة المختارة. يوضح الشكل (٥, ٥) كيفية قياس التغاير باستخدام نقاط البيانات الخمس نفسها. بالنسبة لكل نقطة بيانات (ممثلة بدائرة) نقيس المسافات بين المتوسطين. يُمثّل المتوسطان بالخط العمودي والخط الأفقي المنقطين بنقاط على شكل شرط. تُضرب هذه المسافات أولاً ثم تُجمع. يمكننا ملاحظة أن جميع نقاط البيانات - ما عدا نقطة البيانات H04 - موجودة في الربع الأول أو الربع الثالث الذي أنشأه المتوسطان على شكل خطوط منقطة بنقاط تشبه الشرطات.

هذا يعني أن المسافتين من المتوسطين إما أن تكونا موجبتين معاً (الربع الأول) وإما سالبتين معاً (الربع الثالث) ومن ثم فإن نتيجة ضرب المسافتين ثم المجموع الكلي هو رقم موجب. أما إذا كانت معظم نقاط البيانات في الربعين الثاني والرابع (كما في الشكل ٥, ٢) فستكون إحدى المسافتين موجبة والأخرى سالبة، وستكون نتيجة الضرب والجمع رقماً سالباً. إذا كانت نقاط البيانات مبعثرة عشوائياً في جميع الأرباع الأربعة (كما في الشكل ٥, ٣) فسنحصل على كلٍّ من النتائج الإيجابية والسلبية لعملية الضرب التي يلاشي بعضها بعضاً عند جمعها معاً وسيؤدي ذلك إلى التغاير بالقرب من الصفر.



شكل (٤, ٥). خمس نقاط بيانات للارتباط.



شكل (٥, ٥). التباين للارتباط.

يمكن حساب تغاير الأسماء والصفات في النصوص الخمسة بالشكل (٥, ٥) على النحو الآتي:

$$(٥, ٣) \quad 1518 = \frac{(-27.8 \times -101.5) + (-11 \times -43) + (12.3 \times -7.2) + (51.5 \times 29) + (80.7 \times 17)}{5.1} = \text{التغاير}$$

نستخرج من المعادلة التالية ارتباط بيرسون (ر) بقسمة نتيجة التغاير على نتيجة ضرب الانحراف المعياري^١ والانحراف المعياري^٢، وقد تطرقنا في الفصل الثاني إلى حساب الانحرافات المعيارية في المعادلة (١١, ٢).

$$(٥, ٤) \quad 0.9 = \frac{1518}{73.3 \times 23} = ر$$

في هذه الحالة يكون الارتباط بين الأسماء والصفات في النصوص الخمسة موجباً (كما هو واضح من الشكل ٤, ٥)، والنتائج الذي ظهر كبير جداً: (0.9) ويُرواح معامل الارتباط دائماً من (-1) إلى (+1)، وتشير الأرقام السالبة إلى الارتباطات السلبية، وتشير الأرقام الموجبة إلى الارتباطات الإيجابية، أما الصفر فيدل على عدم وجود علاقة خطية بين المتغيرين المعنيين. تُستخدم القيم الفاصلة التالية من أجل التقريب للإشارة إلى حجم الارتباط (Cohen 1988: 79-80) وهي مقاييس لحجم التأثير^(١):

• (0): لا تأثير

• (±0.1): تأثير صغير

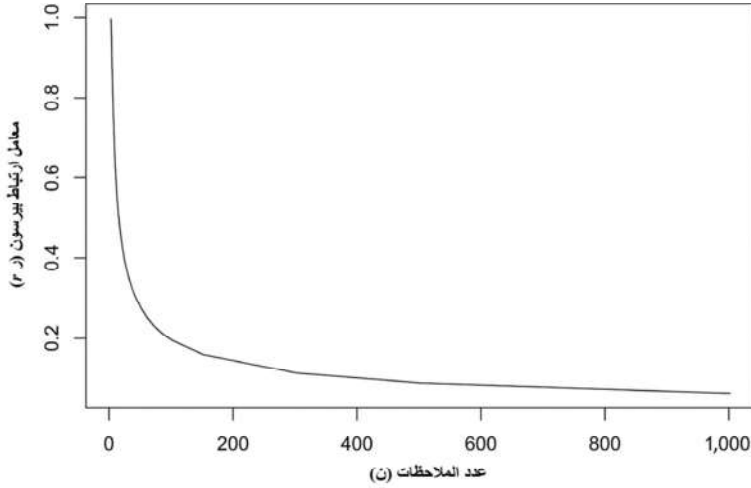
• (±0.3): تأثير متوسط

• (±0.5): تأثير كبير

ينبغي استكمال معامل الارتباط بقيمة الاحتمال أو مجال الثقة (CI) للإشارة إلى ما إذا كان هناك دليل كافٍ في المدونة لتعميم العلاقة على المجموع اللغوي. قيمة الاحتمال هي نتيجة اختبار يقيّم الفرضية الصفرية التي تنص على أن الارتباط في المجموع اللغوي يساوي (0) أي: لا يوجد ارتباط. ترتبط الدلالة الإحصائية للعلاقة ارتباطاً مباشراً بعدد الملاحظات (الحالات). في حال وجود أعداد كبيرة من الحالات في أبحاث المدونة تكون الارتباطات الصغيرة ذات دلالة إحصائية

(١) ينبغي عدم تطبيق التفسير المقترح لحجم تأثير (ر) آلياً؛ بل ينبغي استخدامه دليلاً عاماً مع الانتباه إلى الآثار العملية لقيم حجم التأثير المختلفة بالنسبة للحقل العلمي. لمزيد من المعلومات انظر البحث (٨, ٤).

كما يتضح من الشكل البياني في الشكل (٣، ٥)، وهذه الارتباطات ليست مهمة دائماً من المنظور اللغوي. على سبيل المثال: إذا كان الارتباط ضئيلاً بين الضمائر وأدوات العطف كما في هذا الشكل، ويساوي (-0.029)، فهو ليس دالاً إحصائياً مع ٥٠٠ نص، ولكنه دالٌ إحصائياً مع ٥٠٠٠ نص، ومن ثمّ تصبح قيمة الارتباط نفسها ذات دلالة إحصائية عند مستوى $p < 0.05$ على الرغم من أن التأثير اللغوي فعلياً لا يكاد يذكر.



شكل (٦، ٥). الدلالة الإحصائية ($p < 0.05$) لارتباطات بيرسون وعلاقتها بعدد المشاهدات (الملاحظات).

يوضح الشكل (٦، ٥) العلاقة بين عدد الملاحظات وارتباطات بيرسون ($P < 0.05$) ذات الدلالة الإحصائية، والاتجاه السائد في هذا الشكل هو أنه بزيادة عدد الملاحظات نحتاج إلى ارتباطات أصغر للوصول إلى الدلالة. على سبيل المثال: إذا كان لدينا عشر ملاحظات فقط من نصوص مكتوبة أو حوارات فنحن بحاجة إلى ارتباط كبير قدره (0.63) للوصول إلى الدلالة الإحصائية، وهذا يعني أن جميع نقاط البيانات تحتاج إلى إظهار العلاقة بوضوح في حالة وجود كمية صغيرة فقط من التغيرات المقبولة من أجل اعتبار الأدلة مرضية لإجراء استنتاج موثوق عن المجموع اللغوي. في حالة وجود ١٠٠ ملاحظة يكون المزيد من التغير مقبولاً، وتنخفض القيمة الحرجة الضرورية لارتباط بيرسون إلى (0.2). في حالة وجود ١٠٠٠ ملاحظة سيكون هناك ارتباط ضئيل بقيمة (0.06) وذو دلالة إحصائية. من المنظور اللغوي فإن للارتباط بقيمة 0.06 (أو 0.2 هنا) تأثيراً

عملياً ضئيلاً للغاية من جهة قوة العلاقة بين المتغيرات. لا يزال من المهم الإفصاح عن هذا الارتباط بوصفه نتيجة للتحليل لأنه يمكن أن يكشف عن كيفية عمل اللغة، وعليه فالارتباطات هي مثال آخر على كيفية قياس أحجام التأثير (معامل الارتباط r) والدلالة الإحصائية (قيمة الاحتمال p -value) لشئيين مختلفين للغاية لا ينبغي الخلط بينهما (انظر: ٤، ١ و ٤، ٨ للاطلاع على مناقشة مفصلة عن هذه النقطة).

لنعد إلى مثالنا الخاص بالنصوص الخمسة من مدونة BE06 المعروضة في الشكل (٤، ٥)، فنظراً لأن النصوص تُظهر العلاقة بين المتغيرات بوضوح شديد فإن قيمة الاحتمال المرتبطة بارتباط بيرسون هي 0.033. يشير هذا إلى أنه على الرغم من وجود خمس حالات فقط فإن الارتباط (0.9) ذو دلالة إحصائية. وبعبارة أخرى: لدينا أدلة كافية في البيانات لرفض الفرضية الصفرية التي تنص على أن الارتباط في المجموع اللغوي يساوي (0). إذا كنا نريد أن نكون أكثر تحديداً ونقدّر مدى صغر حجم الارتباط أو كبره المحتمل في المجموع اللغوي فنحن بحاجة لحساب مجالات الثقة $CI^{(٣)}$ للارتباط. لا تحتاج إلى معرفة التفاصيل $^{(٣)}$ الرياضية لأن الآلة ستعالجها بسهولة بالغة. لذلك لنركز في كيفية تفسير النتائج. يُراوح مجال الثقة ٩٥٪ في مثالنا من (0.128) إلى (0.994). لاحظ أنه لا يتضمن (0) الذي يتماشى مع قيمة الاحتمال ذات الدلالة. يشير هذه المجال إلى أنه مهما كان الارتباط محدداً في المجموع اللغوي فإنه سيكون إيجابياً على أي حال بعد الصفر. يحتوي مجال الثقة على الأرجح على القيمة الحقيقية للارتباط في المجموع اللغوي (أي: في جميع النصوص موجودة). هذا يعني أنه من خلال ما لدينا من أدلة (خمسة نصوص فقط) لا يمكننا التأكد من أن الارتباط قد ارتفع فعلياً إلى 0.9 لأنه يمكن أن ينخفض إلى 0.128، أو يمكن أن يكون أعلى أيضاً من 0.9 (0.994)، لذلك نحن بحاجة إلى التعامل مع هذه النتيجة بحذر.

(٢) إن نسيت مفهوم مجال الثقة فعد إلى البحث (١، ٣).

(٣) لمن يريد الاطلاع على المعادلة

الحد الأدنى = $z_0 - 1.96 \times \frac{1}{\sqrt{\text{عدد الملاحظات} - 3}}$ ؛ الحد الأعلى = $z_0 + 1.96 \times \frac{1}{\sqrt{\text{عدد الملاحظات} - 3}}$ حيث $z_0 = 0.51$ المجموع

اللغوي $\left(\frac{+1}{-1}\right)$ ؛ وبعد إنهاء هذه الحساب يُحوّل الحد الأدنى والأعلى إلى قيمة الارتباط (r).

يُستخدم ارتباط سبيرمان (r_s) -الذي يشار إليه أحياناً أيضاً بالحرف اليوناني (ρ) - مع البيانات الترتيبية (الرتب) أو مع بيانات المقياس عند انتقاض الافتراضات البارامترية. عند تحقق هذا الانتقاض نظراً لعدم وجود متوسطات أو انحرافات معيارية أو مسافات من المتوسط تُحوّل بيانات المقياس إلى رتب، ويُقاس التغير من خلال النظر في الاختلافات بين الرتب، ثم يُحسب ارتباط سبيرمان، وذلك على النحو الآتي:

$$r_s = 1 - \frac{6 \times \text{مجموع مربع اختلافات الرتب}}{(\text{عدد الحالات}) \times (\text{عدد مربع الحالات} - 1)} \quad (٥, ٥)$$

لنأخذ مرة أخرى النصوص الخمسة من الشكل (٤, ٥)، لكن في هذه المرة ينبغي تحويل القيم على المحورين السيني والصادي (أي تكرارات الأسماء والصفات على التوالي) إلى رتب كما في الجدول (٥, ١).

جدول (٥, ١). رتب الأسماء والصفات في خمسة نصوص من مدونة BE06.

الأسماء - الصفات		الصفات		الأسماء		
المرتبة	فرق المرتبة	المرتبة	النسبي (لكل التكرار 1000)	المرتبة	النسبي (لكل التكرار 1000)	الملف
0	4-4 = 0	4	86.5	4	212.7	BE_E09
0	3-3 = 0	3	90.0	3	267.9	BE_H04
1	2-1 = 1	1	126.1	2	307.2	BE_J42
1	1-2 = -1	2	113.8	1	336.3	BE_J68
0	5-5 = 0	5	69.3	5	154.1	BE_p19

عند إدخال الاختلافات في الترتيب التريعي من الجدول (٥, ١) وعدد الحالات خمسة في المعادلة نحصل على ما يأتي:

$$r_s = 1 - \frac{(0+1+1+0+0) \times 6}{(1-5^2) \times 5} = 0.9 \quad (٥, ٦)$$

في هذه الحالة يكون الارتباط غير البارامتري (r_s) كبيراً جداً. نستخدم القيم الفاصلة نفسها كما هو الحال بالنسبة لارتباط بيرسون، أي: 0.1 و 0.3 و 0.5 للتفسير التقليدي لقوة التأثير. ظهرت

قيمة الاحتمال أكبر من 0.05 ($p = 0.083$)، أي أننا تقليدياً نستنتج أنه ليس لدينا ما يكفي من الأدلة لرفض الفرضية الصفرية التي تشير إلى أن الارتباط في المجموع اللغوي هو (0). لا تقدم النصوص الخمسة أدلة كافية على أن ارتباط سيرمان ليس معدماً، فما نراه قد يكون نتيجة للصدفة. يمكن تفسير الفرق بين قيمة الاحتمال المرتبطة بارتباط بيرسون وتلك المرتبطة بارتباط سيرمان بغياب بعض المعلومات عند تحويل القيم الفعلية إلى رتب. إذن، فارتباط بيرسون هو الأفضل مع متغيرات المقياس، وارتباط سيرمان متناسب مع المتغيرات الترتيبية.

أولاً: يمكن استخدام معامل ارتباط بيرسون لحساب مقدار التغيرية في متغير واحد يتقاسمه المتغير الآخر، ولذلك يحتاج المعامل إلى التربيع، ويُطلق عليه: معامل التحديد **coefficient of determination** ويحسب باستخدام المعادلة البسيطة أدناه:

$$\text{معامل التحديد (ر}^2\text{)} = \text{مربع عامل ارتباط بيرسون (٥, ٧)}$$

لنأخذ، على سبيل المثال، الأسماء والصفات في مدونة BE06 (الشكل ١, ٥). معامل الارتباط r هو (0.523). والتأثير فيه كبير. عندما نقيّم هذه القيمة نحصل على 0.27 أو 27%.

$$r^2 = 0.523^2 = 0.27 \quad (٥, ٨)$$

يمكن تفسير هذا على أنه تنوع مشترك بنسبة ٢٧٪ بين التكرارات النسبية للأسماء والتكرارات النسبية للصفات في النصوص. هل ٢٧٪ عدد كبير أم صغير؟ ضع في الحسبان ما يأتي: ٢٧٪ من التنوع المفسّر يترك ٧٣٪ من التنوع غير مفسّر، ويُفسّر من متغيرات أخرى لم نأخذها في الحسبان لأننا نظرنا في الارتباط بين هذين المتغيرين فقط. تُعدّ اللغة نظاماً معقداً للغاية ذا عدد كبير من المتغيرات المهمة، لذلك إذا كان يمكن استخدام أحد المتغيرات للتنبؤ بقرابة ٣٠٪ من التنوع في متغير آخر فإن هذا أمر جدير بالملاحظة.

ثانياً: عند التعامل مع متغيرات لغوية متعددة يمكننا حساب الارتباطات الزوجية (ارتباط بيرسون أو ارتباط سيرمان). عادة ما تُعرّض في صورة جداول (مصفوفات الارتباط) أو سلسلة من مخططات التشتت أو مصفوفات التمثيل البياني. على سبيل المثال: لنجعل الأسماء والصفات

والأفعال والضمائر وأدوات العطف في مدونة BE06 متغيرات لغوية رئيسية. ما العلاقة بين هذه المتغيرات اللغوية الخمسة؟ فيما يأتي ثلاثة أوضاع مختلفة لتقديم نتائج التحليل نفسه.

يمكن استخدام الشكل (٥, ٧) للاستكشاف الأولي للاتجاهات الرئيسية في البيانات. يمكننا ملاحظة أنه ما عدا أدوات العطف (مثل and و but و or) تظهر المتغيرات الأربعة الأخرى عند النظر فيها أزواجاً على أنها علاقات إيجابية أو سلبية. من ناحية أخرى: لا تتنبأ أدوات العطف بدقة بورود فئات الكلمة الأخرى كما يتضح من الصف الأخير أو العمود الأخير.

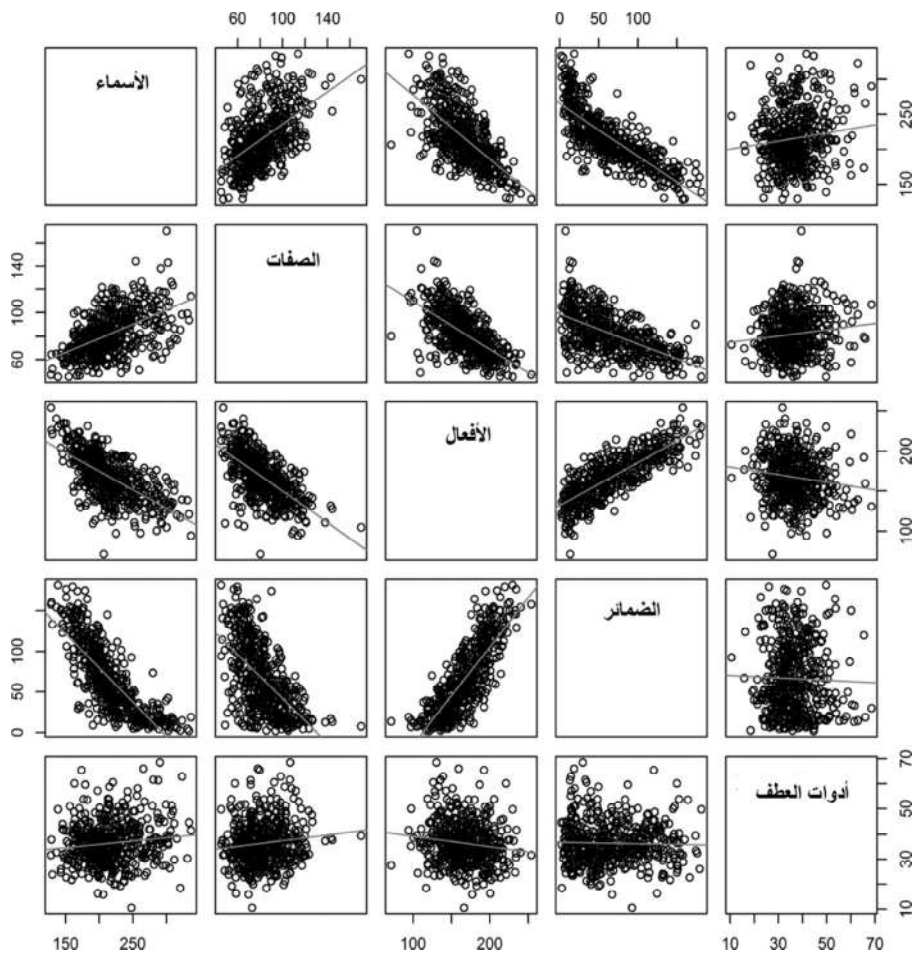
يمكن تأكيد الملاحظة الأولية بالنظر إلى ارتباطات بيرسون الموضحة في الجدول (٥, ٢) التي تُراوح من الصغيرة (0.12) إلى الكبيرة (0.81). يمكن تجاهل الارتباطات الكاملة (1.00) على الخط المائل لأنها من الواضح أن قيم المتغير فيها ترتبط بأنفسها. أصغر قيمة في الجدول قريبة جداً من الصفر (-0.03) مما يشير إلى عدم وجود علاقة بين المتغيرات. في الخمسة نصي تكون حتى الارتباطات الصغيرة (مثل: 0.12) ذات دلالة عند المستوى $0.01 < p$. القيمة الوحيدة غير الدالة هي الارتباط القريب من الصفر كما في القيمة (-0.03).

بالنسبة لأولئك الذين يفضلون المزيد من تمثيل النتائج الواردة في الجدول (٥, ٢)، تتوفر مصفوفة ارتباط (الشكل ٥, ٨). استخدمت درجات التظليل في المصفوفة لتبيين قوة الارتباط، فكلما كانت العلاقة أقوى كان الظل أعمق. في النسخ الملونة تتميز الارتباطات الإيجابية والسلبية أيضاً باستخدام ألوان مختلفة.

جدول (٥, ٢). جدول الارتباط (ارتباطات بيرسون) للأسماء والصفات والأفعال والضمائر وأدوات العطف.

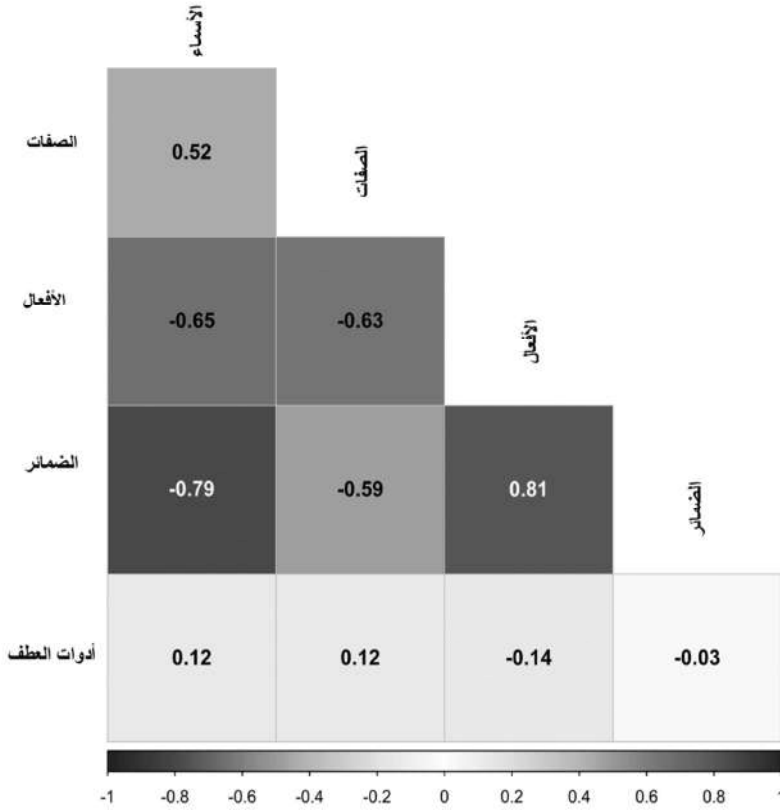
	الأسماء	الصفات	الأفعال	الضمائر	أدوات العطف
الأسماء	1.00**	0.52**	-0.65**	-0.79**	0.12**
الصفات	0.52**	1.00**	-0.63**	-0.59**	0.12**
الأفعال	-0.65**	-0.63**	1.00**	0.81**	-0.14**
الضمائر	-0.79**	-0.59**	0.81**	1.00**	-0.03
أدوات العطف	0.12**	0.12**	-0.14**	-0.03	1.00**

** تدل هذه العلامة على قيم احتمال أصغر من $p < 0.01$



شكل (٧، ٥). مخطط التبعية المتعدد.

إذن فالارتباط هو تقنية قوية لدراسة العلاقة بين المتغيرات في المدونات. ومع المدونات الكبيرة التي تتكون من مئات أو آلاف الملفات ستكون حتى الارتباطات الصغيرة (الذي يكون أصغر من 0.1) ذات دلالة إحصائية. لذلك نحن بحاجة إلى النظر بعناية شديدة في معنى الارتباطات بالنسبة للعلاقات اللغوية بين الخصائص اللغوية المختلفة.



شكل (٨, ٥). مصفوفة الارتباط: الأسماء والصفات والأفعال والضمائر وأدوات العطف.

التقرير الإحصائي: الارتباطات

١- ما يجب الإفصاح عنه

يشتمل التقرير على معامل الارتباط (r ، r_s) مع قيم الاحتمال ذات الصلة. يكون الإفصاح عادةً عن مجالات الثقة (CIs) بدلاً من قيم الاحتمال هو الأفضل لأن مجالات الثقة تقدم تقديرًا أكثر دقة عن القيمة الفعلية للارتباط في المجموع اللغوي. من المهم أيضًا تفسير حجم الارتباط (حجم التأثير)، ومن الأفضل مقارنة الارتباط الملحوظ بالارتباطات المماثلة في البيانات أو تلك الواردة في الأدبيات.

تعتمد طريقة الإفصاح عن الارتباطات المتعددة كما في الجدول (٢، ٥) بإضافة نجمة واحدة (*) أو نجمتين (***) بجوار معامل الارتباط. هذا يعني تقليدياً أن النجمة الواحدة تشير إلى أن القيمة أكبر من 0.05 (أي: $p < 0.05$) وأن النجمتين تشيران إلى أن القيمة أكبر من 0.01 (أي: $p < 0.01$). يمكننا كتابة قيم الاحتمال على الرغم من أن القيم تكون منخفضة جداً نظراً لحجم المدونات الكبيرة، أو يمكن أيضاً تحديد مجالات الثقة.

٢- مثال الإفصاح في التقرير

• هناك ارتباط إيجابي قوي لـ ($r = 0.52$)، مجال الثقة ٩٥٪ ([.46, .58]) بين عدد الأسماء وعدد الصفات المستخدمة في النصوص الإنجليزية. وهذه القيمة ليست كبيرة مثل قيمة الارتباط بين الأفعال والضمائر ($r = 0.81$)، مجال الثقة ٩٥٪ ([.775, .836]) التي تفسر تواردهما في ثلثي الحالات ($r^2 = 66\%$).

• في التوزيع التكميلي يوجد ارتباط سلبي قوي جداً ($r = -0.83$ و $p < 0.01$) بين عدد الأسماء وعدد الضمائر المستخدمة في النصوص الإنجليزية.

• بين الأفعال والصفات الإنجليزية في النصوص المكتوبة علاقة تناسبية عكسية ($r = -0.63$)، وكذلك بين الأفعال والأسماء ($r = -0.65$). الارتباط السلبي بين الأفعال وأدوات العطف صغيرة ($r = -0.14$) دون تأثير ملحوظ يذكر على أسلوب الكتابة.

(٣، ٥) التصنيف: تحليل التجميع المتكامل الهرمي

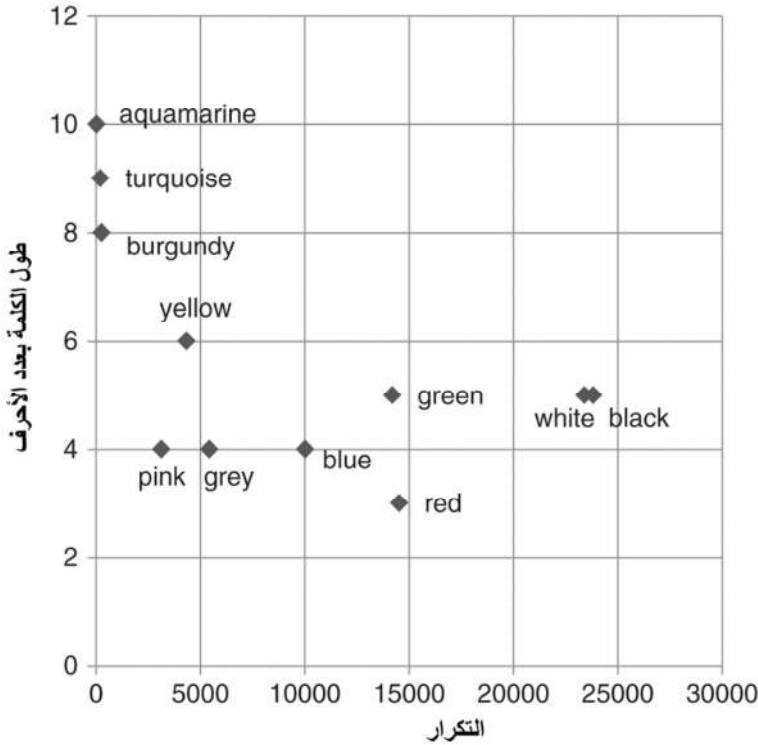
فكّر...

قبل قراءة هذا المبحث فكر في السؤال التالي:

أي مصطلحات اللون في الشكل (٩، ٥) ينتمي بعضها إلى بعض على أساس تكرارها في المدونة الوطنية البريطانية وطول الكلمة؟ ارسم دوائر حول المجموعات.

في المبحث السابق نظرنا في العلاقة بين المتغيرات اللغوية. لاحظنا أن كثيراً منها مرتبط بدرجة ما. في هذا المبحث سنحوّل التركيز إلى كلمات أو جمل أو نصوص أو متحاورين. يمكن وصفها باستخدام متغيرات لغوية متعددة. بدلاً من النظر إلى العلاقة بين الخصائص اللغوية سنبحث

في العلاقة بين الموضوعات التي تتميز بهذه الخصائص. السؤال الذي سنطرحه يسيراً نسبياً وهو موضح في مهمة (فكر)، وهذا السؤال هو الآتي: كيف يمكننا تصنيف الموضوعات بناءً على المتغيرات اللغوية؟ هنا استخدمنا مثلاً يسيراً عن مصطلحات الألوان التي تتميز بتكرار الاستخدام وطول الكلمة. لقد رأينا أن white و black هما أكثر الكلمات تكراراً، أما المصطلحات الأكثر اختصاصاً مثل aquamarine و turquoise و burgundy فهي أقل تكراراً بكثير. المصطلحات الأخيرة هي الأطول وربما بالنسبة للبعض أكثر صعوبة في النطق. إذن، كيف ستمكن من تصنيف مصطلحات الألوان هذه بناءً فقط على ما نعرفه عن عدد مرات تكرارها وعدد أحرفها؟ تتمثل إحدى الطرق الواضحة للقيام بذلك في استخدام المخطط (الشكل ٩، ٥) وإلقاء نظرة على المسافات بين النقاط الفردية.



شكل (٩، ٥). مصطلحات الألوان في المدونة الوطنية البريطانية.

نلاحظ بأن النقاط الأكثر تقاربًا تنتمي إلى المجموعة نفسها. إذا اتبعنا هذا النهج فقد نتعرف على أربع مجموعات رئيسية في المخطط: الأولى: black (أسود) و white (أبيض)، والثانية: blue (أزرق) و red (أحمر) و green (أخضر)، والثالثة: pinky (وردّي) و grey (رمادي) و yellow (أصفر)، والرابعة: aquamarine (زبرجديّ) و turquoise (فيروزيّ) و burgundy (عناي). ولكن كيف نقيس المسافة بالضبط؟ خذ على سبيل المثال المسافة بين green و blue. في الشكل البياني لكلمة green فالإحداثيات هي: [5, 14205]، أما إحداثيات الكلمة blue فهي [4, 10035]. إذا تأملنا الشكل البياني سنلاحظ أيضًا وجود تعارض بين المقياس على المحور السيني (يُراوح بين 0 و 30,000) والمحور الصاديّ (يُراوح بين 0 و 12)؛ لأننا على المحور السيني قد مثّلنا تكرارات الكلمات في المدونة الوطنية البريطانية وهي أعداد كبيرة جدًا، أما على المحور الصاديّ فمثّلنا عدد الأحرف من كلمة كلمة. إذا استخدمناها بمقياس بين الصفر والثلاثين ألف للمحورين سنمنح ثقلًا أكبر لتكرارات الكلمات ونتجاهل تقريبًا طول الكلمات. إذن، ينبغي أن نختصر المخطط ليكون أحادي البعد وليكون تمثيله تمثيلًا خطيًا ممكنًا. للتغلب على مشكلة عدم توافق المقاييس هذه يمكننا تحويل القيم إلى قيم-ز (z-scores₂). تشير الحاشية إلى أننا نستخدم مفهوم القياس (قيمة-ز) بمعنى مختلف عن الفصل الثالث ليدلّ على قيم قياسية لتغير المقياس المعني، ويُعبر عن عدد الانحرافات المعيارية للقيمة بدءًا من المتوسط. تُستخدم قيم-ز مكررة في المواقف التي يكون فيها متغيرات مُقاسة على مستويات مختلفة ولكنها تريد إعطاء كل متغير الأهمية نفسها في التحليل. تُحسب قيم-ز بأخذ كل قيمة فردية للمتغير ثم طرحه على المتوسط وقسمته على الانحراف المعياري (للمعلومية: يُحسب المتوسط باستخدام المعادلة (١, ١) من الفصل الأول، ويُحسب الانحراف المعياري وفقًا للمعادلة (١١, ٢) في الفصل الثاني):

(٥, ٩)

$$\text{قيمة-ز}_2 = \frac{\text{القيمة-المتوسط}}{\text{الانحراف المعياري (العينة)}}$$

إحداثيات قيمة-ز_٢ الجديدة للونين هي: green [-0.32, 0.58] و blue [-0.76, 0.11].

بالعودة إلى السؤال عن كيفية حساب المسافة بين green و blue في الشكل البياني في الشكل (٩, ٥) يمكننا ملاحظة أن أقصر الطرق من نقطة (أ مثلاً) إلى أخرى (ب مثلاً) هي عبر خط مباشر بين النقطتين. وهذا ما يسمى المسافة الإقليدية **Euclidean distance**، ونُحسب وفقاً للمعادلة التالية:

$$\text{معادلة المسافة الإقليدية Euclidean distance:} \quad (٥, ١٠) \\ \text{الجذر التربيعي لنتائج: } (س_ب - س_أ)^2 + (ص_ب - ص_أ)^2 + (زاي_ب - زاي_أ)^2 \text{ إلخ.}$$

حيث (س_أ) هي الإحداثية الأولى للنقطة (أ) و(س_ب) هي الإحداثية الأولى للنقطة (ب)، و(ص_ب) هي الإحداثية الثانية للنقطة (ب) ... إلخ. يمكننا الاستمرار في إضافة المتغيرات التي تحدد موضع الكلمة والانتقال من حيز ثنائي الأبعاد إلى حيز متعدد الأبعاد. في حيز ثنائي الأبعاد الذي أنشأه المتغيران اللغويان كما في مثالنا سنستخدم اثنين فقط من الإحداثيات: (س_أ و س_ب) و(ص_أ و ص_ب).

$$\text{المسافة الإقليدية (green, blue):} \quad (٥, ١١) \\ \text{الجذر التربيعي لنتائج: } (0.58 - 0.11)^2 + [(0.76 -) - (-0.32)]^2 = 0.64$$

هناك طرق بديلة لحساب المسافة بين أ و ب. تخيل أنك في مدينة كبيرة مثل نيويورك وتريد الانتقال من (أ) إلى (ب). ما لم يكن لديك طائرة هليكوبتر فلن يمكنك الذهاب مباشرة. بدلاً من ذلك سيتعين عليك دخول أحد الشوارع ثم الاستدارة ٩٠ درجة إلى شارع آخر. المسافة بين (أ) و(ب) المقاسة على هذا النحو التي تتبّع مسار الزوايا القائمة تسمى مسافة مانهاتن **Manhattan distance**. نُحسب مسافة مانهاتن وفقاً للمعادلة التالية:

$$\text{مسافة مانهاتن (أ, ب) = } |س_ب - س_أ| + |ص_ب - ص_أ| + |ع_ب - ع_أ| \quad (٥, ١٢) \\ \text{حيث } |س_ب - س_أ| \dots \text{ إلخ هي القيمة المطلقة، أي: عدد موجب للفرق بين إحداثيات (أ)}$$

و(ب). بالنسبة لـ green و blue في مثالنا سنحصل على الآتي:

مسافة مانهاتن (blue و green) (١٣, ٥)

$$0.91 = |(-0.76) - -0.32| + |0.58 - 0.11|$$

المتوقع هو أن مسافة مانهاتن أكبر من المسافة الإقليدية المباشرة على الرغم من أنه عند استخدامها لتحليل التجميع فإن كلاً من القياسين للمسافة يحقق نتائج متشابهة إلى حد ما. على الرغم من ذلك فإن مسافة مانهاتن تكون أكثر قوة عند التعامل مع القيم المتطرفة (انظر المبحث ٣, ١ للاطلاع على تعريف القيم المتطرفة). تشمل الأنواع الأخرى من مقاييس المسافة مسافة كانبرا Canberra distance التي هي شكل موحد للمسافة الإقليدية ومانهاتن التي تضع مزيداً من التركيز في الأشياء الأبعد، والنسبة المئوية لعدم التوافق التي تُستخدم عند التعامل مع المتغيرات الفئوية (Everitt et al. 2011).

في هذه المرحلة لدينا كل ما نحتاج إليه لبدء تحليل التجميع. هناك العديد من أنواع التحليل التجميعي المختلفة (Everitt et al. 2011). النوع المستخدم عادةً في لغويات المدونة الحاسوبية للبيانات غير التعاقبية^(٤) هو تحليل التجميع المتكامل الهرمي Hierarchical agglomerative cluster analysis (Gries 2013a: 336ff). المصطلح الذي قد يبدو مخيفاً إلى حد ما هو في الواقع ملخص وصفي جيد جداً للعملية. إننا نأخذ نقاط البيانات الفردية وفي إجراء (هرمي) خطوة بخطوة نضم (أي نُكْتَل) الأقرب منها حتى ننشئ تجميعاً واحداً كبيراً يحتوي على جميع نقاط البيانات. يوضح الشكل (١٠, ٥) هذه العملية، إذ لكل لوحة خطوة فردية إجرائياً.

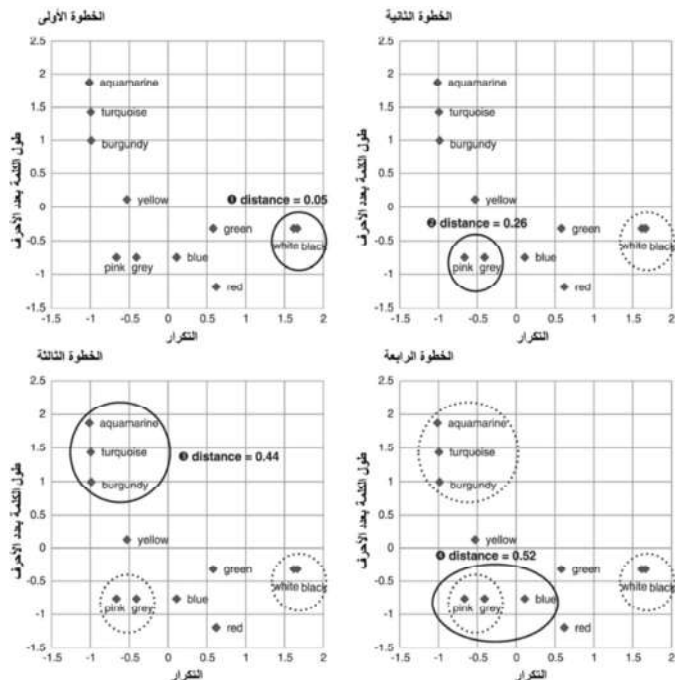
ويتضح من الشكل (١٠, ٥) يُضم black و white في تجميع صغير ثم pink و grey و aquamarine و turquoise و burgundy. بعد ذلك يُنشأ تجميع أكبر بالضم بين pink و grey وبين العنقود: blue. إذا أسرعنا في إجراء العملية فسنحصل على المخطط التجميعي الموضح في الشكل (١١, ٥) الذي يبين جميع الخطوات الفردية في إجراء التجميع وكذلك النتيجة النهائية لتجميع واحد كبير لكل الألوان يحتوي على جميع نقاط البيانات.

(٤) بالنسبة للبيانات التعاقبية (التاريخية) يُستخدم نوع آخر من تحليل التجميع يسمى تحليل التجميع المجاور. نوقشت هذه الطريقة في المبحث (٤, ٧).

مع وجود تجميعات متعددة فإنه سرعان ما يزداد المخطط التجميعي تعقيداً مع صعوبة في قراءته. لهذا السبب تُعرض نتائج إجراء التجميع عادةً على أنها مخطط تفرعي هرمي dendrogram. يُظهر المخطط التفرعي الموضح في الشكل (١٢، ٥) الخطوات الفردية لإجراء التجميع مع عدم عرض الأرقام عادة، إذ تتحول الفروع الأصغر (التجميعات) تدريجياً إلى تجميع واحد كبير. أفضل طريقة لقراءة المخطط التفرعي هي من أسفل إلى أعلى بدءاً من التجميعات الأكثر تقارباً كما هو محدد بمقياس المسافة المختار (المسافة الإقليدية، مانهاتن، كانبرا). تُعرض قيمة المسافة distance بين نقاط البيانات/ التجميعات في صورة ارتفاع المخطط التفرعي، وكلما زاد الارتفاع زادت المسافة بين نقاط البيانات أو التجميعات التي اندمجت. على سبيل المثال: في الشكل (١٢، ٥) نلاحظ المسافة الإقليدية بين black و white على مقياس قيمة-ز، قريبة من الصفر كما هو موضح في المحور الصادي، في حين أن المسافة بين تجميع white و black وبين بقية نقاط البيانات قريبة من القيمة (١).

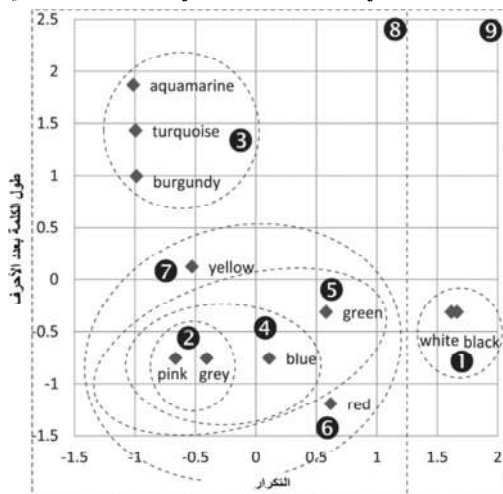
افتراضنا حتى الآن أن طريقة ضم التجميعات الصغيرة إلى تجميعات أكبر هي طريقة واضحة ومباشرة. في إجراء التجميع نحتاج إلى تحديد كيفية القيام بذلك بدقة. السؤال الذي نحتاج إلى طرحه هو: ما نقاط البيانات التجميعية الصغيرة التي ينبغي تمثيلها لحالة التجميع بأكمله؟

يمكن تقديم أربع إجابات. الأولى: يُسمى أقرب نقطة إلى التجميع المجاور الذي نريد دمج تجميعنا الأصلي معه بطريقة الربط المنفرد single linkage SLINK. الثانية: يُسمى أبعد نقطة إلى التجميع المجاور التي نريد دمج تجميعنا الأصلي معها بطريقة الربط التام complete linkage CLINK. الثالثة: لا شيء، يُنظر في المسافات المتبادلة لجميع نقاط البيانات بأخذ القيمة المتوسطة (طريقة الربط المتوسط average linkage method). الرابعة: لا شيء، تُراعى فيه المسافات المتبادلة لجميع نقاط البيانات بحساب مجموع المسافات المربعة (طريقة وارد Ward). من المهم أن ندرك أن الطرق المختلفة لدمج التجميعات تُنتج نتائج مختلفة كما يتضح من الأشكال (١٣، ٥ و ١٤، ٥ و ١٥، ٥ و ١٦). لاحظ استخدام طريقة الربط المنفرد لإنتاج مخطط تفرعي في الشكل (١٢، ٥).

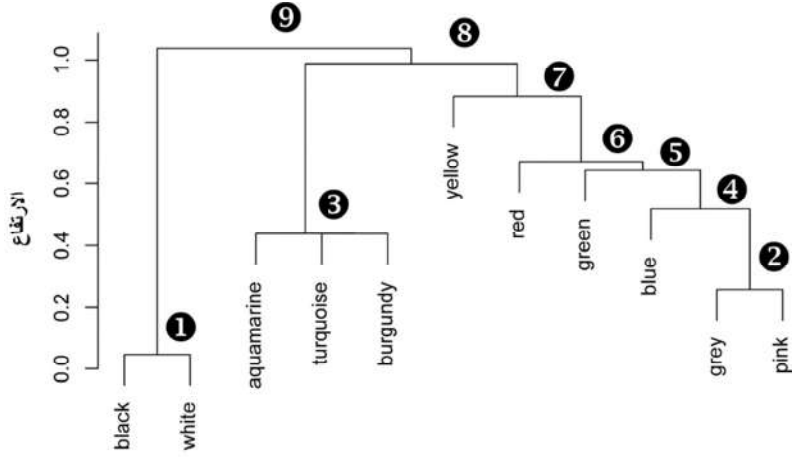


شكل (١٠، ٥). الخطوات الأربع لإنشاء التجميعات.

black (أسود) وwhite (أبيض)، وblue (أزرق) وred (أحمر) وgreen (أخضر) وpinky (وردي) وgrey (رمادي) وyellow (أصفر)، وaquamarine (زبرجدي) وturquoise (فيروزي) وburgundy (عنابي)

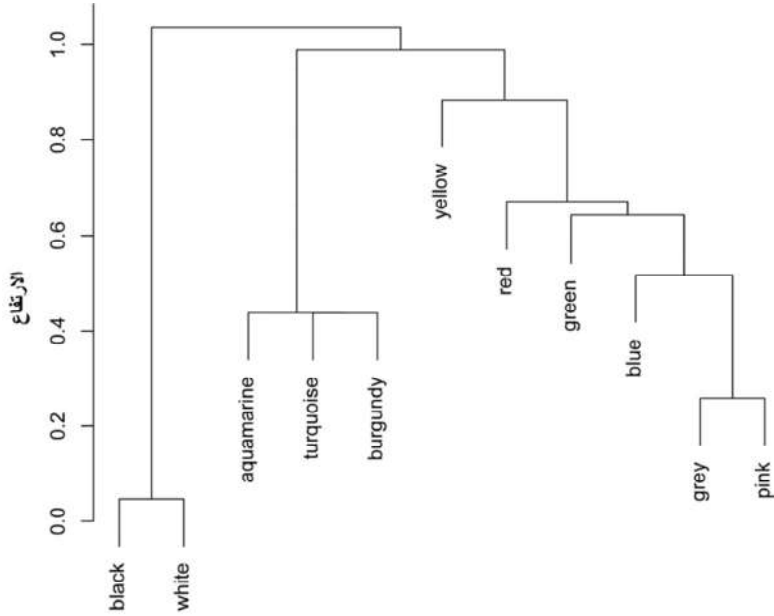


شكل (١١، ٥). النتيجة النهائية لإنشاء التجميعات.



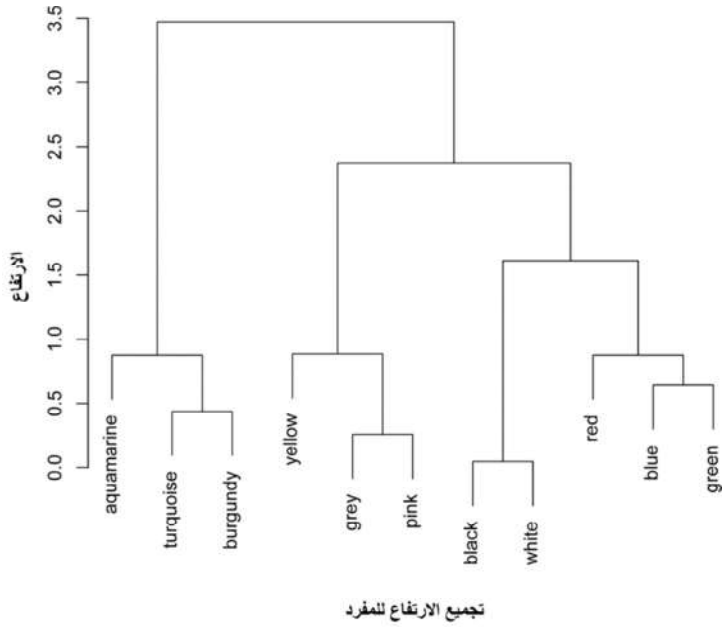
تجميع الارتفاع للمفرد

شكل (١٢, ٥). مصطلحات الألوان: مخطط تفرعي بقيمة-زرة المعيرة والمسافة الإقليدية وطريقة الربط المنفرد.

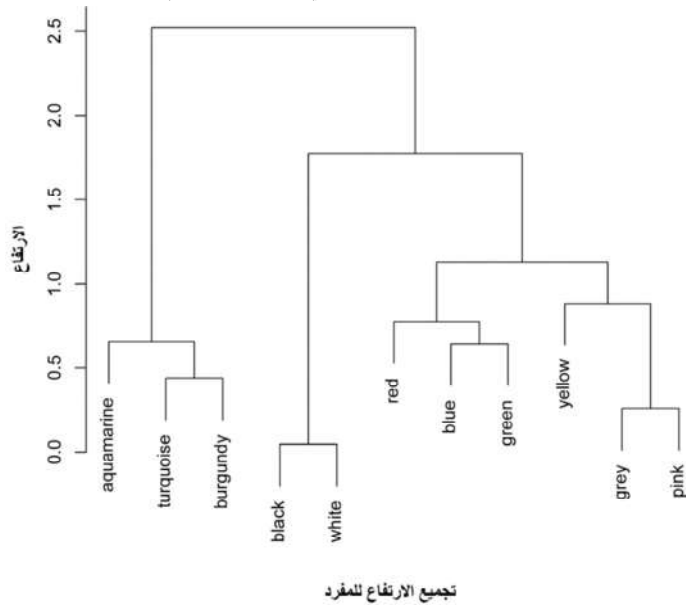


تجميع الارتفاع للمفرد

شكل (١٣, ٥). مخطط تفرعي لطريقة الربط المنفرد.



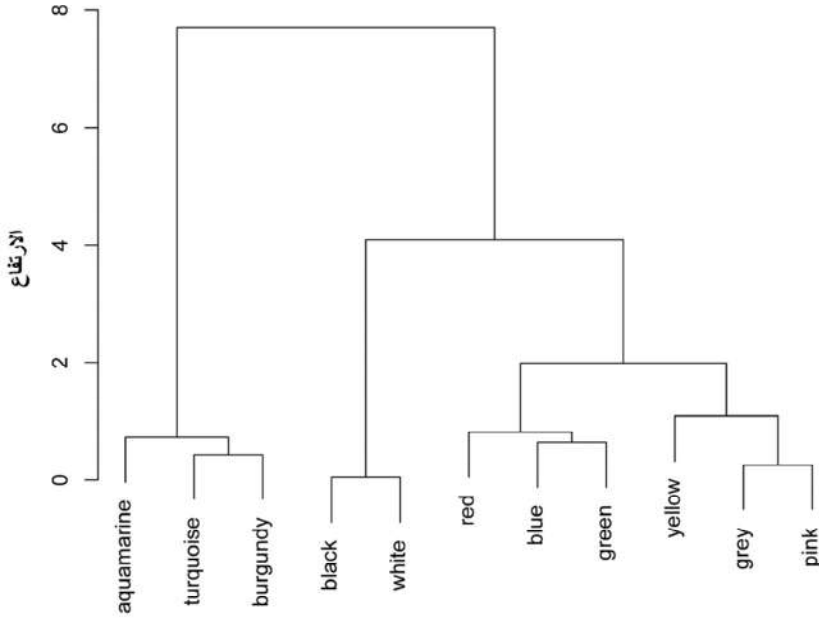
شكل (١٤, ٥). مخطط تفرعي لطريقة الربط التام .



شكل (١٥, ٥). مخطط تفرعي لطريقة الربط المتوسط .

لكل طريقة مزاياها وعيوبها كما هو الحال دائماً، فطريقة الربط الفردي على سبيل المثال سهلة للغاية، لكنها تؤدي إلى تأثير (تسلسلي) مرئي في الفرع الأيمن من المخطط التفرعي كما في الشكل (١٣، ٥)، فينشأ عنها هيكل تجميعي مهلهل نوعاً ما، وليس فيه مجموعات تجميعية بارزة. في لغويات المدونة الحاسوبية يبدو أن أسلوب Ward هو الأشهر لأنه يُنشئ تجميعات مدمجة يسهل تفسيرها نسبياً (راجع Divjak & Gries 2006 و Gries et al. 2009).

ينبغي وضع نقطة أخيرة عن التجميعات، ففي هذا المثال نظرنا في متغيرين لغويين في حيز ثنائي الأبعاد، ويمكن أيضاً استخدام تقنية التجميع مع العديد من المتغيرات اللغوية. يشبه هذا الإجراء المثال اليسير ذكرناه ويكمن الفرق الوحيد في أنه بالنظر إلى وجود متغيرات متعددة فإننا ننظر إلى المسافات بين نقاط البيانات في حيز متعدد الأبعاد. يرد مثال على استخدام تقنية التجميع مع ٤٤ متغيراً لغوياً في المبحث (٥، ٥).



تجميع الارتفاع للمفرد

شكل (١٦، ٥). مخطط تفرعي وفق منهج Ward.

التقرير الإحصائي: تحليل التجميع المتكامل الهرمي

١- ما يجب الإفصاح عنه

تحليل التجميع هو طريقة بصرية استكشافية تُمثل التنميط في البيانات تمثيلاً بيانياً. ينبغي الإفصاح عن كلا البارامترين لتحديد التجميع في تحويل البيانات، وقياس المسافة، وطريقة جمع الكتلة، وكذلك لنتائج التحليل في المخطط التفرعي. يصف مبحث الطريقة في التقرير البحثي الإجراء التحليلي والبارامترات المستخدمة. في النتائج والمناقشة. ينبغي مناقشة كل مخطط تفرعي بعناية. السؤال الرئيسي الذي يتطلّب تناوله هو: كم عدد العوامل المهمة التي يمكن ملاحظتها في المخطط؟

٢- مثال الإفصاح في التقرير

حللت البيانات باستخدام تقنية التجميع المتكامل الهرمي (تحويل بيانات قيمة-ز، ومسافة مانهاتن، وطريقة وارد Ward).

للاطلاع على مثال عن الإفصاح عن نتائج تحليل التجميع ومناقشتها انظر: المبحث (٥، ٥).

(٥، ٤) التحليل متعدد الأبعاد

فكّر...

قبل قراءة هذا المبحث فكر في السؤال التالي:

فيم يختلف النصّ أدناه المأخوذان من المدونة الوطنية البريطانية من جهة اللغة المستخدمة؟

(أ) النص

Margaret: We shall go back, erm after Easter

Bob: Yes

Margaret: hoping, permitting, you know, if it's not too expensive, it got very dear you know

Bob: Yes, that is also a thing to

Margaret: Yeah

Bob: erm, I'm, I'm feeling a bit hard up at the moment, I had a

Margaret: Yes

Bob: a bill for repairing the car for two hundred and thirty pounds and <pause> so

[الترجمة]:

مارغريت: سنعود أدراجنا، إم بعد عيد الفصح

بوب: أجل

مارغريت: على أمل، إن سمحت لي، كما تعرف، إذا لم تكن باهظة الثمن، فستكون أكثر كما تعرف
يا عزيزي

بوب: أجل، هذا أمر يحتاج

مارغريت: نعم

بوب: إيم، أنا، لدي شعور قوي في هذه اللحظة، كان لدي

مارغريت: نعم

بوب: فاتورة لإصلاح السيارة بمئتين وثلاثين جنيهًا إسترلينيًا و <توقف> لذا

النص (ب)

MacCulloch (1819), and later, Murchison and Geikie (1861), provided the first accounts of the geology of Lewis and Harris and the other islands which constitute the north east-south west chain of islands called the Outer Hebrides, which lies some 70 Ian west of the northern Scottish mainland. In 1923, Jehu and Craig produced the first detailed account of the geology of this region, and followed it up with further accounts between 1925 and 1934. Thereafter, several research papers were published on the Lewisian complex, including Dearnley (1962), Myers (1970, 1971), Coward (1972, 1973) and Graham (1973); [الترجمة]:

قدم ماكوللو (١٨١٩) ومن بعده مورشيسون وجيكي (١٨٦١) أولى التفسيرات الجيولوجيا لويس وهارس وجزر أخرى تشكل شمال شرق وجنوب غرب سلسلة الجزر التي تعرف باسم أوتر هيبيريديس. جافيو وكريغ قدما تعليلًا مفصلاً عن جيولوجيا هذه المنطقة، واستمرا في شروحهما بين عامي ١٩٢٥-١٩٣٤م. تبع ذلك عدد من الأبحاث التي نشرها مجمع لويزيان لكل من ديرنلي (١٩٦٢)، ومايرس (١٩٧٠، ١٩٧١)، وكاوارد (١٩٧٢، ١٩٧٣)، وغراهام (١٩٧٣).

بوصفنا مستعملين للغة فنحن متعدّدو الأساليب فيها، وهذا يعني قدرتنا على توظيف مجموعة متنوعة من أساليب الكلام والكتابة حسب الموقف. نستخدم نوعًا مختلفًا من اللغة عندما نتحدث تحدّثًا غير رسمي مع الأصدقاء أو عندما يُطلَبُ منا كتابة تقرير بحثي. هذه القدرة على تغيير أسلوب التحدث أو الكتابة تنعكس مباشرة في الأوعية أو السجلات اللغوية التي ننتجها كما يتضح في المثالين من مهمة (فكر). لعلك لاحظت أن الحوار المنطوق (النص أ) يستخدم بنى نحوية أقصر

(تسمى الأقوال utterances) من النص الأكاديمي (النص ب) الذي يستخدم جملاً طويلة ومعقدة نسبياً. نرى النص (ب) منمقاً بدرجة أكبر بكثير، ويخلو من التردد والتمتات (erm) والتكرار (I'm, I'm feeling) والبدايات الخطأ التي لم تكتمل (Yes, that is also a thing to ... erm, I'm). هناك العديد من الخصائص اللغوية الأخرى التي تميز الكلام غير الرسمي مثل الاستخدام المتكرر للضمائر الشخصية والاختصارات وعلامات الخطاب discourse markers في حين أن الكتابة الأكاديمية تستخدم عادة عدداً كبيراً من الأسماء وحروف الجر وصيغة المبني للمجهول.

إذن فالسألة التي نحن بصددتها في التحليل اللغوي للسجلات اللغوية هي كيفية فهم مقدار التنوع الوظيفي الكبير في البيانات، وكيفية تمييز السجلات اللغوية بالنظر في المبادئ الأساسية للتنوع المنهجي (Biber 1988; Conrad & Biber 2001). تستخدم السجلات اللغوية المختلفة باختلاف وظائفها العديد من الخصائص اللغوية لتحقيق الهدف التواصل، وهو التفاعل بنجاح في ضوء الاصطلاحات المتوقعة اجتماعياً أو أكاديمياً أو غير ذلك. يُعدّ التحليل متعدد الأبعاد إجراءً معقداً وضعه بايبر (1988)، ويتناول عدداً كبيراً من المتغيرات اللغوية، ويحدد المبادئ الأساسية للتنوع الوظيفي بالنظر في كيفية توارد المتغيرات اللغوية الفردية في النصوص. إنه يبدأ بملاحظة يسيرة في أنّ السجلات اللغوية تختلف بطرق متعددة لأن لها وظائف مختلفة.

للتحليل الكامل المتعدد الأبعاد أربع مراحل رئيسية. الأولى: تحديد المتغيرات ذات الصلة. الثانية: استخلاص العوامل من المتغيرات. الثالثة: التفسير الوظيفي للعوامل بوصفها أبعاداً. الرابعة: إدراج السجلات اللغوية في الأبعاد. يمكننا أيضاً إجراء مقارنة واضحة مع الأبعاد الأصلية لبايبر (1988) بتخطي المرحلة الأولى والثانية والثالثة والتركيز في مقارنة السجلات اللغوية في مدونتنا مع السجلات اللغوية المحددة في دراسة بايبر الأصلية (Conrad & Biber 2001). لتفحص المراحل الأربع واحدة تلو الأخرى (انظر أيضاً في Friginal & Hardy 2014).

الخطوة الأولى: تحديد المتغيرات ذات الصلة

قبل إجراء التحليل الإحصائي في الخطوة الثانية نحتاج إلى تحديد عدد كبير من المتغيرات اللغوية في نصوص مدونتنا. ينبغي أن تشمل المدونة على سجلات لغوية مختلفة (مثل الكلام غير الرسمي، والتقارير الإخبارية، والكتابة الأكاديمية، والرواية الشعبية وما إلى ذلك). نحتاج إلى

تحديد المتغيرات ذات الصلة، وهي المتغيرات التي يمكن أن تميز بين السجلات اللغوية في مدونتنا. على سبيل المثال: لجميع المتغيرات المذكورة في المبحث (٢, ٥) أثرٌ مهم في تحديد السجل اللغوي. نُحدد في العادة عشرات المتغيرات (من ٤٠ إلى أكثر من ١٤٠ متغيراً). استخدم باير ٦٧ متغيراً كما هو موضح في الجدول (٣, ٥).

يمكن العثور على قائمة أكثر شمولاً (١٤١ عنصراً) تجمع بين الخصائص النحوية المعجمية والدلالية في دراسة شياو Xiao (٢٠٠٩). يتبع التحليل متعدد الأبعاد تصميم بحث لنص فردي أو لنص حوار (المبحث ٤, ١). يعد الشكل (١٧, ٥) مثالاً لجزء من مجموعة البيانات المستخدمة للتحليل، ويمثل كل صف ملفاً فردياً، وتُظهر الأعمدة بين الثالث والثمان التكرارات النسبية لمتغيرات باير (١٩٨٨) في الملفات. عند إعداد البيانات للتحليل المتعدد الأبعاد نحتاج إلى التأكد من أن لدينا ملفات كافية من عينات متنوعة في المجموعة لإلقاء نظرة مفيدة على التنوع.

جدول (٣, ٥). المجموعة الكاملة لخصائص باير (١٩٨٨) القائمة على كونراد وباير (٢٠٠١: ١٨-١٩).

١- الزمن الماضي
٢- الجهوية التامة perfect aspect
٣- الزمن المضارع
٤- ظروف المكان (مثل: محليّ locally، الدور السفلي downstairs، خلف behind)
٥- ظروف الزمان (مثل: الوقت الحاضر nowadays فوراً immediately، أخيراً eventually)
٦- ضمائر المتكلم
٧- ضمائر المخاطب
٨- ضمائر الغائب الشخصية (ما عدا it)
٩- الضمير it
١٠- ضمائر الإشارة (الضمائر: ذلك that، هذا this، هذه these، أولئك those)
١١- ضمائر الإبهام (أي أحد anyone، كل أحد everybody، لا شيء nothing)
١٢- الفعل: فعل do الذي لا يتكرر مرة أخرى عند ذكره لأول مرة [هل فعلتَ ذلك؟ نعم (فعلت)].
١٣- أسئلة Wh-
١٤- التحويل الاسمي بإضافة اللواحق الدالة عليه: (-tion, -ment, -ness, -ity)
١٥- اسم المصدر gerund (تصريف الفعل بإضافة -ing لتشبيهها بالاسم)

تابع جدول (٣، ٥).

١٦- مجموع الأسماء الأخرى	
١٧- المبني للمجهول الخالي من الفاعل	
١٨- الجمل المبني للمجهول مع نائب فاعل بإضافة (من قبل by)	
١٩- فعل bc الأساسي	
٢٠- الضمير الوجودي: هناك there	
٢١- المكملات من الجمل الفعلية لـ that (مثل: شعرنا بأننا نحتاج إلى دعم مالي We felt <u>that</u> we needed a	(financial base
٢٢- المكملات الوصفية لـ that (مثل: من الواضح جداً بأن أموراً محددة يمكن أن تكون ذات ارتباط جيني It's	(quite obvious <u>that</u> certain things can be sexlinked
٢٣- عبارات wh- (مثل: تساءلت عما ستفعله I wondered <u>what</u> to do	
٢٤- المصادر infinitives	
٢٥- العبارات الظرفية مع التصريف الرابع (مثل: يصرخ غضباً عندما ركض على الدرج Screaming with rage,	(he ran up the stairs
٢٦- العبارات الظرفية مع التصريف الثالث (مثل: بمعطيات هذه الخصائص ليس مفاجئاً أن... <u>Given</u> these	(characteristics, it is not surprising that ...
٢٧- عبارة التصريف الثالث المختصرة بعد الأسماء (مثل: مطلوب لحجم الهواء العادم شبكة بطول ٦ أقدام وعرض ٤ أقدام	(the exhaust air volume <u>required</u> by the 6-ft. x 4-ft. grid
٢٨- عبارة التصريف الرابع المختصرة بعد الأسماء (مثل: التيارات المتلاقية تحوم تحت السطح the currents of	(dissent <u>swirling</u> beneath the surface
٢٩- العبارات الموصولة بـ that في موقع الفاعلية (مثل: الأوراق التي على الطاولة the papers that are on the table	
٣٠- العبارات الموصولة بـ that في موقع المفعولية (مثل: الأبحاث التي اعتقدت أنها ستكون مثيرة للاهتمام the	(papers <u>that</u> she thought would be interesting
٣١- موصولات wh- في موقع الفاعلية (مثل: الناس الذين يعرفونه (people <u>who</u> know him	
٣٢- موصولات wh- في موقع المفعولية (مثل: الناس الذين يعرفهم (people <u>who</u> he knows	
٣٣- العبارات الموصولة الاجتالية (مثل: الطريقة التي بها يُهضم الأكل (the way <u>in which</u> food is digested	
٣٤- موصولات الجملة (مثل: انتظرنا ست ساعات وكان ذلك سخيفاً <u>which</u> was ridiculous (We waited for six hours,	
٣٥- الرابط الظرفي السببي: because	
٣٦- الروابط الظرفية التناقضية مثل: though و although	
٣٧- الروابط الظرفية الشرطية مثل: if و unless	

تابع جدول (٣, ٥).

٣٨-	روابط ظرفية أخرى مثل: while و such that و insomuch as
٣٩-	الجار والمجرور
٤٠-	الصفات الوصفية (مثل: الغرفة الصغيرة the small room)
٤١-	الصفات الإسنادية (مثل: الغرفة صغيرة the room is small)
٤٢-	بقية جميع الظروف
٤٣-	نسبة عدد الكلمات النوعية إلى عدد الكلمات الفعلية
٤٤-	متوسط طول الكلمة
٤٥-	أدوات الربط مثل: بدلاً من ذلك alternatively, مع ذلك nevertheless, لذلك therefore
٤٦-	المُلقَّفات مثل: باعتدال mildly, جزئياً partially, إلى حد ما somewhat
٤٧-	تعبيرات الاحتراز (مثل: تقريباً almost, ربما maybe, نوعاً ما sort of [عدا ما يكون اسماً حقيقياً])
٤٨-	التضخيم (مثل: تماماً لـ completely و totally و utterly)
٤٩-	المؤكدات (مثل: كثيرٌ من a lot of و مؤكَّدٌ for sure و حقّاً really)
٥٠-	أدوات خطابية كمستهلات الجمل: على أي حال anyhow, وحسناً well, الآن now
٥١-	أسماء الإشارة
٥٢-	الأفعال الناقصة الاحتمالية (can, could, may, might)
٥٣-	الأفعال الناقصة الإلزامية (must, ought, should)
٥٤-	الأفعال الناقصة التنبؤية (shall, will, would)
٥٥-	الأفعال المتعلقة بالآخرين مثل: يشكو complain, يشرح explain, يعدد promise
٥٦-	الأفعال المتعلقة بالذات مثل: يعتقد believe, يفكر think, يعرف know
٥٧-	الأفعال الإقناعية مثل يأمر command, يفرض propose, يوصي recommend
٥٨-	الفعل (يبدو) appear و seem
٥٩-	الاختصارات كالفاصلة العلوية في: don't
٦٠-	حذف that المكملة (مثل: أظن أنه اختفى مسبقاً I think [Ø] he's gone already)
٦١-	حروف الجر المقطوعة (مثل: الشخص الذي كنت أتحدث إليه the person that I was talking to)
٦٢-	المصدر المؤوّل المفعول (مثل: I want to completely convince you that)
٦٣-	الفعل المساعد المفعول (مثل they have apparently sold it all)
٦٤-	التركيب البسيط المعطوف مثل: [اسم+اسم]، و[صفة+صفة]، و[فعل+فعل]، و[ظرف+ظرف]
٦٥-	أداة عطف جملة جديدة مستقلة مثل: and التي يُبتدأ بها في الجمل الاستهلاكية
٦٦-	النفي الاحتمالي (مثل: لم يُعثر على دليل No evidence was found)
٦٧-	النفي المؤكد (مثل: هذا غير صحيح That's not true)

من الناحية المثالية ينبغي أن يكون عدد الملفات خمسة أضعاف عدد المتغيرات اللغوية على الأقل، ولكن في بعض الأحيان قد يكفي^(٥) عدد ملفات أقل من ذلك (Friginal & Hardy 2014: 303).

اسم الملف	السجل اللغوي	ماض	تأم	مضارع	المكان	الزمان	ضمير المتكلم
BE API	أخبار-تحقيق صحفي	4.35	0.89	3.86	0.45	0.59	1.58
BE_A02	أخبار-تحقيق صحفي	3.15	0.94	3.64	0.33	0.77	0.72
BE A03	أخبار-تحقيق صحفي	4.12	1.52	5.11	0.34	0.34	0.39
BE A04	أخبار-تحقيق صحفي	6.68	0.95	2.79	0.4	0.75	0.35
BE_A05	أخبار-تحقيق صحفي	4.36	0.62	4.15	0.67	0.88	1.66
BE_A06	أخبار-تحقيق صحفي	3.45	1.18	4.28	0.05	0.41	0.31
BE_A07	أخبار-تحقيق صحفي	2.99	1.42	4.35	0.3	0.51	0.56
BE A08	أخبار-تحقيق صحفي	4.67	0.68	3.46	0.47	0.84	1.89
BE A09	أخبار-تحقيق صحفي	5.44	0.64	5.79	0.25	0.74	1.68
BE_A10	أخبار-تحقيق صحفي	6.85	0.77	3.76	0.39	0.82	1.74
BE_A11	أخبار-تحقيق صحفي	3.96	0.84	3.96	0.5	0.74	0.45
BE_A12	أخبار-تحقيق صحفي	5.32	1.27	5.27	0.61	0.2	2.68
BE_A13	أخبار-تحقيق صحفي	4.16	1.57	4.31	0.51	0.25	0.66
BE_A14	أخبار-تحقيق صحفي	2.78	0.36	4.84	0.36	0.21	1.08

شكل (١٧، ٥). مجموعة بيانات للتحليل متعدد الأبعاد (مقتطف صغير).

(٥) يحدث هذا عندما يكون هناك عدد من الارتباطات القوية بين المتغيرات اللغوية (> 0.5) والأنماط واضحة حتى مع نقاط عينات أقل.

الخطوة الثانية: استخلاص العوامل من متغيرات متعددة

تُستخلص العوامل باستخدام تقنية إحصائية تسمى تحليل العوامل. تحليل العوامل **factor analysis** هو إجراء رياضي معقد يختزل عددًا كبيرًا من المتغيرات اللغوية إلى عدد صغير من العوامل، يجمع كل منها متغيرات لغوية متعددة. يحصل ذلك من خلال النظر في الارتباطات بين المتغيرات (انظر: المبحث ٢، ٥، للاطلاع على شرح الارتباط). تُعد المتغيرات التي ترتبط إيجابيًا أو سلبًا مكونات للعامل نفسه نظرًا إلى ما بينها من اتصال. تعني الارتباطات الإيجابية أن المتغيرات تُظهر نمط الوجود نفسه في البيانات، أما الارتباط السلبى فيشير إلى توزيع تكميلي، أي: إذا ظهر أحد المتغيرات بتكرار مرتفع يظهر المتغير الآخر بتكرار ضئيل والعكس بالعكس. العامل هو مجموعة من المتغيرات اللغوية ذات الصلة التي تلخص نزعة عامة وبعدها أساسيًا في البيانات (انظر الخطوة الثالثة). هذا الترابط بين المتغيرات اللغوية التي تصوّرها العوامل هو انعكاس مباشر للخصائص الوظيفية الأساسية للسجلات اللغوية، وهو يساعدنا على فهم كيفية جمع المتحدثين أو الكتاب بين خصائص لغوية مختلفة تناسب مع موقف معين. على سبيل المثال: نحن نعلم أن الضمائر الشخصية والصيغ المختصرة والأفعال المضارعة والأفعال من قبيل *believe* و *think* و *know* تُستخدم للتعبير عن الطابع غير الرسمي والتواصل الشخصي *personal involvement* للمتحدث، وهو أمر نموذجي في السجلات اللغوية مثل الرواية والكلام غير الرسمي.

يوضح الشكل (١٨، ٥) المبدأ الرئيسي لتحليل العوامل باستخدام البيانات اللغوية الحقيقية، ولكن بعدد أقل من المتغيرات والعوامل مقارنةً بالاعتاد (انظر الخطوة الأولى)^(٦). يوضح أيضًا عشرة متغيرات لغوية مدمجة في عاملين. يتكون العامل الأول من أفعال خاصة (مثل *believe* و *think*)، وحذف *that*، والأفعال المضارعة، والاختصارات، وضمائر المخاطب. أما العامل الثاني فيشتمل على أفعال ماضية، وضمائر غائب، وصيغ الزمن التام (الجهوية التامة)، والأفعال العامة (مثل *complain* و *promise*) والنفي التركيبي *synthetic negation*. يمثل الخط الأفقي (المحور السيني) العامل الأول، أما العامل الثاني فيمثله الخط العمودي (المحور الصادي). تُظهر نقاط البيانات المتغيرات اللغوية الفردية على شكل معين.

(٦) الرسم التوضيحي مستوحى من فيلد وآخرين (Field et al 2012: 753ff).



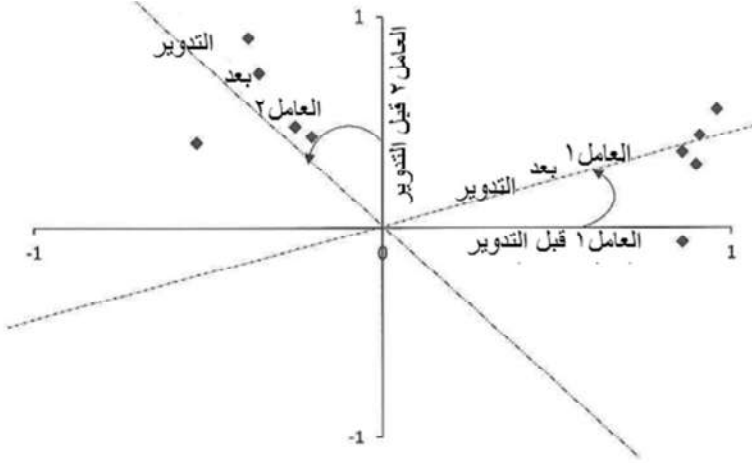
شكل (١٨, ٥). اختزال البيانات من ١٠ متغيرات إلى عاملين.

يجب ذكر ثلاثة تفاصيل مهمة قبل الانتقال إلى الخطوة الثالثة. التفصيل الأول: قبل إجراء تحليل العوامل نحتاج إلى التحقق من أن المتغيرات اللغوية الفردية مرتبطة ارتباطاً معقولاً ببعض المتغيرات الأخرى في مجموعة البيانات، ويكون ذلك الارتباط أعلى من (0.3)، وإلا فإننا لن نستطيع الجمع بين المتغيرات في العوامل. يمكن إجراء الفحص بالنظر إلى ارتباطات بيرسون بين أزواج المتغيرات. تُعرض الارتباطات عادةً على شكل مصفوفة ارتباط، يرتبط فيها كل متغير بباقي المتغيرات في مجموعة البيانات^(٧). يمكن استبعاد تلك المتغيرات التي لا ترتبط بمتغيرات أخرى من متابعة التحليل.

التفصيل الثاني: لتحسين نتائج تحليل العوامل تُدوّر العوامل التي يمثلها المحوران في الشكل (١٨, ٥)، وتمر عبر منتصف المتغيرات ذات الصلة. بهذه الطريقة تُمثّل العوامل المتغيرات الأساسية تمثيلاً أكثر جودة. هناك أنواع مختلفة من التدوير يمكن تطبيقها (فاريماكس varimax وبروماكس promax وغيرهما). نظراً لأننا نتعامل مع بيانات لغوية يمكننا أن نفترض أن العوامل مرتبطة كما هو

(٧) بالنظر في مصفوفة الارتباط، يُجرى اختبار بارتليت Bartlett، وهو اختبار إحصائي رسمي يتحقق من الحد الأدنى للعلاقة بين المتغيرات كأن تكون قيمة المتغير مختلفاً اختلافاً دالاً من القيمة صفر.

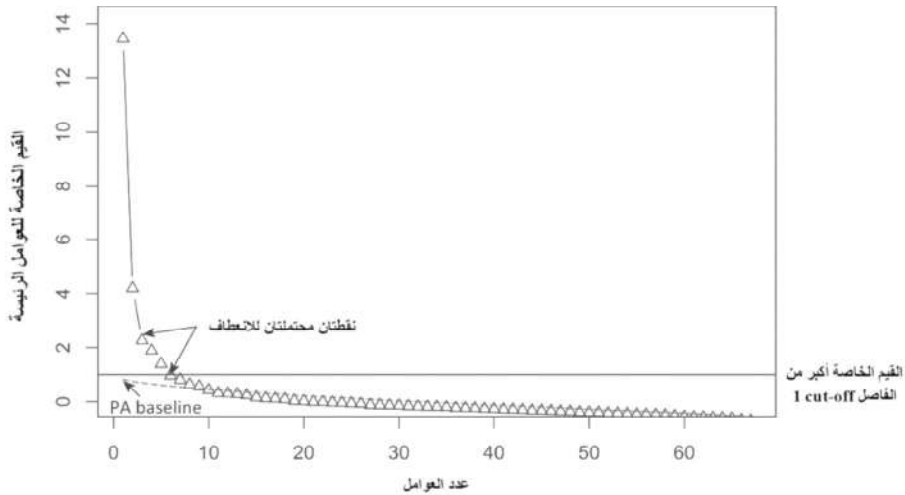
الحال تقريباً بكل شيء في اللغة، ومن ثمّ فالتدوير المناسب لهذا النوع من التحليل هو بروماكس (أي: التناوب المائل) لوجود زاوية مائلة بين الخطوط التي تمثل العوامل. من الممكن أيضاً اختيار فاريماكس (التدوير العمودي) وهو الافتراضي في أنواع التحليل غير اللغوية. يمكن الاطلاع على مبدأ التدوير في الشكل (١٩، ٥).



شكل (١٩، ٥). تدوير بروماكس للعامل.

التفصيل الثالث: علينا أن نقرر عدد العوامل التي نريد استخلاصها. تتمثل الفكرة في استخلاص أقل عدد ممكن من العوامل التي تفسر أكبر قدر ممكن من التنوع في البيانات، لأن النقطة الأساسية لتحليل العوامل هي اختزال عدد كبير من المتغيرات اللغوية إلى عدد قليل من العوامل الأساسية. نستخدم هنا أسلوب التمثيل الذي يُنتج مخططاً ركامياً يساعدنا على اتخاذ القرار. **المخطط الركامي scree plot** عبارة عن رسم بياني موضح في الشكل (٢٠، ٥)، ويقدم مؤشراً على عدد العوامل التي ينبغي أن نستخلصها، ويعرض عدد العوامل، ويمثل كلاً منها مثلث صغير على المحور السيني والقيم الكاملة على المحور الصادي. القيم الكاملة **Eigenvalue** مقياس لمدى التنوع في البيانات التي يوضحها العامل، فكلما زادت القيمة كان ذلك أفضل. من الناحية الرياضية: القيم الكاملة هي مجموع العوامل المربعة لجميع المتغيرات. إذا كان للمتغير عوامل مرتفعة العدد فذلك يُعدّ جزءاً مهماً من العامل. إذن، كيف تُعرف عدد العوامل التي ينبغي استخلاصها؟

لهذا السؤال إجابات مختلفة. يُنظر عادةً في العوامل ذات القيم الكاملة التي تزيد على (1)، ولمزيد من التحليل انظر (Kaiser 1960). إذا طبقنا هذا المعيار على المخطط الركامي في الشكل (٥, ٢٠) فسنستخلص خمسة عوامل. من المعايير الأخرى أن ننظر إلى نقطة الانعكاس في المخطط الركامي حيث ينتهي الانخفاض الحاد في القيم الكاملة، ويبدأ منحنى المخطط في العودة إلى الاعتدال. في الشكل (٥, ٢٠) نقطتان واضحتان للانقلاب تشيران إلى حل ثلاثي أو سداسي العوامل. يُستخدم التحليل الموازي parallel analysis في بعض الأحيان لإنشاء منطلق للقيم الكاملة التي حُصل عليها في المحاكاة الحاسوبية (مونت كارلو) مع متغيرات (عشوائية) غير ذات صلة. تحتاج العوامل المستخلصة إلى تحديد القيم الكاملة بوضوح فوق هذا المنطلق (Valero- Ledesma & Mora 2007؛ Hayton et al. 2004). يُمكننا في الشكل (٥, ٢٠) ملاحظة ما يصل إلى تسعة عوامل تفي بهذا المعيار. ينبغي أن ندرك بأننا نتعامل مع موقف يتطلب تعويض شيء بشيء. إذا احتجنا إلى مزيد من العوامل فعلينا أن نُفسّر ذلك بالتنوع. في المقابل، كلما تناقصت العوامل الكاملة كان تفسيرها صعباً (انظر: الخطوة الثالثة أدناه).



شكل (٥, ٢٠). استخلاص العامل: مخطط ركامي.

الخطوة الثالثة: التفسير الوظيفي للعوامل بوصفها الأبعاد

تتضمن هذه الخطوة التفسير الوظيفي للعوامل المستخلصة بوصفها أبعادًا. نحتاج هنا إلى أخذ كل من العوامل المستخلصة والنظر في عناصر العوامل factor loadings للمتغيرات الفردية التي نُخرجها تحليلات العامل. تُعدّ هذه العناصر مقياس مهمة للمتغيرات الفردية لعامل معين، وتُراوح عادة من (-1) إلى (+1)^(٨). لنوضح تفسير العوامل مع عرض العامل (1) في الجدول (٤، ٥). يمكننا ملاحظة ثلاثة أنواع من المتغيرات لها تأثير في هذا العامل. المجموعة أ: متغيرات ذات عناصر موجبة عالية تزيد على 0.35 في الخصائص من رقم ١ إلى ٥. المجموعة ب: متغيرات ذات عناصر سلبية عالية تزيد على 0.35 بالقيمة المطلقة في الخصائص من ٨ إلى ١٢. المجموعة ج: متغيرات ذات عناصر منخفضة للعوامل الإيجابية والسلبية، أي: أصغر من 0.35 بالقيمة المطلقة في الخاصية رقم ٦ و٧.

يمكن تفسير العامل (1) على النحو الآتي. أولاً، لننظر إلى المتغيرات ذات العناصر الإيجابية العالية في المجموعة أ. تتوارد هذه المتغيرات في النصوص، ويمكننا افتراض أن لديها وظيفة اتصالية مشتركة. عندما ننظر إلى مجموعة المتغيرات في هذه المجموعة يمكننا ملاحظة أن هذه المتغيرات تشير إلى لغة غير رسمية، بل تفاعلية للغاية مع الاستخدام المتكرر لضائير المتكلم والمخاطب، والاختصارات، والظروف، وthat المحذوفة. ثانياً، ننظر إلى المتغيرات ذات العناصر السلبية للعامل في المجموعة ب التي لها توزيع تكميلي للمتغيرات ذات التحويلات الموجبة. نجدها ترد وروداً غير متكرر في النصوص التي ترد فيها المتغيرات من المجموعة (أ) بتكرار مرتفع والعكس بالعكس. تشترك المتغيرات من المجموعة (ب) في وظيفة التواصل التي يمكن أن توصف بأنها (وصف من النوع الأكاديمي) مع صيغ البناء للمجهول والأسماء المحوَّرة والصيغ الاسمية التي تنتهي باللاحق -tation و-ment و-ness و-ity. يفسر بايبر (1988) العامل الأول استناداً إلى نتائج مماثلة كبعد ذي إنتاج لغوي تواصلِي INVOLVED (متغيرات المجموعة أ) مقابل إنتاج لغوي معلوماتي INFORMATIONAL (متغيرات المجموعة ب).

يمكن تجاهل المتغيرات ذات العناصر المنخفضة في المجموعة (ج) لأنها لا تسهم إسهاماً كبيراً في هذا العامل. يكون لهذه المتغيرات عادة تحميلات عالية على عوامل أخرى، ويُعدّ كل متغير وارد مرة واحدة فقط جزءاً من العامل الذي يكون له أعلى حمولة يزيد على 0.35 بالقيمة المطلقة.

(٨) لاحظ أنه في حالة تدوير بروماكس لكونه يُسمح للعوامل بالترابط، يمكن أن تكون عناصر العامل أكبر من 1 بالقيمة المطلقة.

جدول (٤، ٥). العامل 1: تمحيلات المتغيرات الفردية.

المتغيرات	عناصر العامل	الخصائص
المجموعة أ	0.906	١- الاختصارات
	0.67	٢- ضمير المتكلم
	0.605	٣- ضمير المخاطب
	0.576	٤- الظروف
	0.546	٥- حذف that
المجموعة ج	0.219	٦- ضمائر الملكية
	-0.227	٧- Gerund (اسم المصدر)
المجموعة ب	-0.408	٨- التحويل الاسمي
	-0.465	٩- الصفة الوصفية
	-0.518	١٠- المبني للمجهول
	-0.607	١١- الجاز والمجرور
	-0.778	١٢- أساء أخرى

اختُصرت قائمة المتغيرات لأجل التوضيح.

الخطوة الرابعة: وضع السجلات على الأبعاد

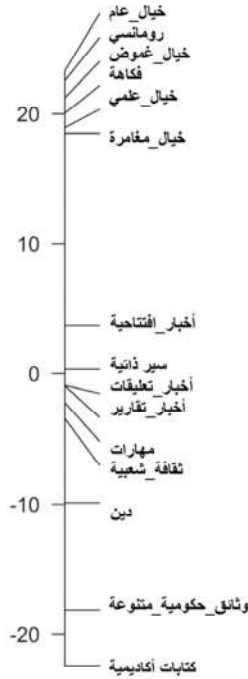
تتضمن الخطوة الأخيرة وضع السجلات على الأبعاد. يمكن أن تساعد هذه الخطوة أيضًا على تفسير الوظائف التواصلية للأبعاد من الخطوة الثالثة بالنظر في أنواع النصوص التي ترد فيها المتغيرات التي تُحمل بتحميلات أكثر على كل عامل. نحتاج أولاً إلى توحيد مجموعة البيانات (الشكل ١٧، ٥) بحساب قيم-ز٢. قيم-ز٢ هي قيم قياسية تستخدم الانحراف المعياري بوصفه وحدة قياس، وتُستخدم لإعطاء كل متغير الثقل (التأثير) نفسه بغض النظر عما إذا كانت الخاصية اللغوية متكررة أو نادرة أو مُقاسة على المقياس نفسه أو على مقياس مختلف. تُحسب قيم-ز٢ باستخدام المعادلة (٩، ٥) بالمبحث (٣، ٥). بعد ذلك تُحسب قيم البعد dimension scores لكل نص على النحو الآتي:

(٥، ١٤)

قيمة البعد للنص = المتغير 1 المجموعة 1 + المتغير 2 المجموعة 1 ...

-المتغير 1 المجموعة ب - المتغير 2 المجموعة ب

قيمة البُعد هي مجاميع قيم-ز₂ للمتغيرات ذات التحميلات أو العناصر الإيجابية العالية (المجموعة أ) مطروحاً منها قيم-ز₂ للمتغيرات ذات التحميلات أو العناصر السلبية العالية (المجموعة ب). أخيراً، يُحسب متوسط قيمة البعد لكل سجل لغوي بأخذ جميع درجات أبعاد النص التي تنتمي إلى السجل اللغوي نفسه ثم حساب متوسط القيمة. يوضع كل سجل على مقياس أحادي البعد وفقاً للقيمة المتوسطة الناتجة لقيمة البعد كما هو ظاهر في الشكل (٢١، ٥).



اختبار التباين الأحادي: ف (14, 485) = 64.42، قيمة الاحتمال < 0.001 (م. ر. ت. ر. ب. = 65.03%)

شكل (٢١، ٥). متوسط قيم السجلات اللغوية الموضوع على البُعد 1: التواصلية مقابل المعلوماتية.

يمكننا ملاحظة أن السجلات اللغوية ذات القيم العالية للبعد 1 هي أنواع مختلفة من الرواية يحدث فيها كثير من التفاعلات، لذلك توضع في اتجاه الطرف التواصلية من البعد. من ناحية أخرى تتمتع الوثائق الحكومية والكتابة الأكاديمية بقيمة منخفضة للبعد، وتتجمع نحو الطرف المعلوماتية للبعد. أما السجلات اللغوية الأخرى فتقع في مكان ما في الوسط، بعضها أقرب إلى الطرف التواصلية وبعضها الآخر أقرب إلى الطرف المعلوماتية.

لاختبار الفروق الدلالية الإحصائية بين السجلات اللغوية الموضوعة على البعد 1 تُحسب تحليل التباين الأحادي one-way ANOVA (انظر: المبحث ٣, ٦ لمزيد من المعلومات)؛ ويتم رصد (ر²) أيضا لبيان مقدار التنوع في السجلات اللغوية الذي يوضحه هذا العامل. تتكرر عملية وضع السجل مع كل العوامل المستخلصة.

التقرير الإحصائي: التحليل متعدد الأبعاد

١ - ما يجب الإفصاح عنه

يتضمن التحليل متعدد الأبعاد عدداً من الخطوات المعقدة والإجراءات الإحصائية التي ينبغي الإفصاح عنها من أجل التماثل الكامل للنتائج. يعتمد نجاح التحليل متعدد الأبعاد أيضاً على موثوقية التحديد التلقائي للمتغيرات اللغوية (الوسم) في المدونات. وعليه، ينبغي الإفصاح عن المعلومات التالية:

أولاً: يجب إدراج جميع المتغيرات اللغوية مع المعلومات الخاصة بمقدار تداخلها مع المتغيرات الأصلية لباير (1988). يمكن استخدام رقم من قائمة هذه المتغيرات في الجدول (٣, ٥) بالإحالة إلى متغيرات باير، كالرقم (٣) مثلاً الذي يشير إلى الزمن المضارع. لاحظ أن هذا التحليل المتغير الأبعاد ما هو إلا إجراء عام لا يخص اللغة الإنجليزية وحدها، لأنه من الممكن تحديد متغيرات مماثلة ذات صلة بلغات أخرى.

ثانياً: إجراء الواسم أيا كان هذا الواسم أو استخدام الواسمات الجاهزة عند الاقتضاء مع التحقق من موثوقيته.

ثالثاً: نوع التحليل لمتعدد الأبعاد، أي: التحليل متعدد الأبعاد الكامل أو المقارنة بأبعاد باير (1988).

رابعاً: دوران تحليل العوامل (مثل بروماكس أو فاريماكس) وعدد العوامل المستخلصة.

خامساً: النتائج: (أ) تحميلات أو عناصر العامل، (ب) مخططات البعد، (ج) تحليل التباين ور².

يصف مبحث الطريقة في التقرير البحثي الإجراءات التحليلية والبارامترات المستخدمة. ينبغي مناقشة كل مخططات البعد بدقة في النتائج. السؤال الرئيسي الذي يتعين معالجته هو: ما التنوع الوظيفي الأساسي في السجلات اللغوية التي كشفت عنه الطريقة المستخدمة؟

٢- مثال الإفصاح في التقرير

انظر المبحث (٥, ٥).

(٥, ٥) التطبيق: السجلات اللغوية في الإنجليزية النيوزيلندية

هذه الدراسة الدقيقة مستوحاة مباشرة من استكشاف ريتشارد شياو متعدد الأبعاد للهجات اللغة للإنجليزية في العالم (2009 Xiao). نظر شياو إلى خصائص اللغة الإنجليزية المستخدمة في بريطانيا وهونج كونج والهند والفلبين وسنغافورة، وتركز هذه الدراسة في اللغة الإنجليزية في نيوزيلندا.

جدول (٥, ٥). السجلات اللغوية في مدونة ICE-NZ.

السجل اللغوي	عدد النصوص
1. حوار_خاص_منطوق (محادثات)	100
2. حوار_عام_منطوق	80
3. مونولوج_مرتجل_منطوق	70
4. مونولوج_مكتوب_منطوق	50
5. كتابة_طالب_مكتوبة_غير مطبوعة	20
6. خطابات_مكتوبة_غير مطبوعة	30
7. نص_أكاديمي_مكتوب_مطبوع	40
8. نص_شعبي_مكتوب_مطبوع	40
9. تحقيق_صحفي_مكتوب_مطبوع	20
10. نص_تعليمي_مكتوب_مطبوع	20
11. نص_إقناعي_مكتوب_مطبوع	10
12. نص_إبداعي_مكتوب_مطبوع	20

تستند هذه الدراسة إلى مدونة ICE-NZ (Vine 1999)، وهي مكونة من مليون كلمة من اللغة الإنجليزية، ٤٠٪ منها مكتوبة و ٦٠٪ منها منطوقة وأنتجت في نيوزيلندا. يوضح الجدول (٥, ٥) توزيع السجلات في هذه المدونة.

تبدأ نقطة الانطلاق بأخذ متغيرات باير (1988) السبعة والستين في الحسبان. اختير من بينها أربعة وأربعون متغيراً غير مرتبط ببنية الجملة في المدونة لتفادي مشكلات تحديد المتغيرات في المدونات الفرعية المنطوقة من مدونة ICE-NZ^(٩). فيما يأتي قائمة هذه المتغيرات، وتشير الأرقام الموجودة بين قوسين إلى ترتيب باير:

- الزمن الماضي (1)، الجهوية التامة (2)، الزمن الحاضر (3)، ظروف المكان (4)، ظروف الزمان (5)، ضمائر المتكلم (6)، ضمائر المخاطب (7)، ضمائر الغائب الشخصية (8)، الضمير it (9)، ضمائر الإشارة (10)، الضمائر المبهمة (11)، فعل do النيابي (12)، التحويلات الاسمية (14)، مجموعة الأسماء الأخرى (16)، الجمل المبنية للمجهول دون نائب فاعل (17)، الجمل المبنية للمجهول مع نائب فاعل (18)، فعل be الرئيسي (19)، الضمير الوجودي there (20)، الروابط الظرفية السببية (35)، الروابط الظرفية التناقضية (36)، الروابط الظرفية الشرطية (37)، الروابط الظرفية الأخرى (38)، أسماء الجر والمجرورات (39)، الصفات الوصفية (40)، الصفات الخبرية (41)، مجموعة الظروف (42)، نسبة الكلمة النوعية / الكلمة الفعلية (43)، متوسط طول الكلمة (44)، ظروف الوصل (45)، المُلطفات (46)، تعبيرات الاحتراز (47)، المقويات (48)، المؤكدات (49)، أسماء الإشارة (51)، الأفعال الناقصة الاحتمالية (52)، الأفعال الناقصة الدالة على الضرورة (53)، الأفعال الناقصة التنبؤية (54)، الأفعال العامة (55)، الأفعال الخاصة (56) الأفعال الإقناعية (57)، appear و seems (58)، الاختصارات (59)، النفي التركيبي (66)، النفي التحليلي (67)

جرى النظر في العديد من المتغيرات اللغوية النموذجية الأخرى للسياق النيوزيلندي إلا أنها استُبعدت في النهاية. على سبيل المثال: وجود كلمات ماورية (Māori: كلمات نيوزيلندية) في الجزء المنطوق من المدونة. لم يكن هذا المتغير متعلقاً بالمتغيرات الأربعة والأربعين الأخرى، كما أن جميع الارتباطات لها صغيرة جداً، أي: أقل من 0.1^(١٠).

(٩) على سبيل المثال: المتغيرات مثل روابط الجملة، وأدوات الخطاب الاستهلاكية للجملة، وحروف الجر في الموضع النهائي، وكلها تعتمد على وضع علامات واضحة لحدود الجملة، ولا تتوفر في المدونة الفرعية المنطوقة من مدونة ICE-NZ لأنها تعتمد على أدوار المتكلم.

(١٠) يرتبط وجود كلمات الماورية بعرق المتحدث الذي هو من أصل ماوري (r = 0.436. مجال الثقة 95٪ [0.34, 0.524]).

قبل وسم أجزاء الكلام عولجت ملفات المدونة مسبقاً وحذفت جميع الوسوم الكبيرة mega-tags والتعليقات. إضافةً إلى ذلك، في المدونة الفرعية المنطوقة، أزيلت أيضاً الأصوات المصاحبة للغة مثل er و um و mm والكلمات غير المكتملة لمنع أخطاء وسم أجزاء الكلام وتضخم عدد المتغيرات^(١١). في الجزء المنطوق وُضعت جميع أمثلة ضمير المتكلم بحرف كبير بدلاً من الحرف الصغير المعتاد إجراءه عادة في معالجة النصوص الإنجليزية، فحوّل الضمير (i) إلى (I) لأجل تقليل عدد أخطاء وسوم ستانفورد لأقسام الكلام، إذ تجعل هذه الوسوم الحرف الصغير (i) كلمة أجنبية.

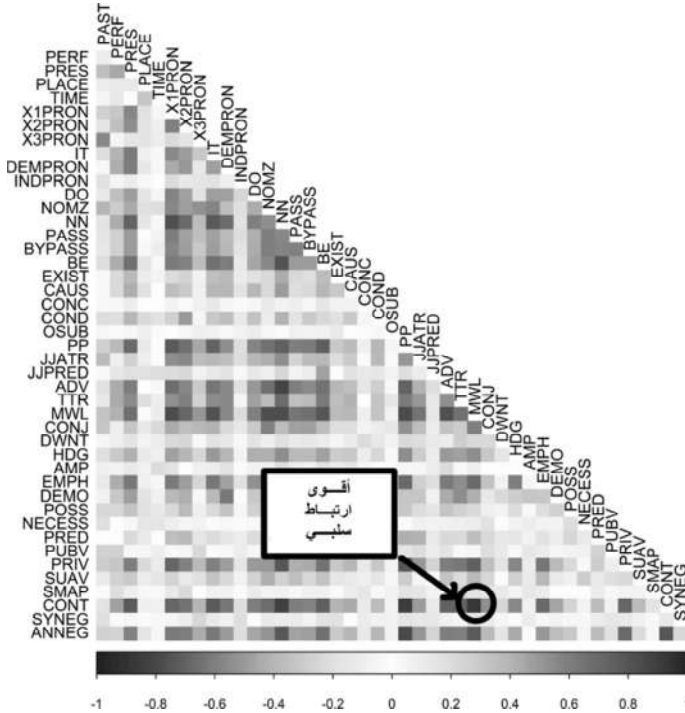
استُخدم محلل مات MAT (Nini 2015)، وهو وحدة مبنية على واسم ستانفورد، لتحديد المتغيرات الأربعة وأربعين في المدونة. فُحصت نتائج إجراء الوسم بحثاً عن أي أخطاء منهجية في الوسم أو أخطاء في تحديد المتغير (انظر: Leech et al. 1994 و Manning 2011 و Vine 2011 للاطلاع على مناقشة أخطاء الواسمات)، ثم عولجت البيانات باستخدام أدوات تحليل الارتباط وأدوات تحليل التجميع وأدوات التحليل المتعددة الأبعاد من أدوات لانكستر الإحصائية Lancaster Stats Tools.

عند التعامل مع متغيرات لغوية متعددة فإن الخطوة الأولى هي النظر إلى الارتباط بين هذه المتغيرات، ولهذا الغرض أُنتجت مصفوفة ارتباط 44×44 (الشكل ٢٢، ٥) تُصوّر علاقات بيرسون بين المتغيرات الأربعة والأربعين. تدل شدة التظليل على قوة الارتباط. بالنظر إلى الشكل فإن عدداً كبيراً من الارتباطات معتدلة أو قوية، وهذا أحد المتطلبات الأساسية لإجراء تحليل العوامل بنجاح. في مصفوفة الارتباط حُدّد أقوى ارتباط سلبي بين متوسط طول الكلمة وبين استخدام الاختصارات ليكون: ($r = -.885$ ، مجال الثقة ٩٥٪ $[-.903, -.865]$).

يشير هذا الارتباط السلبي إلى أن النصوص الموجودة في المدونة التي تحتوي على كلمات أطول في متوسطها (وفقاً لعدد الأحرف) لا تكاد تحتوي على صيغ مختصرة والعكس صحيح. هذا الارتباط منطقي من الناحية اللغوية عندما تزداد الكلمات الأكثر حروفاً في النصوص الرسمية التي تستخدم الصيغ المختصرة استخداماً نادراً. تحتوي السجلات المنطوقة غير الرسمية على العديد من الاختصارات والكلمات الأقصر. يمكن تصور العلاقة بين الاختصارات ومتوسط طول الكلمة باستخدام مخطط التبعثر في الشكل (٢٣، ٥).

(١١) على سبيل المثال: توّسم في الأغلب علامات التردد er و um بواسم خطأ كأن تُجعل أسماء أو صفات أو أفعال في السياق النحوي.

هنا يمكننا ملاحظة أن هذين المتغيرين وحدهما يمكنهما المساعدة على التمييز بين السجلات اللغوية المختلفة. في الزاوية اليسرى العليا من الحيز ثنائي الأبعاد جُمعت المحادثات غير الرسمية. عندما نتحرك لأسفل المائل يمكننا ملاحظة تجميعات متداخلة من الحوارات العامة، والمونولوجات المرتجلة، والرسائل والكتابات الإبداعية (يسارًا) وكذلك المونولوج المكتوب (يمينًا)، وهي تتميز بتناقص عدد الاختصارات تدريجيًا وزيادة متوسط طول الكلمة.

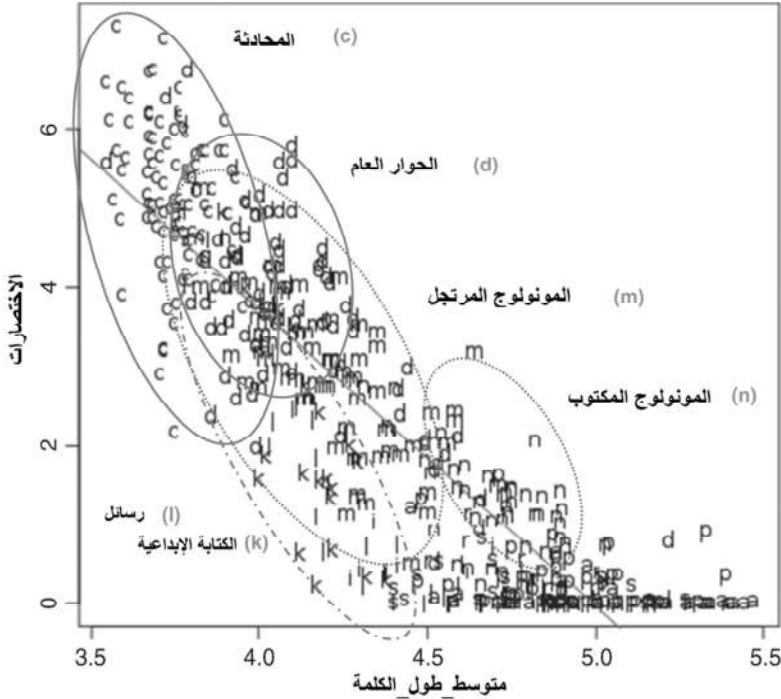


شكل (٢٢، ٥). مصفوفة الارتباط: ٤٤ متغيرًا.

عندما نتجاوز الحيز الثنائي الأبعاد بالشكل (٢٣، ٥) ونستخدم جميع المتغيرات الأربعة والأربعين فإنه يمكننا التمييز بين السجلات اللغوية بمزيد من الدقة. للقيام بذلك استُخدمت أولاً تقنية التجميع المتكامل الهرمي^(١٢) مع مجموعة البيانات بأكملها (٤٤ متغيرًا بدون تحويل، ومع مسافة مانهاتن أو طريقة وارد Ward) ليظهر بعد ذلك المخطط التجميعي الناتج كما في الشكل (٢٤، ٥).

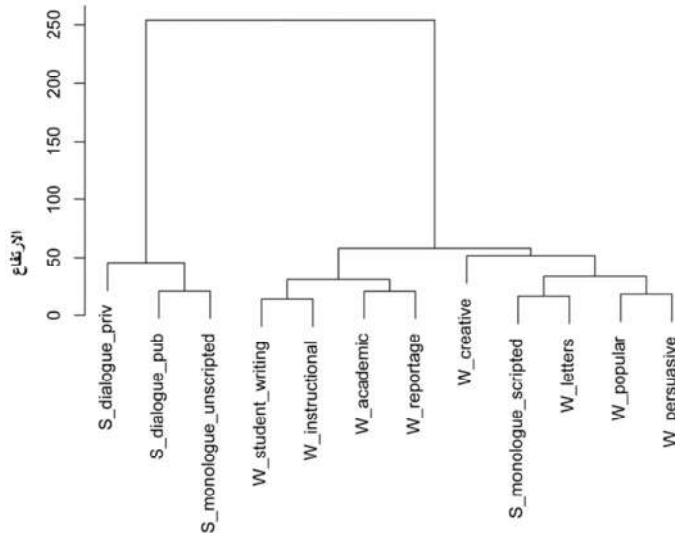
(١٢) تعتمد التجميعات على القيم المتوسطة لكل سجل لغوي، وقد حُسبت هذه المتوسطات قبل تحليل التجميع.

يُمكننا في الشكل (٢٤، ٥) ملاحظة انقسام واضح بين الكلام والكتابة ما عدا المونولوج المنطوق المكتوب الذي يتكثّل مع الحروف. من المتوقع أن يتقارب المونولوج المكتوب من الكتابة لأن الكلام المكتوب نوع معين من أنواع النصوص المكتوبة. ظهرت ثلاث تجميعات رئيسية؛ الأول: تجميع منطوق (حوار خاص وعام ومونولوج مرتجل)، والثاني: تجميع كتابة أكثر رسمية (كتابة طلاب، وكتابة تعليمية، وكتابة أكاديمية، وكتابة التقارير)، والثالث: كتابة أقل رسمية مع تجميع المونولوج المكتوب (الكتابة الإبداعية، وكتابة الرسائل، والكتابة الشعبية، والكتابة الإقناعية والمونولوج المكتوب). يمكننا أيضًا أن نرى الهيكل الداخلي لهذه التجميعات بالنظر إلى الارتفاع على المحور الصادي من المخطط، وهو ما يدل على المسافة بين السجلات اللغوية في مساحة تتضمن أربعاً وأربعين بعداً. على الرغم من أن تقنية التجميع فعّالة فهي لا تفصل بين المجموعات المختلفة من المتغيرات اللغوية (العوامل) التي تميز السجلات اللغوية الفردية. يُختزل تحليل التجميع تلك المساحة من أربعة وأربعين بعداً إلى بعدٍ واحد فقط يجمع بين إسهامات جميع المتغيرات الأربعة والأربعين، ثم تُمثّل في صورة مخطط فقرعي.



شكل (٢٣، ٥). الارتباط بين متوسط طول الكلمة والاختصارات: تجميعات السجل اللغوي.

أُجري تحليل متعدد الأبعاد لاستكشاف مزيد من الأبعاد في البيانات، وبسبب قيود مساحة الكتاب سيجري الإفصاح عن مجموعة أول عاملين أو بُعدين فقط من نتائج هذا التحليل. لم تكن مجموعة البيانات لمتغيرات باير البالغ عددها سبعة وستين متاحة، وعليه أُجرى تحليل العامل باستخدام تدوير بروماكس promax rotation لاستخراج أربعة عوامل فُسرت بعد ذلك بوصفها أبعادًا. استند الإفصاح بشأن عدد العوامل المراد استخلاصها إلى نقطة الانعكاس الظاهر في المخطط الركامي، والجدول (٦، ٥) يوضح العاملين 1 و2 بما في ذلك تحميلات المتغيرات الفردية التي تزيد على 0.35 بالقيمة المطلقة.



تجميع الارتفاع للمفرد

[الترجمة]:

dialogue_priv	حوار خاص	monologue_unscripted	مونولوج مرتجل	instructional	تعليمي	S	منطوق
dialogue_pub	حوار عام	monologue_scripted	مونولوج مكتوب	student_writing	كتابة الطالب		
academic	أكاديمي	reportage	تحقيق صحفي	creative	إبداعي	W	مكتوب
letters	خطابات	popular	شعبي	persuasive	إقناعي		

شكل (٢٤، ٥). مخطط التجميع: سجلات لغوية في الإنجليزية النيوزيلندية.

يمكننا ملاحظة أن العامل الأول يتضمن متغيرات أعلى من 0.35 بالقيمة المطلقة من العامل الثاني. تشتمل التحويلات الإيجابية للعامل (1) على أفعال مضارعة، والفعل be، والاختصارات، والظروف، والأفعال الخاصة، والضائر، إلخ. تشمل التحويلات السلبية استخدامًا مكثفًا للاسم، وحروف الجر، والكلمات الأطول، والكلمات النوعية مع الكلمات الفعلية، واستخدام المبني للمجهول، إلخ. تشير الخصائص الإيجابية من الناحية الوظيفية إلى تواصلية المتحدثين involvement العالية والأسلوب غير الرسمي. أما الخصائص السلبية فتُظهر في سجلات لغوية أكاديمية أكثر رسمية، وثرية بالمحتوى المعلوماتي. نظرًا لأن توزيع الخصائص يتوافق تمامًا مع البعد (1) لبابير (١٩٨٨) فقد احتُفظ بمصطلحه لهذا البعد: (تواصلية مقابل معلوماتي involved vs informational). يوضح الشكل (٢٥، ٥) توزيع السجلات اللغوية الفردية على طول البعد (1) عندما تكون السجلات اللغوية المنطوقة (ما عدا المونولوج المكتوب^(١٣)) أقرب إلى الطرف التواصلية للبعد من السجلات اللغوية المكتوبة. وكما هو متوقع فالسجل اللغوي الأكثر تواصلية هو الحوار الخاص المنطوق (المحادثات)، والسجل اللغوي الأكثر معلوماتية هو التقارير الصحفية مع الكتابة الأكاديمية. تجدر الإشارة أيضًا إلى أن للاختصارات ومتوسط طول الكلمة في المتغيرين ذوي العلاقة السلبية الأقوى المذكورة سابقا (الشكل ٢٢، ٥)، تحميلات عالية جدًا على البعد (1). هذا يعني أن هذين المتغيرين يمكنهما التمييز بفعالية بين السجلات اللغوية الفردية بمصطلحات يسيرة للغاية، وعليه فالشكل (٢٥، ٥) امتداد لمخطط التبثر البسيط من الشكل (٢٣، ٥).

جدول (٦، ٥). نتائج تحليل العامل للغة الإنجليزية النيوزيلندية: تحميلات العامل.

الخصائص	تحميلات العامل 1	الخصائص	تحميلات العامل 2
الزمن المضارع (3)	0.879	الأفعال الناقصة التنبؤية (54)	0.75
الفعل be الرئيسي (19)	0.848	الشرطية مقابل عطف النسق (37)	0.7
الاختصارات (59)	0.846	الأفعال الإقناعية (57)	0.57
مجموعة الظروف (42)	0.802	الأفعال الناقصة الاحتمالية (52)	0.56
المؤكّدات (49)	0.798	الأفعال الناقصة الإلزامية (53)	0.53
الضمير it (9)	0.746	ضائر المخاطب (7)	0.358]

(١٣) لاحظ أن المونولوج المكتوب يتكامل مع السجلات اللغوية المكتوبة في المخطط التفرعي في الشكل 5.24.

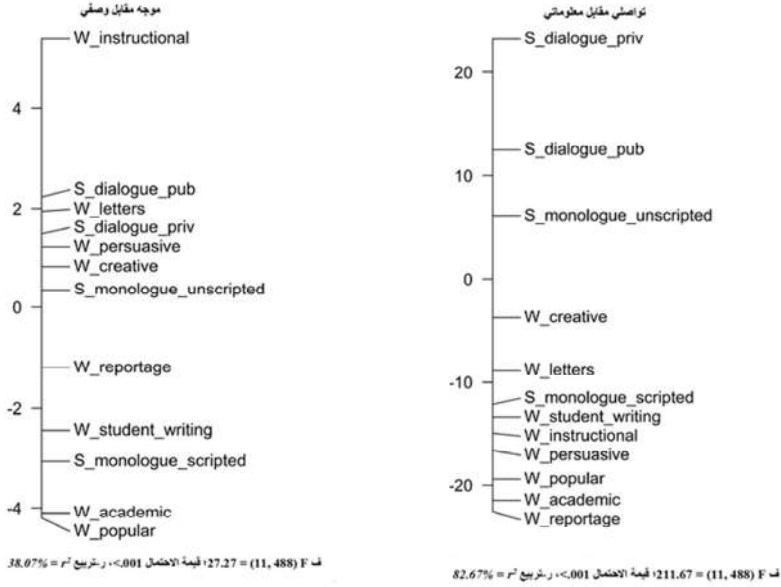
تابع جدول (٦, ٥).

الخصائص	تحميلات العامل 1	الخصائص	تحميلات العامل 2
ضمائر الإشارة (10)	0.739	الصفات الوصفية (40)	-0.453
الأفعال الخاصة (56)	0.713		
ضمائر المتكلم (6)	0.67		
النفي التحليلي (67)	0.631		
أسماء الإشارة (51)	0.629		
التحرزات (47)	0.596		
فعل do النيابي (12)	0.572		
ضمائر المخاطب (7)	0.537		
السببي مقابل عطف النسق (35)	0.511		
There الوجودية (20)	0.472		
المبني للمجهول مع نائب الفاعل (18)	-0.492		
التحويلات الاسمية (14)	-0.512		
الأفعال الإقناعية (57)	-0.519		
المبني للمجهول دون نائب الفاعل (17)	-0.558		
الجهوية التامة (2)	-0.632		
نسبة الكلمة النوعية/ الكلمة الفعلية (43)	-0.692		
مجموع الجار والمجرور (39)	-0.693		
متوسط طول الكلمة (44)	-0.74		
مجموع الأسماء الأخرى (16)	-0.875		

^١ تشير الأرقام بين الأقواس إلى ترتيب عرض متغيرات بايبر (١٩٨٨) الأصلية التي تستند إليها هذه القائمة (انظر: الجدول ٣, ٥).

بالنظر إلى العامل (2) تشتمل أكبر التحميلات الإيجابية على أنواع مختلفة من الأفعال الناقصة، و if و unless الظرفيتين الشرطيتين، وكذلك على أفعال مثل command و propose و recommend. استبعدت ضمائر المخاطب لأن لها حمولة أعلى على العامل (1). الحمولة السلبية الوحيدة هي الصفات الوصفية. تتمثل الوظيفة الأساسية للخصائص الإيجابية في الإنتاج الموجه (الأفعال الناقصة) الذي هو من سمات النصوص التعليمية المكتوبة، أما الجانب الآخر من البعد فهو الإنتاج الوصفي الذي هو من سمات الكتابة الأكاديمية والنصوص المعلوماتية الشائعة. يمكننا

استخدام التسمية (إنتاج موجه مقابل إنتاج وصفي). لاحظ أن هذا البعد مختلف تمامًا عن البعد (2) لباير (1988) الذي يتضمن الخطاب السردي مقابل غير السردية، وخصائص باير السردية وغير السردية بارزة في اللغة الإنجليزية النيوزيلندية وتظهر أيضًا في العامل 3 (لم تناقش هنا).



شكل (٥, ٢٥). البعد الأول: الإنجليزية النيوزيلندية شكل (٥, ٢٦). البعد الثاني: الإنجليزية النيوزيلندية – تحليل كامل متعدد الأبعاد.

dialogue_priv	حوار خاص	monologue_unscripted	مونولوج مرتجل	instructional	تعليمي	S	منطوق
dialogue_pub	حوار عام	monologue_scripted	مونولوج مكتوب	student_writing	كتابة الطالب		
academic	أكاديمي	reportage	تحقيق صحفي	creative	إبداعي	W	مكتوب
letters	خطابات	popular	شعبي	persuasive	إقناعي		

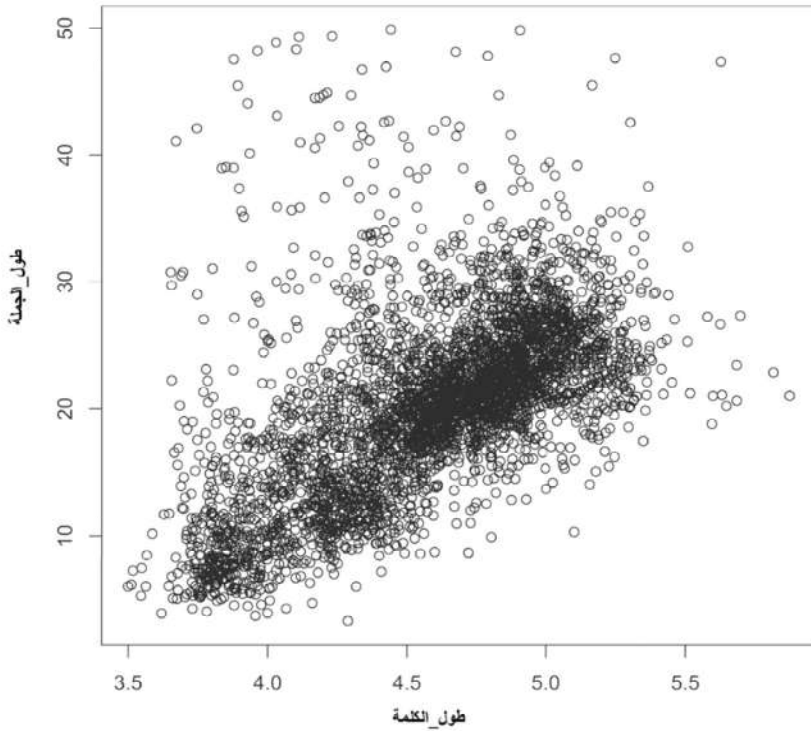
أخيراً، تجدر الإشارة إلى أن كلا البعدين يُظهران فروقاً ذات دلالة إحصائية بين السجلات اللغوية على النحو الذي حدده تحليل التباين الأحادي. في حالة البعد (1) يُفسَّر ٨٣٪ تقريباً من التنوع في قيم البعد للنصوص الفردية من خلال مفردات السجل الخاصة به، أما في حالة البعد (2) فهذا الرقم أقل من ٤٠٪، ومن ثم فالبعد (1) يعد مؤشراً أقوى على تنوع السجل اللغوي من البعد (2).

(٥, ٦) تمارين

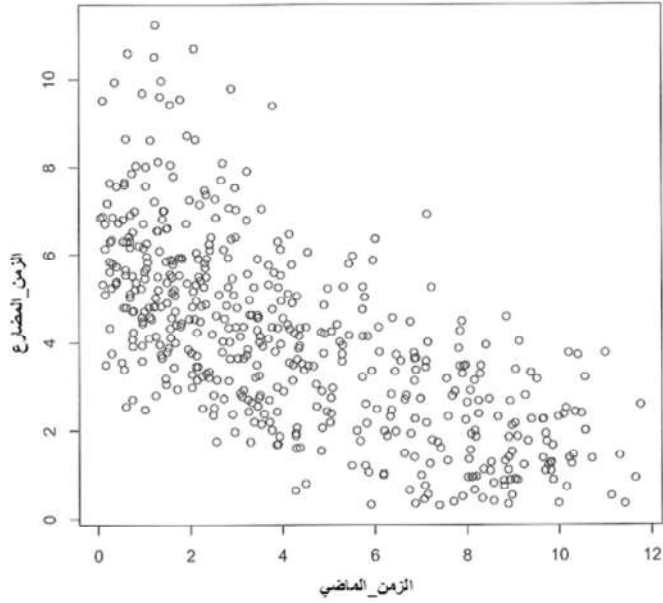
١- احسب ارتباطات بيرسون وسبيرمان يدويًا بين الأفعال والصفات في عشرة نصوص مختارة عشوائيًا من مدونة BE06.

الأفعال	165.2	218.0	213.9	140.7	190.8	163.1	183.0	161.7	135.0	169.9
الصفات	76.5	74.4	60.3	68.4	77.6	98.8	76.5	91.9	102.6	96.0

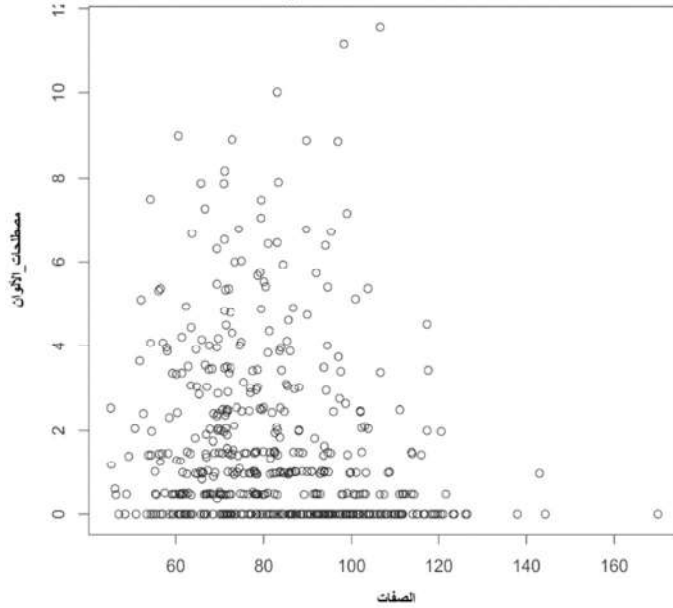
٢- ماذا يمكنك أن تقول عن العلاقة بين المتغيرات في الرسوم البيانية الأربعة في الأشكال (٥, ٢٧) و(٥, ٢٨) و(٥, ٢٩) و(٥, ٣٠)؟



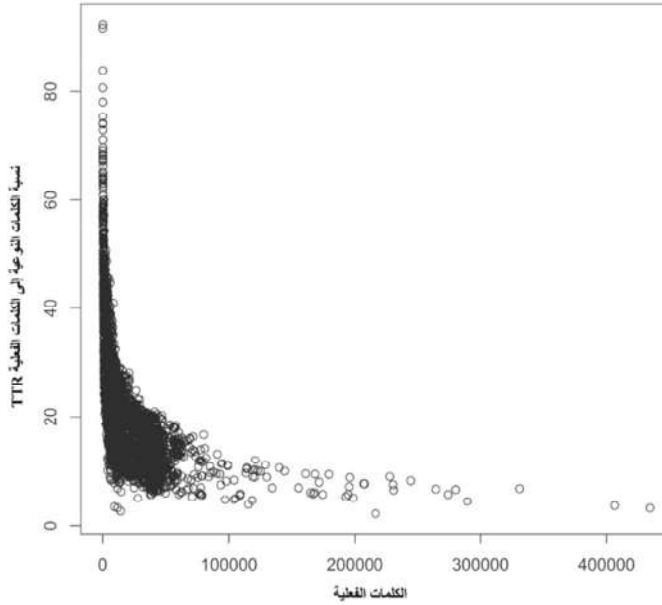
شكل (٥, ٢٧). العلاقة بين متوسط طول الكلمة (عدد الأحرف) ومتوسط طول الجملة (عدد الكلمات) في المدونة الوطنية البريطانية .



شكل (٢٨، ٥). العلاقة بين استخدام الماضي والمضارع في مدونة BE06.



شكل (٢٩، ٥). العلاقة بين استخدام الصفات ومصطلحات اللون في مدونة BE06.



شكل (٣٠، ٥). العلاقة بين طول النص (الكلمات الفعلية) ونسبة الكلمات النوعية/ الفعلية (TTR) في المدونة الوطنية البريطانية

٣- تنقسم كل مدونة من عائلة براون إلى ١٥ وعاءاً مختلفاً من النصوص المدرجة أدناه (انظر أيضاً في المبحث ٤، ١).

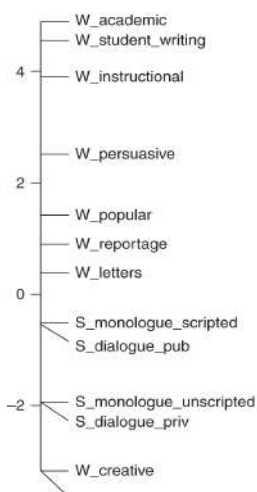
أ (صحفي: تحقيق صحفي)، ب (صحفي: افتتاحية) ج (صحافة: تقييمات نقدية)، د (ديني)، هـ (مهارات ومهن وهوايات)، و (الثقافة الشعبية)، ز (الأدب البحث، والتراجم، والمقالات)، ح (وثائق حكومية متنوعة، تقارير مؤسسية، تقارير صناعية، جدول جامعي، مجلة دار الصناعة)، ط (كتابات مستنيرة وعلمية)، ي (كتابات مستنيرة وعلمية)، ك (الغموض والرواية البوليسية)، ل (الخيال العلمي)، م (المغامرات والرواية الغربية)، ن (القصص الرومانسية والغرامية)، س (الفكاهة).

هذا التصنيف مفيد جداً وقد يكون مفصلاً لغرض ما. جمع أنواع النصوص الفردية في فئات أكبر بناءً على تشابهها الوظيفي، ثم صمم دراسة يمكنك من خلالها التحقق من مجموعتك.

٤- يقدم الجدول (٥,٧) تحميلات العامل (3) و(4) استنادًا إلى التحليل المتعدد الأبعاد للغة الإنجليزية النيوزلندية بالمبحث (٥,٥). أوردنا مخططات البعد أيضًا في الشكلين (٥,٣١) و(٥,٣٢). فسر كل عامل بوصفه بُعدًا وظيفيًا. ضع تسميات لهذه الأبعاد.

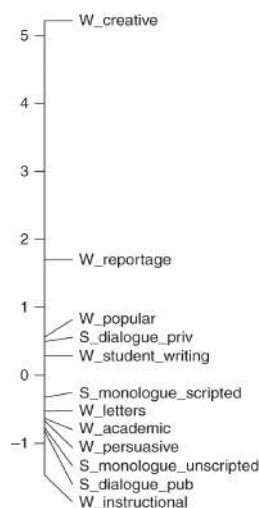
جدول (٥,٧). نتائج تحليل عوامل الإنجليزية النيوزلندية: تحميلات العامل (3) و(4).

التحميلات العامل 4	الخصائص	تحميلات العامل 3	الخصائص
0.618	التحويلات الاسمية (14)	1.099	الزمن الماضي (1)
0.499	ظروف الوصل (45)	0.461	ضمائر الغائب الشخصية (8)
0.36	المبني للمجهول دون نائب فاعل (17)	-0.304	الصفات الوصفية (40)
0.347	المبني للمجهول مع نائب فاعل (18)	-0.583	الزمن المضارع (3)
-0.444	ظروف الزمان (5)		
-0.504	ظروف المكان (4)		



لف F (11, 488) = 144.69، قيمة الاحتمال < .001، رتب مربع $r^2 = 56.18\%$

شكل (٥,٣٢). البعد الرابع.



لف F (11, 488) = 25.25، قيمة الاحتمال < .001، رتب مربع $r^2 = 36.27\%$

شكل (٥,٣١). البعد الثالث.

Dialogue_ priv	حوار_خاص	Monologue_ unscripted	مونولوج_مرئجل	instructional	تعليمي	S	منطوق
Dialogue_ pub	حوار_عام	Monologue_ scripted	مونولوج_مكتوب	Student_ writing	كتابة_الطالب	W	مكتوب
academic	أكاديمي	reportage	تحقيق صحفي	creative	إبداعي		
letters	خطابات	popular	شعبي	persuasive	إقناعي		

٥- استخدم البيانات المقدمة على موقع الويب المصاحب للكتاب وأداة التحليل المتعدد الأبعاد لمقارنة السجلات اللغوية باللغة الإنجليزية البريطانية والأمريكية الحالية.

تذكر الآتي

- تُستخدم الارتباطات للتحقيق في العلاقة بين متغيرين في آنٍ واحد.
- ارتباط بيرسون مناسب لمتغيرات المقياس، في حين يناسب ارتباط سيرمان المتغيرات الترتيبية (الرتب). يمكن أيضًا استخدام ارتباط سيرمان مع متغيرات المقياس إذا كانت المتوسطات بوصفها قياسات لنزعة المركزية لا تمثل البيانات جيدًا (توزيعات منحرفة للغاية).
- يستخدم تحليل التجميع المتكامل الهرمي لتصنيف الكلمات والنصوص والسجلات اللغوية وما إلى ذلك. نتيجة هذا التحليل تكون في مخطط تفرعي.
- أشد أنواع التحليل تعقيدًا من بين الثلاثة التي نوقشت في هذا الفصل هو التحليل المتعدد الأبعاد (MD) الذي يُحلل عددًا كبيرًا من المتغيرات اللغوية ويختزلها في عدد صغير من العوامل التي تُفسّر بأنها أبعاد مختلفة، ويمكن إضافة سجلات لغوية أخرى مع هذه الأبعاد.

قراءات إضافية

- Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press.
- _____. (2012). Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory*, 5(1), 9-37.
- Biber, D., Conrad, S., Reppen, R., Byrd, P. & Helt, M. (2002). Speaking and writing in the university: a multidimensional comparison. *TESOL Quarterly*, 36(1), 9-48.
- Chen, P. Y. & Popovich, P. M. (2002). *Correlation: parametric and nonparametric measures*. Thousand Oaks, CA: Sage.
- Conrad, S. & Biber, D. (2001) Multidimensional methodology and the dimensions of register variation in English. In S. Co mud & D. Biber (eds.), *Variation in English: multidimensional studies*, pp. 18-19. Harlow: Pearson Education.
- Divjak, D. & Gries, S. Th. (2006). Ways of trying in Russian: clustering behavioral profiles. *Corpus Linguistics and Linguistic Theory*, 2(1), 23-60.
- Romesburg, C. (2004[1984]). *Cluster analysis for researchers*. Raleigh, NC: Lulu Press.

- Sardinha, T. B. & Pinto, M. V. (eds.) (2014). *Multi-dimensional analysis, 25 years on: a tribute to Douglas Biber*. Amsterdam: John Benjamins.
- Sheskin, D. J. (2007). *Handbook of parametric and nonparametric statistical procedures*. Boca Raton, FL: Chapman & Hall/CRC.
- Xiao, R. (2009). Multidimensional analysis and the study of world Englishes. *World Englishes*, 25(4), 421—50.

أدوات لانكستر الإحصائية

- ١- يمكن استنساخ جميع التحليلات الواردة في هذا الفصل باستخدام أدوات لانكستر الإحصائية. تشمل الأدوات المتاحة على وجه الخصوص:
 - حاسبة الارتباط.
 - التجميعات.
 - التحليل متعدد الأبعاد.
- ٢- يوفر الموقع أيضًا موادًا إضافية للطلاب والمعلمين.

اللغويات الاجتماعية وعلم

الأسلوب: التنوع الاجتماعي والفردية

(١, ٦) ما الذي يتناوله هذا الفصل؟

يناقش هذا الفصل الإجراءات الإحصائية المختلفة والمتاحة لتحليل التباين اللغوي الاجتماعي والأسلوبي في المدونات، ويستعرض الطرق المختلفة لذلك التنوع. وتشمل الإحصاءات التي نوقشت في هذا الفصل اختبار-ت، وتحليل التباين، واختبار مان-ويتني، واختبار كروسكال-والاس، وتحليل التناظر، ونمذجة التأثير الممتزج. يركز كل من هذه الإحصاءات في نوع مختلف من التحليل اللغوي الاجتماعي ويتطلب تصميمًا بحثيًا محددًا. بالنسبة للمقارنات بين المجموعات يُستخدم اختبار-ت، وتحليل التباين، واختبار مان-ويتني، واختبار كروسكال-والاس. يمكن استكشاف الأسلوب اللغوي الفردي باستخدام تحليل التناظر. يمكن الحصول على التحليل اللغوي الاجتماعي (اللابوفيان: نسبة إلى اللغوي لافوف) التقليدي مع التركيز في التنوع في سياق لغوي محدد باستخدام نمذجة التأثير الممتزج، وهي تقنية لا تزال جديدة إلى حد ما في اللغويات الاجتماعية القائمة على المدونة.

سنُجيب في هذا الفصل عن الأسئلة الخمسة الآتية:

- كيف يمكننا التعامل مع التنوعات (اللغوية الاجتماعية) وتوظيفها؟ (المبحث ٢, ٦).
- كيف يمكننا مقارنة مجموعات المتحدثين بناءً على متغيرات اجتماعية مختلفة مثل الجنس أو الطبقة الاجتماعية؟ (المبحث ٣, ٦).

- كيف يمكننا تلخيص المتغيرات اللغوية المتعددة وإنشاء ملفات تعريف للمتحدثين؟ (المبحث ٤, ٦).
- كيف يمكننا استكشاف المتغيرات اللغوية الاجتماعية اللابوفية؟ (المبحث ٥, ٦).
- كيف يمكن استخدام التقنيات التي نوقشت في هذا الفصل في البحث؟ (المبحث ٦, ٦).

(٦, ٢) الأسلوب الفردي والتنوع الاجتماعي: أين تبدأ المتغيرات اللغوية الاجتماعية؟

فكّر... قبل قراءة هذا المبحث، انظر في أربع عينات لأربعة متحدثين، اثنان حقيقيان واثنان خياليان. بناءً على ما قرأته، هل يمكنك تخمين نوع الجنس والعمر التقريبي والخصائص الاجتماعية الأخرى؟ سنتناول هذه الأمثلة قريباً.	
SPEAKER 1 He's a <u>fucker</u>, I can't stick him he's the most snobbish little <u>cunt</u> I've ever known I'd like to see what stuff he had if he had to pay for it himself hard <u>luck</u>, no I'd like to see what he'd do with his stuff if he had to pay with it, pay for it himself. Cos he's well out of order, when <u>you</u>, were you there when he was were <u>you</u> there when he was like slamming his nine iron into one of those brick posts, that's sad he's a twat	المتحدث الأول إنه سخيف، لا أطيعه، إنه أكبر مغرور عرفته على الإطلاق، ليت شعري ما الذي يستحق أن يجعل نفسه من أجله عرضةً للتهلكة. إنه الحظ، ولكنني أود أن أرى ماذا سيفعل بها دفع نفسه ثمناً له. لأنه كان مخطئاً، عندما كنت، كنت هناك عندما كان هناك كنتَ هناك عندما كان يجب ضرب المضرب في صندوق بريد المبنى، هذا أمر مؤسف إنه شديد الحمق
SPEAKER 2 What boy? What in my school? Nick? went out with him. I dumped him. Yeah. I dumped him on Wednesday. I <u>said</u> hey, Mac! Not this Wednesday, last Wednesday I was saying, I just said I don't think we should go out <u>no more</u>. <u>He said</u> alright.	المتحدث الثاني أي صبي؟ ماذا، الذي كان في مدرستي؟ من اسمه نك؟ خرجتُ معه. أنهيت علاقتي به. بلى. أنهيت علاقتي به يوم الأربعاء.

I said bye. I said I hope we can still be friends. He said yeah. I said alright I'll meet you for school, bye. <laugh>I went erm T hope we can still be friends.	قلتُ مهلاً، مالك! ليس هذا الأربعاء، إنه الأربعاء الماضي، قلت فقط لا أعتقد أننا يجب أن نخرج ليس أكثر. قال حسناً. قلت وداعاً. قلت آمل ألا نفقد علاقتنا كأصدقاء. قال نعم. قلت حسناً سألتقي بك في المدرسة، وداعاً. <صوت ضحك> ذهبْتُ آآآ آمل أن نظل أصدقاء.
SPEAKER 3 Come hither Nurse. What is yond gentleman? ... Go ask his name. If he be married, <u>My grave is like to be my wedding bed.</u> My only love sprung from my only hate.	المتحدث الثالث تعالى هنا أيتها الممرضة. ماذا هنالك أيها السيد؟ ... اذهبي فاسأليه عن اسمه. إن كان متزوجاً، فإن قبري سيكون فراش زفافي. انبثق جبي الوحيد من كراهيتي الوحيدة.
SPEAKER 4 And fire-ey'd fury be my conduct now! Now, Tybalt, take the villain back again That late <u>thou</u> gav'st me, for Mercutio's soul Is but a little way above our heads, staying for thine to keep him company. Either <u>thou</u>, or I, or both, must go with him. This shall determine that. O, I am fortune's fool.	المتحدث الرابع ولأتبع صوت الغضب القادم من أعماق النار بعين ملتتهية اسمع يا تيبالت! إني لأرد لك الكلمة - إذ أنت الوغد وليس أنا! ما زالت روح صدقي ميركتيو فوق رؤوس القوم ترفرف تبغي أن تصحب روحك في رحلتها لا بد أن تذهب أو أذهب أو يذهب كل منا معه! يا أيها الغلام يا مسكين! قد كنت صاحبه هنا ولسوف تدركه هناك! هذا سيحكم بيننا! يا لحظي السيء

فكرة الأسلوب محورية بالنسبة للتحليلات في هذا الفصل، فوفقاً لتعريف كوبلاند Coupland (٢٠٠٧: ٢) العام فالأسلوب طرائق للتحدث تكون مرتبطة سياقياً^(١) بالمجموعات الاجتماعية والأوقات والأماكن. سننظر في أثر خلفية المتحدث والمجتمع اللغوي في اللغة التي ينتجها المتحدثون. الأسلوب هو فكرة موحدة تربط ما بين اللغويات الاجتماعية (الأسلوب الاجتماعي) والأسلوبيات (الأسلوب الأدبي) واللغويات القضائية (الأسلوب الفردي). بغض النظر عن طبيعة اللغة الطبيعية في مجموعة البيانات وإن كانت روايات فإن الإجراءات الإحصائية التي نوقشت في هذا الفصل ستساعدنا على تحديد التنوع في أسلوب الكتابة وإدراكه. تُظهر المتغيرات اللغوية التي ينطوي عليها هذا التنوع اختلافات منهجية وفقاً لكل من المتحدثين الفرديين اللذين يميزان بين الأساليب الفردية individual styles والمجموعات التي تميز بين اللهجات الاجتماعية sociolects. ولكن كيف يمكننا تحديد هذه المتغيرات؟

في اللغويات الاجتماعية التنوعية التي ابتكرها لابوف Labov (١٩٦٦، ١٩٧٢) يُعرّف المتغير اللغوي الاجتماعي بأنه "طرق مختلفة لقول المعنى نفسه" (Labov 2010: 368). على سبيل المثال، عند النظر إلى المتحدث الأول في مهمة (فكر) بافترض أنه يمكننا الوصول إلى نسخة صوتية أو أننا يمكن أن نسمع التسجيل الأصلي يمكننا أن نخمن تخميناً مدروساً إلى أي جزء من بريطانيا ينتمي المتحدث بالنظر في نطق (u) في كلمات مثل cunt و fucker و luck التي يمكن نطقها عموماً /ʊ/ في بعض اللهجات في شمال إنجلترا أو /ʌ/. لاحظ أننا نقصد الكلمات نفسها بالمعنى نفسه، والاختلاف إنما هو (صوتي) فقط. من الأمثلة أيضاً على هذا النوع من المتغيرات استخدام نفي النفي كما في قول المتحدث الثاني: I don't think we should go out no more الذي كان يمكن فيه استخدام نفي واحد I don't think we should go out any more دون أن يتغير المعنى. نظراً لأن المعنى الافتراضي propositional meaning لهذين الخيارين واحد يمكننا التركيز في المعنى الاجتماعي، أي الآثار الاجتماعية المترتبة على استخدام صيغة دون أخرى. عندما نوسع التنوع ليشمل البيانات التاريخية (انظر الفصل السابع عن تحليل التغير اللغوي) فكل من you الذي استخدمه المتحدث الأول و thou الذي استخدمه المتحدث الرابع، واللذين يشاران إلى المفرد المخاطب، ينطبق عليهما تعريف لابوف للمتغير اللغوي الاجتماعي. لتعريف المتغير اللغوي الاجتماعي اللابوفي

(١) أي تشير إليه.

Labovian sociolinguistic variable ينبغي رسم حدود المتغير *circumscribe the variable context* بعناية. بعبارة أخرى: تحديد ظرف التنوع *envelope of variation* (المبحث ٢، ٤). هذا يعني أننا بحاجة إلى البحث في النص عن المواضع التي يمكن فيها التباين بين اثنين أو أكثر من الأشكال المختلفة للصيغ اللغوية مع بقاء المعنى كما هو (Tagliamonte 2006). المتغير اللغوي الاجتماعي من هذا النوع هو فئة مغلقة لها عدة مستويات، وعادة لها مستويان (شكلان مختلفان). لمزيد من المناقشة انظر: المبحث (٥، ٦).

على الرغم من أن هذا التعريف الشكلي للمتغير اللغوي الاجتماعي يمنحنا سيطرة كاملة على التحليلات اللغوية والاجتماعية التي يحويها التنوع فإنه سرعان ما ينتهي أثره عندما ننظر إلى أنواع أخرى من التنوع المنهجي التي تتجاوز علم الأصوات والنحو البسيط (Lavandera 1978). على سبيل المثال، يستخدم المتحدث الأول عددًا من الكلمات النابية القوية التي تُعدّ خصائص لغوية موصومة بالعار اجتماعيًا والتي تحمل بوضوح معنى اجتماعيًا واضحًا، إلا أنها لا تفي بتعريف لابوف للمتغير اللغوي الاجتماعي؛ وليس لها بديل معروف من الكلمات التي تُستبدل بـ *fucker* و *cunt* في خطاب مهذب. ويُستخدم العديد من الكلمات النابية كمقويات للمعنى يمكن أن تظهر في أي موضع في الخطاب (انظر المناقشة عن المتغيرات المحيطة في المبحث ٢، ٤).

على النقيض من منهج لابوف الشكلي يبحث بايبر (مثل Biber & Conrad 2009) في التنوع الوظيفي، إذ يكون التركيز التحليلي في اختيارات المتحدثين والكتاب "من المجموع الكامل للخصائص المعجمية النحوية للغة واختيار الخصائص اللغوية الأكثر ملاءمة وظيفيًا لمواقفهم وأغراضهم التواصلية" (Biber & Conrad 2009: 265). يمكننا ملاحظة عدد من هياكل إعداد التقارير (*I said, he said, I went*) في عينة للمتحدث الثاني، تحدّد بوضوح أسلوب المتحدث على الرغم من عدم وجود منافسة واضحة للحفاظ على المعنى بين الأشكال اللغوية المختلفة التي يمكن تحليلها بسهولة على طريقة لابوف. على النوال نفسه فاللغة المجازية للمتحدث الثالث (*my grave is like to be my wedding bed*) لها بلا شك أثر خاص من الأسلوب، ولكن لا يمكن تحليلها إلا وظيفيًا. من المهم أن ندرك أن النهج الوظيفي، بالرغم من تحقيق مجموعة متنوعة من الخصائص التي تعد أكبر مقارنة بنهج لابوف الشكلي، ليس دائمًا واضحًا، ويتطلب التفكير الناقد في نوع الأدلة ومصادرها في البيانات (المدونات). يعتمد هذا النهج على القابلية المباشرة لمقارنة المدونات (الفرعية) التي نستخدمها، ويفترض أن المدونات التي نقارنها متشابهة في جميع الجوانب المهمة ما عدا المتغيرات التي نتحكم بها ونركز فيها مثل الجنس والعمر والسجل اللغوي.

يعتمد النهج الوظيفي على مستوى تحديد المتغير واختياره عند جمع العينات. على سبيل المثال، إذا كانت المدونات التي نقارنها متحيزة بسبب الإفراط أو التفريط في تمثيل موضوع معين فسيكون التحليل ضعيفاً. لتوفير إجابة كاملة عن السؤال في مهمة (فكر) فالمتحدثان الأول والثاني شخصان حقيقيان من المدونة الوطنية البريطانية. المتحدث الأول طالب يبلغ من العمر ١٧ عاماً من منطقة وسط إنجلترا من الطبقة العاملة. المتحدث الثاني طالبة تبلغ من العمر ١٤ عاماً من لندن، ومن إحدى الطبقات الاجتماعية الدنيا. أما المتحدث الثالث (جوليت) والمتحدث الرابع (روميو) فشخصيتان روائيتان من مسرحية شكسبير الشهيرة، وهما يمثّلان في العمر والجنس المتحدثين الأول والثاني لكنهما يختلفان في التنوع التاريخي للغة الإنجليزية التي يتحدثان بها، والطبقة الاجتماعية، إذ ينتمي روميو وجوليت إلى أسر ثرية في فيرونا الروائية لشكسبير، والأهم من ذلك أن روميو وجوليت هما من نتاج خيال شكسبير، وليس شخصيتين حقيقيتين. من المثير للاهتمام أن المتحدثين يتعاملون مع موضوعات متوازية إلى حد ما: الغضب والعدوان عند المتحدث الأول والرابع، والمواعدة والمغازلة عند المتحدث الثاني والثالث.

(٦، ٣) مقارنة المجموعة: اختبار-ت، وتحليل التباين، واختبار مان-ويتني U، واختبار كروسكال-والاس

فكر...

قبل قراءة هذا المبحث فكر في السؤال التالي:

هل يمكنك تحديد جنس المتحدث الأول من المحادثة القصيرة التالية؟

SPEAKER 1: So, she says, oh it don't matter. She says er <pause> will you help me fill them in now? I says, can I ask you a question? She says what? I says, these forms things, I says <pause> d'ya know how to fill them in? She says

SPEAKER 1: Ah.

SPEAKER 1: No. She said, somebody's always done it for me.

SPEAKER 2: Annette's done them.

المتحدث ١: إذن، كما قالت، لا يهم. إنها تقول آآآ <توقف> هل ستساعدني على ملئها الآن؟

أقول، هل يمكنني أن أطرح عليك سؤالاً؟ فتقول ماذا؟ فأقول، هذه تكون الأشياء، أقول >

توقف < هل تعرف كيف تملأها؟ فتقول

المتحدث ١: آها.

المتحدث ١: لا، قالت لقد كان يفعل ذلك شخص ما دائماً من أجلي.

المتحدث ٢: لقد أنجزتها أنيت من أجلي.

المحادثة أعلاه مأخوذة من مدونة BNC64 (Brezina 2013) وهي عينة من خطاب غير رسمي مكون من 1.5 مليون كلمة ومأخوذة من المدونة الوطنية البريطانية. هذه العينة لافئة للنظر لعدد من الأسباب ولا سيما الاستخدام المتكرر للضمائر من المتحدث الأول. وتكرار الضمائر الشخصية بوصفها متغيراً لغوياً اجتماعياً يستحق بعض الاهتمام، فهل يمكن -مثلاً- أن يساعد على التمييز بين جنس المتحدث؟ لنفترض أن استخدام الضمير الشخصي هو علامة لغوية اجتماعية ولنختبر هذه الفرضية إحصائياً.

تتيح لنا مدونة BNC64 التي تتكون من عينات منطوقة من ٣٢ متحدثاً ومتحدثة تتبع الأداء اللغوي للمتحدثين الفرديين^(٢) بسهولة، وتوفّر كمية كبيرة من الأدلة لكل متحدث من جهة الكلمات التي يمكن تحليلها. قبل أن نبدأ التحليل من المهم أن نلاحظ في هذا البحث أننا نعتمد على النهج الوظيفي للتنوع الذي نوقش في البحث (٦، ٢). لا يمكن تفسير الضمائر الشخصية بوصفها متغيرات لغوية اجتماعية لابوية لأننا لا نتعامل مع متغيرين مختلفين تماماً بين شكلين (أو أكثر). نحن إذن مهتمون بمقارنة الأسلوب الذي يستخدم العديد من الضمائر الشخصية -كما هو موضح في مهمة (فكر)- بأساليب مختلفة أو بعدد أقل من الضمائر. بدلاً من افتراض معنى ثابت في المتغيرات اللغوية (وهو ما لا نستطيعه)، نفترض أننا نتعامل مع مدونة متوازنة تختبرنا عن وظيفة التفاعل بالضمائر في الخطاب غير الرسمي، وأن للمتحدثين من جهة المبدأ فرصاً متساوية لاستخدام الضمائر الشخصية في خطابهم.

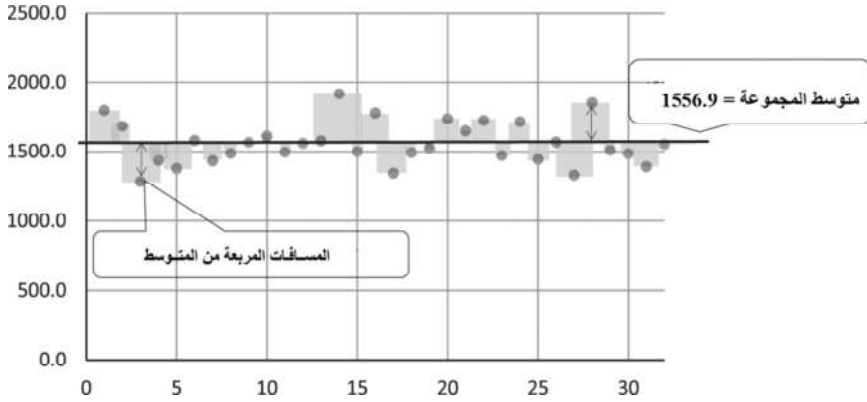
إذن، أي اختبار إحصائي ينبغي أن نستخدمه في هذه الحالة؟ أيسر خيار لدينا هو اختبار-ت. في هذه الحالة سنستخدم نسخة من اختبار-ت لعينات ويلتش المستقلة Welch's independent samples t-test. يقارن هذا الاختبار مجموعتين من المتحدثين (مثل المتحدثين من الذكور والإناث)، ويقارن القيم المتوسطة للمتغير اللغوي، والمتغير في مثالنا هو التكرار النسبي للضمائر الشخصية في عينات المتحدث الفردية. يأخذ هذا الاختبار في الحسبان التنوع الداخلي في كل مجموعة متباينة.

(٢) نستخدم مجموعة الأساليب الإحصائية التي نوقشت في هذا البحث تصميم بحث للنص المكتوب أو الحوارية، وذلك بتتبع ظهور المتغيرات اللغوية فيها. في اللغة المكتوبة أصبح ذلك أسهل نظراً لأن المكتوب عادة ما يكون من مؤلف أو مؤلفين منفصلين، على عكس الكلام الحوارية الذي يتطلب الفصل بعناية بين أدوار المتحاورين.

التباين variance هو مجموع المسافات المربعة للقيم الفردية من متوسط المجموعة مقسوماً على درجة الحرية. هذه هي النسخة المربعة من الانحراف المعياري للعينة (المبحث ٤, ٢) ومن ثمّ الانحراف المعياري^٢.

يُحسب التنوع وفق المعادلة التالية:

$$\text{التباين} = \frac{\text{مجموع المسافات المربعة من المتوسط}}{\text{درجات الحرية}} \quad (٦, ١)$$



شكل (٦, ١). توزيع الضائير الشخصية في مدونة BNC64 للمتحدثات.

يوضح الشكل (٦, ١) كيفية حساب التباين للضائير الشخصية لجميع المتحدثات البالغ عددهن ٣٢ من مدونة BNC64 الفرعية (التكرار النسبي لكل 10 آلاف كلمة). كل دائرة هي متحدث فردي، والخط الأفقي السميكة هو متوسط المجموعة، وتمثل المنطقة المظلمة المسافات المربعة للقيم الفردية من المتوسط.

منطق حساب التباين سهل فهو معرفة مقدار التنوع الكلي. نضيف المسافات المربعة من المتوسط ونستخدم مسافات مربعة بدلاً من مسافات بسيطة لأن المسافات الأقل من المتوسط سالبة وستلغي المسافات الموجبة، ثم نقسم هذا على درجات الحرية. درجات الحرية هي مفهوم معقد يُستخدم عند التعامل مع الحسابات المستندة إلى عينة (مدونة) بدلاً من المجموع اللغوي. تشير درجة الحرية (df) إلى عدد من العناصر المستقلة الحرة في حسابنا، أي العناصر التي لا يمكن التنبؤ بها من

العناصر السابقة. هي في التطبيق العملي عدد الحالات (أي: عدد النصوص أو المتحدثين) أو عدد المجموعات (عند النظر إلى تباين المجموعة) مطروح منها واحد. نطرح واحدًا من عدد الحالات (المجموعات) لأن الحالة الأخيرة يمكن التنبؤ بها دائمًا لأنها الوحيدة بعد استيفاء كل الحالات^(٣). على سبيل المثال، عندما نتعامل مع ٣٢ متحدثًا فستكون درجات الحرية (٣١)، وذلك نتيجة (٣٢-١). عند تطبيق ذلك على مثالنا نحصل على:

$$\text{التباين للإناث} = \frac{(1794.6-1556.9)2+(1681.3-1556.9)2+..(1550.0-1556.9)2}{32-1} = 23820.5 \quad (٦, ٢)$$

لنعد إلى اختبار-ت الذي يستخدم التباين كأحد مكوناته. مثل أي اختبار إحصائي فإن اختبار-ت يحتوي على افتراضات ينبغي مراجعتها قبل إجراء الاختبار. الافتراض الأهم هو استقلال الملاحظات independence of observations، ويشير هذا الاستقلال في السياق اللغوي الاجتماعي إلى كل ملاحظة (نص أو عينة خطاب) تنتمي إلى متحدث واحد أختير عشوائيًا^(٤)، كما يشير إلى كل استخدام للغة من متحدث في العينة لا يتأثر باستخدام اللغة بمتحدث آخر. نحتاج هنا إلى التأكد من تضمين كل متحدث بوصفه ملاحظة (حالة) مرة واحدة فقط حتى لو توافرت نصوص متعددة، وإلا فإننا نحسب المتحدث نفسه مرتين أو عدة مرات، وننقص الجزء الأول من الافتراض. يسهل الحفاظ على الجزء الثاني من الافتراض في النصوص المكتوبة أو المونولوجات، ولكن في المحادثة سينتقص هذا الافتراض إلى حد ما إذا تفاعل المتحدثون فيما بينهم (يعرف بتأثير

(٣) من أجل الاستخدام الناجح للتقنيات الإحصائية التي نوقشت هنا فإن فهم التفاصيل الرياضية لمفهوم (درجات الحرية) ليس مهما. المهم هو معرفة أن درجات الحرية تظهر في معادلات هذه القياسات الإحصائية، ويبلغ عنها تبليغا روتينيًا في تقارير الأبحاث (انظر التقرير الإحصائي في نهاية هذا الفصل). وعلى سبيل المجاز، تصوّر فصلًا يحتوي على ١٥ مقعدًا ويبحث فيه ١٥ عشر طالبًا عن مكان للجلوس. درجة الحرية هي ١٤ (١٥ مطروح منها ١) لأن ١٤ من الطلاب لديهم (حرية) في اختيار مكان ما، والأخير ليس له حرية الاختيار وعليه الجلوس على آخر مقعد فارغ.

(٤) إذا كانت لدينا عينة أو أكثر من كل متحدث وأردنا معرفة الاختلاف في لغتهم بين العينة ١ والعينة ٢ وما إلى ذلك (مثل التغير أو التطور اللغوي) فإننا نتعامل مع ما يسمى بتصميم القياسات المتكررة الذي يتطلب نسخة مختلفة من الاختبار الإحصائي (سيُشرح ذلك لاحقًا).

التهيئة في اللغويات النفسية وبالمواءمة في اللغويات الاجتماعية). ينبغي أن نضع في حسابنا هذا التأثير وتقييم تأثيره في كل حالة على حدة. هناك افتراضان آخران تتضمنهما الكتب المتخصصة (وعادةً ما يساء فهمهما)، وهما التوزيع الطبيعي normal distribution والتفاوت المتماثل homoscedasticity. التوزيع الطبيعي للمتغير اللغوي في المجموع اللغوي هو توزيع متناظر مفترض يُتصوّر على أنه منحنى على شكل جرس (المبحث ٣، ١). التفاوت المتماثل^(٥) هو مصطلح فني يشير إلى المساواة في التباينات بين مجموعتين نريد المقارنة بينهما. كما هو موضح في الدراسات (Boneau 1960; Lumley et al. 2002) فإن اختبار-ت فعال، ويمكن استخدامه حتى مع بيانات منحرفة أو غير موزعة توزيعاً طبيعياً. علاوة على ذلك، فنسخة اختبار-ت المعروفة باسم اختبار عينات ويلتش المستقلة تعوض أيضاً عن التباينات غير المتكافئة أو الكميات المختلفة من التنوع في كل مجموعة من المجموعات المقارنة وذلك بإعادة تعديل درجات الحرية، وعليه فإن انتقاص هذين الافتراضين نادراً ما يمثل مشكلة.

يُحسب اختبار-ت لعينات ويلتش المستقلة وفقاً للمعادلة التالية:

$$(٦, ٣) \quad \text{اختبار-ت لعينات ويلتش المستقلة} = \frac{\text{متوسط المجموعة ١} - \text{متوسط المجموعة 2}}{\sqrt{\frac{\text{تباين المجموعة ١}}{\text{عدد الحالات في المجموعة ١}} + \frac{\text{تباين المجموعة ٢}}{\text{عدد الحالات في المجموعة ٢}}}}$$

كما يتبين من المعادلة هناك ثلاثة عوامل لها تأثير في كون الاختبار ذا دلالة أم لا. العامل الأول: حجم متوسط الفرق. العامل الثاني: التباين في كل من المجموعتين. العامل الثالث: حجم العينة (عدد الحالات، أي عدد المتحدثين أو النصوص، في كلتا المجموعتين). تكون قيمة الاختبار-ت كبيرة والاختبار دالاً إذا كان هناك فرق كبير بين المتوسطات، وكان التباين صغيراً في المجموعات، وكان هناك عدد كبير من الحالات. تشير هذه العوامل مجتمعة إلى أدلة كافية في البيانات تدل على أن المجموعتين مختلفتان فيما يتعلق باستخدام المتغير اللغوي المعني.

(٥) التفاوت المتماثل هو جزء من التنافر، ويمكنك استخدام المصطلح البديل الأكثر وصفيّة: (المساواة في التباينات).

في مثالنا فإن التكرارات النسبية (لكل 10 آلاف كلمة) للضمائر الشخصية بالمدونتين الفرعيتين للذكور والإناث هي 1451.8 و 1556.9 على التوالي، والتباينات هي 22,256.5 و 23,820.5 ولدينا ٣٢ متحدثاً في كل مجموعة. عندما ندخل هذه المعلومات في المعادلة نحصل على النتيجة.

$$(٦, ٤) \quad 2.8 = \frac{105.1}{\sqrt{744.4 + 695.5}} = \frac{1451.8 - 1556.9}{\sqrt{\frac{22256.5}{32} + \frac{23820.5}{32}}} = \text{اختبار-ت لعينات ويلتش المستقلة}$$

في هذه الحالة تكون قيمة اختبار-ت (2.8)، وهي دالة إحصائية، وبدرجات الحرية المناسبة (61.93)، وقد حُسبت درجة الحرية بمعادلة لن نعرضها هنا^(٦). حُسبت قيمة الاحتمال وظهرت (0.007)، وهي أصغر من 0.05 (مستوى ألفا 5% المعتاد^(٧)). هذه الأدلة كافية من المدونة لرفض الفرضية الصفريّة (لا يوجد فرق بين المجموعتين من الجنسين). نستنتج بعد ذلك أن هناك فرقاً ذا دلالة إحصائية بين المتحدثين من الذكور والإناث.

إضافةً إلى الاختبار الإحصائي نحتاج أيضاً إلى حساب مقياس حجم التأثير لتقييم حجم الفرق بين المجموعتين بأسلوب معياري. مع اختبار-ت لدينا العديد من خيارات مقاييس حجم التأثير التي تشمل كوهن-د Cohen's d و (ر) r بوصفهما مقياسين يُستخدمان لقياس حجم التأثير. يُحسب كوهن-د على أنه الفرق بين المتوسطين المُعبر عنهما بوحدات الانحراف المعياري. يوصي كوهن (1988: 40) بالتفسير القياسي التالي لمقياس d (المبحث ٤، ٨):

تفسير d : أكبر من 0.3 يكون التأثير صغيراً. أكبر من 0.5 يكون التأثير متوسطاً. أكبر من 0.8 يكون التأثير كبيراً

بغض النظر عن علامة الجمع أو علامة الطرح أمام القيمة، فهما يشيران إلى ما إذا كانت للمجموعة الأولى التي جرى بحثها في المقارنة قيمة متوسط أكبر (+) أو قيمة متوسط أصغر

(٦) هناك معادلة (معقدة بعض الشيء) يمكن استخدامها لحساب درجات الحرية، ومن ثم تحديد نقطة الفصل لإحصاء اختبار ت لتحديد الدلالة. تُدرج عادةً نقطة الفصل هذه في جداول إحصائية وفق درجات الحرية المناسبة. في الحزم الإحصائية الحديثة، يمكن تخطي هذه الخطوة نظراً لاحتساب درجات الحرية وكذلك قيمة الاحتمال (الدلالة الإحصائية) تلقائياً.

(٧) (مستوى ألفا) هو المصطلح الإحصائي لمستوى الدلالة الإحصائية التي نختارها. في العلوم الاجتماعية يكون مستوى الدلالة عادة 0.05 أي: ٥٪.

(-) من المجموعة الثانية، لذلك فهذا مؤشر على اتجاه التأثير direction of the effect. يمكن أن يختلف تفسير قيم كوهن-د وفقاً لنظام معين. قد لا يكون ما يُعدّ تأثيراً كبيراً في الفيزياء تأثيراً كبيراً في الدراسات اللغوية أو العكس (انظر: المبحث ٢, ٩ لمزيد من المناقشة). يُحسب كوهن-د على النحو الآتي:

$$\text{كوهن-د} = \frac{\text{متوسط ١ المجموعة} - \text{متوسط المجموعة 2}}{\text{مجموع الانحراف المعياري}} \quad (٦, ٥)$$

حيث

$$\text{مجموع الانحراف المعياري} = \frac{\text{الانحراف المعياري الأول} \times 2 \times (\text{الحالات في المجموعة الأولى} - 1) + \text{الانحراف المعياري الثاني} \times 2 \times (\text{الحالات في المجموعة الثانية} - 1)}{\sqrt{\text{جميع الحالات} - 2}}$$

في مثالنا فإن حساب كوهن-د على النحو الآتي:

$$\text{كوهن-د} = \frac{1451.8 - 1556.9}{151.8} = \frac{1451.8 - 1556.9}{\sqrt{\frac{22256.5 \times (32.1) + 23820.5 \times (32 - 1)}{64 - 2}}} = 0.69 \quad (٦, ٦)$$

يمكن تفسير ذلك على أنه تأثير متوسط.

وكما هو الحال مع أي مقياس لحجم التأثير نحتاج أيضاً إلى النظر في مجال الثقة ٩٥٪ لكوهن-د الذي يراوح في مثالنا هنا بين (0.18) و(1.21). يدل هذا على مدى التأثير المحتمل في المجموع اللغوي الذي هو جميع المتحدثين من الذكور والإناث في اللغة الإنجليزية البريطانية. نظراً لأن مجال الثقة واسع للغاية إذ يراوح من الحد الأدنى من التأثير إلى تأثير كبير، لا يمكننا التأكد من الحجم الفعلي للتأثير في المجموع اللغوي، ويرجع ذلك إلى صغر حجم العينة نسبياً الذي يُمثّل ٦٤ متحدثاً. بغض النظر عن الحجم الصغير، فإنه بالنسبة للبيانات الشديدة الانحدار (أي: البيانات التي تنقص بشدة افتراض الحالة الطبيعية) اقترح أسلوب فعال من كوهن-د (Erceg- Hum & Miroseovich 2008) يعمل لاستخدام المعادلة نفسها التي تستخدمها النسخة القياسية في المعادلة

(٥, ٦)، ولكن بتعديل المتوسطات والانحرافات المعيارية التي تدخلها حيث يتم تشذيب المتوسطات أو تسوية الانحرافات المعيارية^(٨).

هناك خيار شائع آخر لحجم التأثير مع اختبار-ت وهو r المقياس المعروف لتحليل الارتباط (المبحث ٢, ٥). يمكن تحويل كوهن-د بسهولة إلى r والعكس (المبحث ٤, ٨). يمتاز r بنطاق قياسي للقيم (من 0 إلى 1)، أما كوهن-د فلا يتميز بحد أدنى أو حد أقصى للقيمة. يوصي كوهن (1988: 79-80) بمقياس r على نحو التفسير الآتي:

تفسير r : أكبر من 0.1 التأثير صغير. أكبر من 0.3 التأثير متوسط. أكبر من 0.5 التأثير كبير في حالتنا فإن $r = 0.33$ ومجال الثقة ٩٥٪: [0.08, 0.53] ويمكن تفسيرها على أنها ذات تأثير متوسط مع الانتباه إلى أن تأثير المجموع اللغوي الحقيقي قد يراوح من الحد الأدنى (0.08) إلى الحد الكبير (0.53).

قارنًا حتى الآن مجموعتين من المتحدثين. لكن ماذا لو كنا بحاجة إلى مقارنة المزيد؟ في مثل هذه الحالة يمكن استخدام اختبار يسمى تحليل التباين الأحادي الاتجاه^(٩) (one-way ANOVA). تخيل أنك مهتم باستخدام الصيغة غير القياسية ain't في الخطاب البريطاني وما إذا كان للحالة الاجتماعية الاقتصادية للمتحدثين (الطبقة الاجتماعية) تأثير بها أم لا. سنستخدم بيانات من مدونة BNC64 مرة أخرى. فيما يأتي بعض الأمثلة على استخدامها:

- (1) cos there ain't [= isn't] no sign of them, is there? (BNC64, M2)
- (2) I've won twice ain't [= haven't] I? (BNC64, F5)

تُصنف مدونة BNC64 المتحدثين إلى أربع فئات تعكس الحالة الاجتماعية والاقتصادية للمتحدثين. الفئة الأولى: (AB) لإداريين وتنفيذيين وخبراء. الفئة الثانية: (C1) لموظفين صغار

(٨) قُدمت هذه العملية للتخلص من القيم المتطرفة (المبحث ٣, ١)، في حين أن وسائل التشذيب تتضمن حذف جزء معين (على سبيل المثال ٢٠٪ على كل جانب من جوانب التوزيع) من القيم القصوى قبل حساب الوسط بالطريقة القياسية (انظر المعادلة ١, ١) فإن التقليل يحصل باستبدال هذه القيم القصوى بأخر قيمة غير متطرفة قبل حساب الانحراف المعياري بالطريقة القياسية (انظر المعادلة ١١, ٢). لمزيد من المعلومات انظر Erceg-Hum & Miroseovich (2008).

(٩) يشير مصطلح الأحادي في اللغة الإحصائية إلى عدد المتغيرات المستقلة وتكون حالة واحدة. باستخدام تحليل التباين الأحادي نقسم المتحدثين إلى مجموعات متعددة استنادًا إلى معيار واحد مثل الحالة الاجتماعية والاقتصادية.

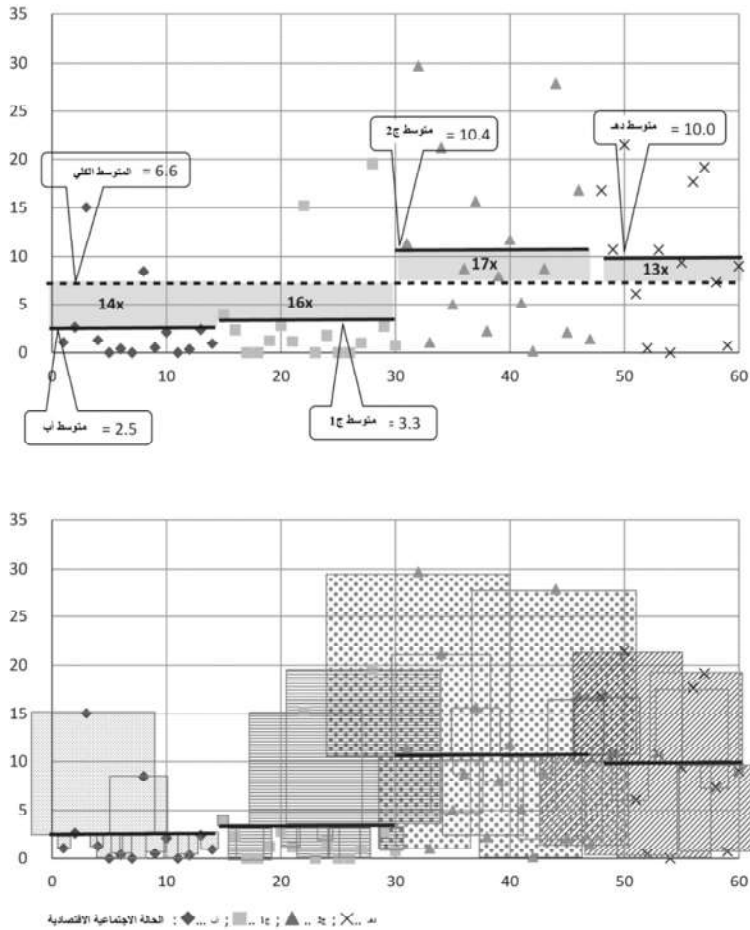
في الإدارة ومشرفين ومحترفين. الفئة الثالثة: (C2) لعمّال حرفيين مهرة. الفئة الرابعة: (CD) لعمّال شبه مهرة أو غير مهرة.

لتحليل تباين الافتراضات التي تشبه افتراضات اختبار-ت: استقلال الملاحظات، والحالة الطبيعية، والتفاوت المتماثل، فإن الأهم منها هو استقلال الملاحظات. تكمن أهميته من أن اختبار تحليل التباين فعّال ضد انتقاض افتراض الحالة الطبيعية (Schmider et al. 2010). إذا انتقضت فرضية التفاوت المتماثل فمن الممكن استخدام أسلوب ويلتش من تحليل التباين الذي يُعدّ فعّالاً ضد انتقاضه (Minitab n.d.). معادلة تحليل التباين الأحادي^(١٠) هي على النحو الآتي:

$$\text{تحليل التباين الأحادي (ف)} = \frac{\text{التباين بين المجموعة}}{\text{التباين داخل المجموعة}} \quad (٦,٧)$$

لتوضيح المنطق وراء تحليل التباين، يُمثّل الشكل (٦,٢) مخططات لستين متحدثاً من مدونة BNC64 وفقاً لاستخدامهم لـ ain't (التكرار النسبي لكل 10 آلاف كلمة) وانتائهم إلى الطبقة الاجتماعية (أشكال نقاط البيانات). استُبعد من التحليل أربعة متحدثين كانت طبقتهم الاجتماعية غير معروفة. تعرض اللوحة العلوية التباين بين المجموعة (المنطقة المظللة مضروبة في الحالات في كل مجموعة، أي مسافة ابتعاد متوسط كل مجموعة عن المتوسط الكلي grand mean). أما اللوحة السفلية فتعرض التباين داخل المجموعة (المنطقة المظللة). إن خط الأفق المنقوط في اللوحة العلوية هو المتوسط الكلي، أي متوسط المدونة بأكملها. تمثل الخطوط الأفقية الأربعة الأقصر متوسطات لمجموعات الطبقة الاجتماعية الفردية. توضح المربعات في اللوحة السفلية المسافات المربعة (المنطقة المظللة) من متوسطات المجموعة (الخطوط الأفقية الأقصر) لكل نقطة بيانات.

(١٠) هذه هي معادلة تحليل التباين الأحادية التقليدية لفيشر التي هي أسهل في التفسير. يتبع تحليل التباين لويلتش المبادئ، نفسها لكن المعادلة تتضمن عدداً من التعديلات (انظر الموقع الإلكتروني المصاحب للكتاب من أجل الاطلاع على معادلة تحليل التباين لويلتش).



شكل (٢, ٦). حساب تحليل التباين: التباين بين المجموعة (العلوي)، التباين داخل المجموعة (السفلي).

لحساب التباين في البيانات الذي توضحه الطبقة الاجتماعية بوصفه متغيرًا مستقلًا بين المجموعة، نحتاج إلى أخذ متوسطات المجموعة وحساب مسافات الترتيبية من المتوسط الكلي، ثم ضرب هذه الأرقام في عدد الحالات الفردية في كل مجموعة ثم قسمة ذلك على درجات الحرية المناسبة.

$$\frac{\text{حالات المجموعة } 1 \times (\text{المتوسط } - 1) + \text{حالات المجموعة } 2 \times (\text{المتوسط } - 2) + \dots + 2}{\text{عدد المجموعات} - 1} = \text{التباين بين المجموعة}$$

التباين داخل المجموعة هو مجموع التباينات الفردية لكل مجموعة، ويُحسب كل منها وفقاً للمعادلة (٦، ١)، ودرجات الحرية هي عدد الحالات ناقص عدد المجموعات.

$$\frac{\text{مجموع المسافات المربعة للمجموعة } 1 + \text{مجموع المسافات المربعة للمجموعة } 2 + \dots}{\text{عدد حالات المجموعات}} = \text{التباين داخل المجموعة}$$

في مثالنا يُحسب التباين بين المجموعة والتباين داخل المجموعة وتحليل التباين (ف) F على النحو الآتي:

$$(٦، ١٠) \quad \frac{14 \times (2.5-6.6)^2 + 16 \times (3.3-6.6)^2 + 17 \times (10.4-6.6)^2 + 13 \times (10.0-6.6)^2}{4-1} = \text{التباين بين المجموعة}$$

$$(٦، ١١) \quad 47.8 = \frac{231.2 + 485.1 + 1341.4 + 621.7}{60-4} = \text{التباين داخل المجموعة}$$

$$(٦، ١٢) \quad 5.6 = \frac{267.35}{45.7} = \text{تحليل التباين الأحادي (ف)}$$

قيمة الاحتمال ذات الصلة هي $p = 0.002$ وبهذا يمكننا أن نستنتج أن النتيجة ذات دلالة إحصائية.

ولأن تحليل التباين هو اختبار شامل يكتشف اختلافاً ذا دلالة إحصائية في أي مكان في البيانات وبين أي من المجموعات فإنه نخبرنا بموضع الاختلاف بالضبط. نحتاج هنا إلى إجراء اختبارات بعدية **post-hoc tests** وهي مقارنات زوجية لمتوسطات المجموعة الفردية مع تصحيح للاختبار المتعدد الذي يشبه اختبار-ت. يزداد احتمال ظهور نتيجة إيجابية خطأً من النوع الأول؛ لأنه مع كل اختبار يُستخدم قيمة الاحتمال تكون على استعداد لقبول أنه في عدد قليل من الحالات (5%) ستكون النتيجة ذات دلالة إحصائية حتى لو كانت الفرضية الصفرية صحيحة (لا يوجد تأثير للمتغير المستقل). بإجراء اختبار متعدد يزيد هذا الاحتمال ازدياداً كبيراً. على سبيل المثال، من خلال أربع مجموعات كما في المثال الخاص بنا، يمكننا إجراء ستة مقارنات زوجية مما سيزيد الخطأ من النوع الأول

إلى 26.5%^(١١). هذا المعدل الخطأ ليس مقبولا ونحتاج إلى تطبيق تصحيحات للاختبارات المتعددة مثل تصحيح بونفيروني (Bonferroni 2000; Cabin & Mitchell 1995; Shaffer 1993; Toothaker 1993). في حال الاختبارات البعدية يُبلغ عادة عن اختبارات-ت مع تصحيح بونفيروني، وهو اختبار يحتاج إلى شيء من التحفظ لأنه قد لا يكون لديه القدرة الإحصائية لاكتشاف الاختلافات الصغيرة. تشمل الخيارات الأخرى اختبار توكي للاختلافات الدالة إحصائيا بدقة (Tuckey's honestly significant difference) أو اختبارات مختلفة تستند إلى التمهيد bootstrapping وهو طريقة لإعادة أخذ عينات متعددة (المبحث ٣، ٧).

م	إناث	ذكور	م
الترتيب النسبي	الترتيب النسبي	التكرار النسبي	الترتيب
1	1794.6	4	57
2	1681.3	10	48
3	1282.5	59	44
4	1435.1	42	50
5	1377.9	53	29
6	1577.8	16	25
7	1432.1	43	49
8	1485.7	34	58
9	1564.5	20	45
10	1609.6	14	24
11	1493.4	32	41
12	1558.1	22	61
13	1575.8	17	38

شكل (٣، ٦). مجموعة بيانات من مدونة BNC64 والتكرارات النسبية والرتب لاستخدام الضمائر الشخصية.

(١١) يُطلق على هذا الخطأ (الخطأ العائلي) لأنه مرتبط باختبارات متعددة استناداً إلى مجموعة من الاستدلالات الإحصائية المتعددة الصلات. يحسب الخطأ العائلي على النحو الآتي: عدد الاختبارات $1 - 0.95$ ، وفي مثالنا هنا فالمعادلة ونتيجتها هما على النحو الآتي: $(1 - 0.95^6 = 0.265)$.

عند الإفصاح عن تحليل التباين والدلالة الإحصائية، ينبغي الإفصاح أيضًا عن حجم التأثير، وحجم التأثير الكلي (الشامل) الذي يُبلغ عنه أحيانًا بإيتا تربيع (η^2). التفسير القياسي لحجم التأثير هذا هو كما يأتي (Kirk 1996؛ Richardson 2011):

تفسير η^2 الجزئي:

أكبر من $0.01 > \eta^2$: التأثير صغير. أكبر من $0.06 > \eta^2$: التأثير متوسط. أكبر من $\eta^2 > 0.14$:

0.14: التأثير كبير

من الخيارات الأفضل والأكثر تحديدًا الإفصاح عن أحجام التأثير للفروق بين المجموعات الفردية إلى جانب الاختبارات البعدية. يُعد كوهن-د أو r أكثر أحجام التأثير شيوعًا في هذه الحالة. لا اختبار-ت وتحليل التباين أيضًا نظيران إحصائيان غير بارامترين non-parametric counterparts وهما: اختبار مان-ويتني U المعروف أيضًا باسم اختبار مجموع رتب ويلكسون Wilcoxon rank-sum test واختبار كروسكال-والاس. يعني مصطلح (غير بارامترى) أننا لسنا بحاجة إلى أي معرفة عن البارامترات كالمتوسط أو الانحراف المعياري للمتغير المعني بالتحليل في المجموع اللغوي (Dodge 2008). لا تحتوي هذه الاختبارات على افتراضات عن التوزيع الطبيعي الأساسي للمتغير اللغوي الاجتماعي في المجموع اللغوي وعن تجانس التباين، ويمكن استخدامها في حالات بيانات المقياس الشديدة الانحراف أو في حالات البيانات الترتيبية. وبدلاً من التكرارات النسبية يعمل هذان الاختباران بصيغة (الرتب)، ويمكن بسهولة تحويل التكرارات النسبية (متغير المقياس) إلى رتب على الرغم من أن هذا يأتي على حساب فقدان بعض المعلومات عن الاختلافات الدقيقة بين الحالات للمتحدثين. يوضح الشكل (٦، ٣) تحول مجموعة البيانات من التكرار النسبي إلى الرتب؛ وتُظهر مجموعة البيانات (المعروضة جزئياً فقط) توزيع الضمائر الشخصية من مجموعتين من الجنسين في مدونة BNC64 التي نوقشت سابقاً في سياق الحديث عن اختبار-ت.

لحساب اختبار مان-ويتني U وكروسكال-والاس ينبغي اتباع ما يأتي:

١- إذا كان المتغير اللغوي متغير مقياس فجميع البيانات تُرتب تسلسلياً دون النظر إلى القيمة الأعلى التي تحوز على الرتبة (1)، وتحوز القيم الدقيقة ترتيباً متوسطاً. على سبيل المثال: إذا كانت القيم العليا الثلاثة هي (10.5) فستكون الرتبة لها 2، 2، 2 وفق نتيجة الحساب الآتي: $2 = (3/3 + 2 + 1)$.

٢- تُقسم البيانات إلى مجموعات وفقاً للمتغير المستقل كالجنس أو الوضع الاجتماعى والاقتصادى، ويُحسب مجموع الرتب لكل مجموعة.

بالنسبة لاختبار مان-ويتني U (مجموع رتب ويلكسون) تُحسب قيمتا $U^{(1)}$ ، واحدة لكل مجموعة، ثم تُؤخذ منها القيمة الأصغر (Mann & Whitney 1947; Kerby 2014):

$$U_1 = \text{مجموع الرتب للمجموعة 1} - \frac{(\text{الحالات في المجموعة 1}) \times (\text{الحالات في المجموعة 1} + 1)}{2} \quad (١٣, ٦)$$

$$U_2 = \text{مجموع الرتب للمجموعة 2} - \frac{(\text{الحالات في المجموعة 2}) \times (\text{الحالات في المجموعة 2} + 1)}{2}$$

الفكرة وراء هذا الاختبار هي الأخذ في الحسبان المجموع الفعلي لرتب كل مجموعة وطرح الحد الأدنى الممكن لمجموع الرتب منها، التي يمكن للمجموعة تحقيقها [الحالات في المجموعة، X (الحالات في المجموعة، $1 + 2 / 2$]. لنفترض أن لدينا مجموعتين من خمسة متحدثين، في المجموعة الثانية جميع الرتب الأصغر (1-5). مجموع الرتب لهذه المجموعة هو $1 + 2 + 3 + 4 + 5 = 15$ ، والحد الأدنى للمجموع المحتمل لهذه المجموعة هو أيضاً (15) نتيجةً لـ: $[2 \times (6 \times 5)]$. الناتج لـ U أصغر الخيارين من المعادلة (١٣، ٦) وهي (0) نتيجةً لـ: $(15 - 15)$. يُظهر منطق الاختبار بأن قيم U الصغيرة تُظهر فروقاً قوية بين المجموعتين كما هو موضح أعلاه.

يعمل اختبار كروسكال-والاس (الذي يُنتج قيم H) على مبدأ شبيه ولكن يأخذ في الحسبان الرتب في مجموعات متعددة تزيد على ثلاث مجموعات (Kruskal & Wallis 1952). تُوفّر المعلومات عن الدلالة الإحصائية لـ U و H تلقائياً بحزم إحصائية، ويمكن الحصول عليها أيضاً من الجداول الإحصائية. اقترحت مقاييس مختلفة لحجم التأثير في الاختبارات غير البارامترية (Kerby 2014). من المحتمل أن يكون الإفصاح عن حجم التأثير الذي يحدد الفرق بين مجموعتين مفيداً وبدلاً عن قياس حجم التأثير الكلي. يمكن إكمال اختبار مان-ويتني U من خلال الارتباط المتسلسل الثنائي الرتبي (Glass 1965). يُحسب الارتباط المتسلسل الثنائي الرتبي (rank biserial correlation (r_{rb}) على النحو الآتي:

(١٢) لاحظ أن المعادلة (١٣، ٦) تقدم نسخة أبسط من التي قُدمت في البحث الأصلي لمان وويتني (Mann & Whitney

$$\frac{\text{متوسط الرتب المجموعة 1} - \text{متوسط الرتب المجموعة 2}}{\text{عدد كل الحالات}} = (r_{tb}) = \text{الارتباط المتسلسل الثنائي الرتبي}$$

كما هو حال معامل ارتباط بيرسون وسيرمان (المبحث ٢, ٥)، يمكن أن تأخذ r_{tb} قيماً تمتد من (-1) إلى (1). كلما كانت القيمة أكبر بالقيمة المطلقة كان الارتباط أقوى. تشير القيمة السالبة إلى أن للمجموعة الثانية قيمة ترتيبية أكبر. من الخيارات الأخرى لقياس حجم التأثير استخدام احتمال الأفضلية (PS) probability of superiority التي نوقشت في المبحث (٤, ٨).

ننوه أخيراً إلى أنه إذا كانت لدينا عينات متعددة لكل متحدث فردي وأردنا أن نعرف، على سبيل المثال، كيفية تطور لغتهم مع مرور الوقت أو كيفية تأثر هذا التطور بمتغير مستقل آخر بين نقطتين أو أكثر من العينات فإنه ينبغي استخدام اختبار القياسات المتكررة repeated measures tests التي نوقشت في هذا الفصل. تتطابق اختبارات القياسات المتكررة مع المتحدثين الفرديين عبر الأوضاع والأحوال، ولا تفترض التوزيع العشوائي للمتحدثين في مجموعات مختلفة (Verma 2016). هناك امتداد آخر للإجراءات الموضحة في هذا المبحث، ألا وهو تحليل التباين المتعدد (تحليل التباين العاملي) (multi-way ANOVA (factorial ANOVA، وهو اختبار يراعي متغيرين أو أكثر من المتغيرات المستقلة في الوقت نفسه كالجنس والطبقة الاجتماعية، ويراعي طبيعة التفاعل بينها (انظر: Field et al. 2012: الفصل الثاني عشر للاطلاع على مثال عملي).

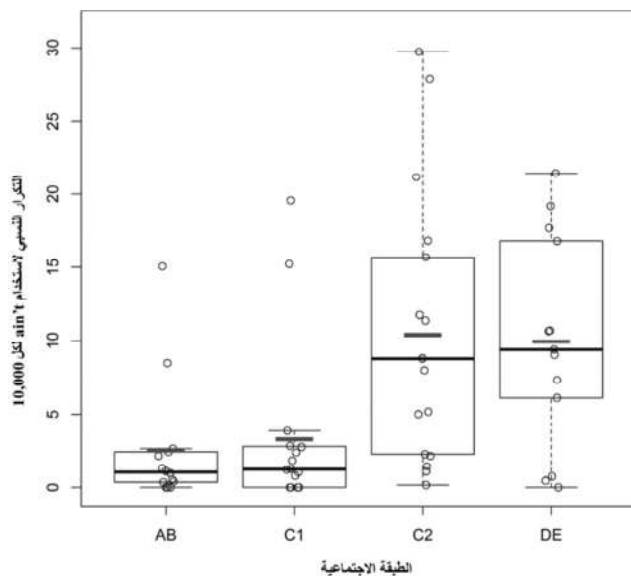
التقرير الإحصائي: اختبار-ت، وتحليل التباين، واختبار مان-ويتني U، واختبار كروسكال-والاس
١- ما يجب الإفصاح عنه

عند مقارنة مجموعات المتحدثين باستخدام اختبار-ت، وتحليل التباين، واختبار مان-ويتني U واختبار كروسكال-والاس، نُشير في التقرير إلى إحصائية الاختبار أولاً، ثم إلى درجات الحرية (للاختبار-ت وتحليل التباين) ثانياً، ثم إلى قيمة الاحتمال ثالثاً، ثم إلى حجم التأثير (بما في ذلك مجال الثقة ٩٥٪) رابعاً.

يُشار في التقرير إلى قيم الاحتمال الدقيقة ما لم تكن أصغر من (0.001). بالنسبة إلى اختبار تحليل التباين وكروسكال-والاس اللذين يُجريان على مجموعات متعددة، ينبغي أيضاً أن يُشار في التقرير إلى الاختبارات البعدية وأحجام تأثيرها ذات الصلة عند الاقتضاء. من الممكن أيضاً تشجيع الاختبارات بتمثيل البيانات كاستخدام مخططات الصندوق أو أعمدة الخطأ التي تُظهر مجالات الثقة أو استخدامها معاً.

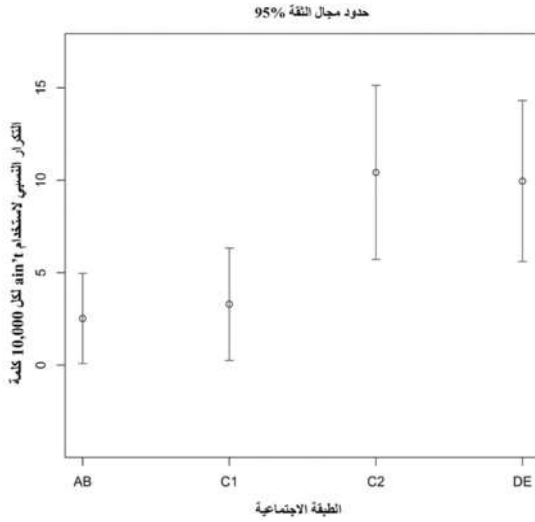
٢- مثال الإفصاح في التقرير

- اختبار-ت كان هناك تأثير ذو دلالة إحصائية للجنس على استخدام الضمائر الشخصية؛ اختبار-ت: (61.93) = 2.77؛ وقيمة الاحتمال: (0.007) إذ يستخدم الإناث (المتوسط = 1556.9 والانحراف المعياري = 154.3) الضمائر الشخصية أكثر من الذكور (المتوسط = 1451.8 والانحراف المعياري = 149.2). حجم التأثير متوسط حيث قيمة كوهن-د = 0.69 ومجال الثقة [0.18, 1.21].
- في مدونة BNC64 كان هناك تأثير كبير للجنس على استخدام الضمائر الشخصية؛ اختبار-ت: (61.93) = 2.77؛ وقيمة الاحتمال: (0.007)؛ وكوهن-د = 0.69 ومجال الثقة [0.18, 1.21].
- كان هناك تأثير كبير للطبقة الاجتماعية على استخدام الصيغة ain't: $F(3,56) = 5.59$ ، وقيمة الاحتمال: (0.002) وتربيع إيتا $\eta^2 = (0.432)$.



تحليل
التباين

شكل (٤، ٦). توزيع ain't في مدونة BNC64 للمتحدثين بتأثير الطبقة الاجتماعية.



شكل (٥, ٦). مجال الثقة لـ ain't في مدونة BNC64.

• كما نرى من الشكل (٤, ٦) هناك انقسام واضح بين المتحدثين (AB) و (C1) للطبقة المتوسطة، وبين المتحدثين (C2) و (DE) للطبقة العاملة. هذه الملاحظة صالحة أيضاً للمجموع اللغوي كما يتضح من مجالات الثقة في الشكل (٥, ٦).

• كان هناك تأثير كبير للطبقة الاجتماعية على استخدام الصيغة ain't وفق المقاييس الآتية: $F(3,56) = 5.59$ ، وقيمة الاحتمال $= 0.002$ ، وإيتا تربيع $\eta^2 = 0.432$. كما هو مبين في الاختبارات البعدية (بونفيروني)، فإن الفرق كبير بين (AB) و (C1، DE) وكذلك بين (C1) و (DE، C2) والدلالة الإحصائية من قيمة الاحتمال للكل أكبر من (0.05).

• يختلف استخدام الضمائر من المتحدثات (الوسيط $= 1,536.6$) اختلافاً كبيراً عن استخدام المتحدثين الذكور له (الوسيط $= 1433.0$)، وقيمة $U = 705$ ، وقيمة الاحتمال: $0.009 = r_{tb}$ (-0.19).

مان-ويتني

U

• كان هناك تأثير كبير للطبقة الاجتماعية على استخدام الصيغة ain't وفق المقاييس الآتية: $H(3) = 15.57$ ؛ وقيمة الاحتمال: (0.001).

كروسكال-

والاس

(١٣) الوسيط هو القيمة المتوسطة (انظر المبحث ١, ٣). في حالة الاختبارات غير البارامترية، يُبلغ عن الوسائط بدلاً من المتوسطات.

(٤, ٦) الأسلوب الفردي: تحليل التناظر

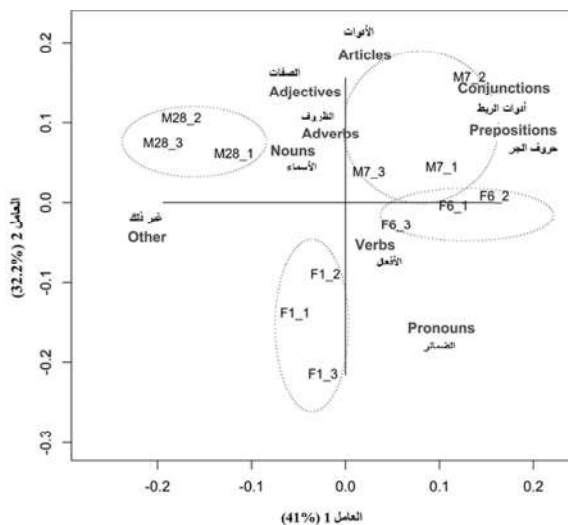
فكّر... قبل قراءة هذا البحث انظر إلى عينات الكلام المكتوب والمأخوذة من متحدثين مختلفين. تنتمي ثلاث عينات إلى أحدهما، وتنتمي العينة الرابعة إلى الآخر. هل يمكنك معرفة العينات التي تنتمي إلى المتحدث نفسه؟ هل هناك أي أدلة لغوية يمكن أن تساعدك على تمييز المتحدثين أحدهما عن الآخر؟	
Sample 2 Yes, exactly. Yeah. Yeah. But no it, it doesn't bother me. I don't mind you know on camera, I don't mind being on video, anything. <u>Warts</u> and all. Oh no no. I know I can't see Mike and Robin making a cake, can you? Not, no no no. العينة الثانية نعم بالضبط. نعم. نعم. لكن لا، هذا، هذا لا يزعجني. لا مانع من معرفتك بوجود الكاميرا، ولا مانع من أن تكون على شريط فيديو أو أي شيء. وارتس والجميع. أو كلا كلا. أعلم أنه لا يمكنني رؤية مايك وروبن يصنعان كعكة، هل يمكنك ذلك؟ لا، لا لا لا.	Sample 1 You lot ain't supposed to know I'm taping. Oh you wanna listen to <u>something</u>, but you don't know what you wanna listen. No! It's <u>nothing</u> to do with this <u>school</u>. No one in this <u>school</u> listens to it. Can I have a <u>look</u> at the <u>bottle</u> please? العينة الأولى ليس من المفترض أن تعرف أنني أسجل على الشريط. أوه، تريد الاستماع إلى شيء ما، لكنك لا تعرف ما تريد الاستماع إليه. لا! لا علاقة لهذا بهذه المدرسة. لا أحد في هذه المدرسة يستمع إليها. أيمكنني إلقاء نظرة على الزجاجاة من فضلك؟
Sample 4 You're fifteen <u>years</u> old and you still wanna watch the <u>Jungle Book</u>. Honestly! Some <u>people</u>! Oh don't start on me you know, saying I can't, there on Tuesday! Don't start because I'll, I'll smash your <u>face</u> in! What am I doing down here? العينة الرابعة تبلغ من العمر خمسة عشر عامًا وما زلت تريد مشاهدة كتاب الأدغال. بكل صراحة! بعض الناس! لا تنتقدي، كما تعلم، أقول لا أستطيع، يوم الثلاثاء! لا تبدأ لأنني سأحطم وجهك! ماذا أفعل هنا؟	Sample 3 And you were saying all this <u>stuff</u>. You were. I've got it on <u>tape</u>. That's what you were saying. And you were saying, oh yeah, we found these <u>porno magazines</u> and we're selling them off to <u>perverts</u>. And I said yeah but you're looking at <u>them</u> and, and you started laughing. العينة الثالثة وكنت تقول كل هذه الأشياء. كنت. لقد حصلت عليه على شريط. هذا ما كنت تقوله. وكنت تقول، أوه، لقد وجدنا هذه المجلات الإباحية، ونحن نبيعها للمنحرفين. وقلت نعم ولكنك تنظر إليه وبدأت في الضحك.
ما تحته خط: اسم. البنط العريض: فعل. الخط المائل: الضمائر.	

جرى التحقيق في كل أسلوب وفق مجموعة الخصائص اللغوية المعجمية والنحوية. لنلق نظرة على تقنية تشبه إلى حد ما تحليل العوامل الذي نُوقش في الفصل الخامس (المبحث ٤، ٥)، وُضع في المقام الأول لتحليل جداول النتائج المتقاطعة مع البيانات الفتوية ونوع البيانات كما نُوقش في الفصل الرابع (المبحث ٣، ٤)^(١٤). يطلق على هذه التقنية تحليل التناظر correspondence analysis، ويُعرف أحياناً أيضاً باسم القياس الأمثل optimal scaling وتحليل التجانس homogeneity analysis (Clausen 1998: 1). تحليل التناظر هو تقنية مختصرة ينتج عنه مخطط تناظر correspondence plot. مخطط التناظر هو تمثيل مرئي لجدولة متقاطعة في مساحة ثنائية الأبعاد في العادة، ويُجرى باستخدام قياس مسافة كاي تربيع مقياس بين قرب أو بعد الفئات المدرجة في الجدول. يعد مخطط التناظر الثنائي الأبعاد (الشكل ٦، ٦) الصورة الأكثر فائدة للواقع المعقد لأنها تقلل من عدد أبعاد التنوع لتجعلها في بُعدين يمكن تمثيلها على المحور السيني والصادي^(١٥).

الخاصية الفريدة لمخطط التناظر هي أنه يصوّر كل فئات الأعمدة والصفوف للجدولة المتقاطعة. فعلى سبيل المثال، يُظهر مخطط التناظر في الشكل (٦، ٦) ثلاثة متحدثين فرديين (F1 و F6 و M7 و M28) وفقاً لاستخدامهم فئات الكلمات المختلفة، وقد عُرِضت عينات المتحدث وفئات الكلمات منها في المخطط. واستخدمت بالنسبة لكل متحدث ثلاث عينات لكل 1000 كلمة تُتَبَّع فيها تكرارات الأسماء والأفعال والصفات. يمكننا ملاحظة أنه في كل مرة تتجمع عينات الكلام الثلاث من المتحدث نفسه وتتقارب، فإن ذلك يشير إلى تفرد الأسلوب اللغوي فيها. يمكن رؤية مجموعة البيانات الكاملة التي يستند إليها مخطط التناظر في الشكل (٦، ٦) وفي الجدول (١، ٦).

(١٤) على الرغم من أن هذه التقنية قد صُمِّمت للبيانات الفتوية مثل أعداد تكرار الخصائص اللغوية المختلفة أو الفئات النحوية المعجمية فإنها أكثر تنوعاً كما وضع ذلك غرينيكر (Greenacre 2007: 182-4) ويمكن استخدامها حتى مع بيانات المقياس (مثل التكرارات النسبية)، وعادةً ما تُحوّل بيانات المقياس إلى رتب أو قيم-ز في حالة وجود مقاييس مختلفة.

(١٥) مخططات التطابق الثلاثية الأبعاد ممكنة أيضاً، لكن تفسيرها أكثر صعوبة لأن تحليل ثلاثة أبعاد في الوقت نفسه يصبح معقداً إلى حد ما. تحليل بُعدين فقط عادةً ما يكون كافياً لتمثيل معظم التباين في البيانات.



شكل (٦، ٦). مخطط التناظر لفئات الكلمات في خطاب المتحدثين الفرديين.

جدول (٦، ١). الجدولة المتقاطعة لفئات الكلمات في خطاب المتحدثين الفرديين [يشير F إلى الأنثى، ويشير M إلى الذكر].

العينة	الأفعال	الأسماء	الضائير	الظروف	الصفات	حروف الجر	أدوات العطف	الأدوات	فئات أخرى
F1_1	257	154	197	42	81	39	26	39	165
F1_2	240	139	182	19	82	65	42	49	182
F1_3	267	101	208	30	83	53	39	26	193
F6_1	256	109	179	29	103	61	69	61	133
F6_2	257	115	181	48	79	78	68	54	120
F6_3	257	152	171	23	97	64	49	49	138
M7_1	250	121	163	31	134	68	57	55	121
M7_2	218	161	148	47	92	87	70	67	110
M7_3	248	130	151	47	107	76	41	53	147
M2B_1	235	154	124	41	108	48	52	49	189
M28_2	219	154	106	42	114	57	42	58	208
M2_8_3	219	141	120	48	100	40	47	66	219

جدول (٦، ٢). الجدولة المتقاطعة للأفعال والأدوات في خطاب متحدثين.

العينة	الأفعال	الأدوات	الإجمالي
F1_1	257	39	(إجمالي الصف) 296
M28_3	219	66	(إجمالي الصف) 285
الإجمالي	(إجمالي العمود) 476	(إجمالي العمود) 105	(الإجمالي الكلي) 581

لنعدّ سريعاً إلى مهمة (فَكَّرَ). تمثل العينة الأولى والثالثة والرابعة المتحدثة (F1) من الشكل (٦، ٦) والجدول (٦، ١) وتمثل العينة الثانية المتحدث M28. عمر المتحدث الشاب (F1) ١٤ عاماً من لندن، والمتحدث M28 كبير في السن من لانكشير. يُميّز هؤلاء المتحدثون بوضوح في التحليل بتكرارات مختلفة يستخدمون بها فئات الكلمات القياسية. قد لا يكون ذلك واضحاً على الفور عند النظر إلى عينات صغيرة لا تتجاوز ٥٠ كلمة كما في مهمة (فَكَّرَ)، ويتطلب عينات أكبر تزيد على ألف كلمة. يوضح هذا المثال موقفاً يكون فيه التقدير الكمي مفيداً بوضوح لأنه يساعدنا على الكشف عن الانتظام في استخدام اللغة، وهو انتظام يظل مخفياً عند استخدام التحليل النوعي فقط.

يرتبط تحليل التناظر من الناحية النظرية باختبار كاي تربيع الذي نوقش في المبحث (٣، ٤)، والذي يمكن إجراؤه على جدول نتائج تقاطعية ثنائي الاتجاه (على سبيل المثال، الجدول ٦، ١) والذي يختبر فرضية تفاوت المتباثل الصفرية، أي يحدد ما إذا كانت القيم تُوزع بالتساوي النسبي في الجدول من ضمن تباين الصدفة أو إذا كانت هناك اختلافات منهجية بين الفئات المعروضة في الجدول (Greenacre 2007: 26-9; Benzécri 1992: 54-5). باستخدام اختبار كاي تربيع نسأل عما إذا كان هناك فرق بين المتحدثين من جهة توزيع فئات الكلمات الفردية في كلامهم. بدلاً من قيمة الاحتمال التي ينتجها اختبار كاي تربيع يُظهر تحليل التناظر العلاقة بين الفئات واضحةً بتمثيل كل المتحدثين وفئات الكلمات في مخطط التناظر نفسه. على الرغم من أن اختبار كاي تربيع يمكن أن يجيب فقط عن سؤال تقريرى سهل (نعم أم لا) عن الدلالة الإحصائية دون الإشارة إلى المكان الذي يكمن فيه الاختلاف بالضبط الأمر الذي يمثل مشكلة مع جداول النتائج المتقاطعة، فيمكن لتحليل التناظر أن يُظهر لنا صورة أكبر للعلاقات المعقدة من التشابهات والاختلافات.

إذن كيف تُترجم المعلومات في الجدول إلى التمثيل البصري في مخطط التناظر؟ لنبدأ بأخذ صفين وعمودين من الجدول (٦، ١). تُعرض المعلومات عرضاً منفصلاً في الجدول (٦، ٢). لاحظ أيضاً أن (F1) و (M28) هما المتحدثان في مهمة (فَكَّرَ). يوضح الجدول (٦، ٢) أيضاً إجماليات الصفوف والأعمدة، والإجمالي الكلي لها.

تتمثل الخطوة الأولى في تحليل التناظر في تحويل المعلومات الموجودة في الجدول إلى ما يسمى بملفات التعريف profiles، وهي نسب مئوية مبنية على إجماليات الصفوف والأعمدة، ويُعبّر عنها في

صورة أرقام عشرية (مثل 0.87) بدلاً من النسب المئوية (المبحث ٣, ٤). تُحسب ملف التعريف بوصفه نظاماً للنسب المئوية (Benzécri 1992: 4) من خلال قسمة الرقم في الخلية على إجمالي الصفوف والأعمدة والصفوف على التوالي لملفات تعريف الصفوف والأعمدة. على سبيل المثال، تُحسب ملفات تعريف الصف للمتحدثين FI و M28 وفق ما هو موضح في الجدول (٣, ٦).

جدول (٣, ٦). ملفات التعريف للمتحدثين.

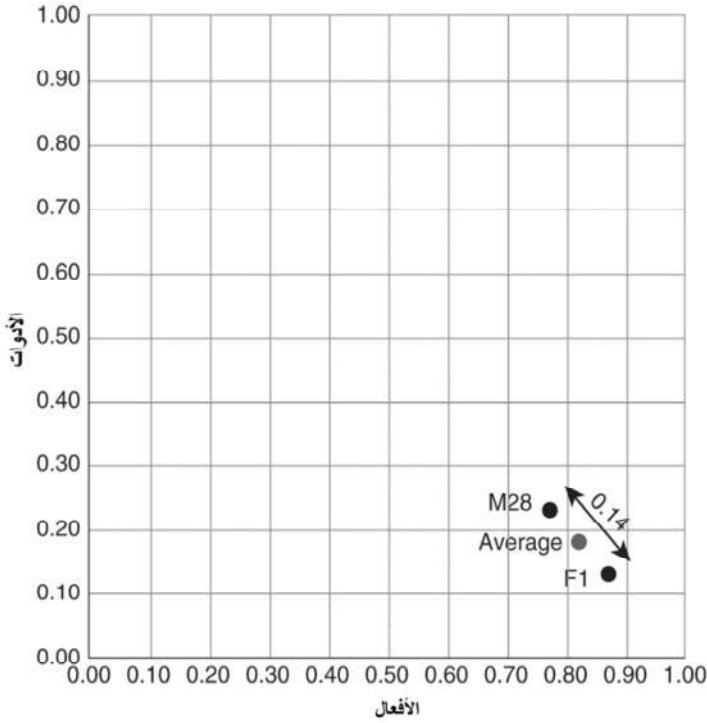
العينة	الأفعال	الأدوات	الإجمالي
FI_1	$257/296 = 0.87$	$39/296 = 0.13$	$296/296 = 1$
M28_3	$219/285 = 0.77$	$66/285 = 0.23$	$285/285 = 1$
تعريف المتوسط	$476/581 = 0.82$	$105/581 = 0.18$	$1=581/581$

تُحسب أيضًا ملفات تعريف الأعمدة وتُحلل في تحليل التناظر، ويمكن تمثيل ملفات الصفوف من الجدول (٣, ٦) بيانًا باستخدام مخطط تشتت بسيط (المبحث ٥, ١) الذي يعرض في مساحة ثنائية الأبعاد حالة تعريف المتحدث، وتمثل ملفات التعريف هذه نسب استخدام الأفعال والأدوات من المتحدثين. في مثالنا السابق فنسب الفعل في إحدائيات المحور السيني هي بواقع (0.87) للمتحدث FI و (0.77) للمتحدث M28، ونسب الأداة في إحدائيات المحور الصادي هي (0.13) للمتحدث FI و (0.23) للمتحدث M28. سيوضع تعريف المتوسط average profile في المتصف وفقًا لإحدائيات تعريف المتوسط (0.82 و 0.18 على التوالي). يمكن رؤية نتيجة المخطط في الشكل (٤, ٦).

يستخدم مخطط التشتت في الشكل (٦, ٧) ما يسمى المسافة الإقليدية بين النقاط في الشكل البياني، وهي المسافة الأكثر مباشرة بين النقاط، أي ما نعنيه عادةً بـ (المسافة) في الاستخدام اليومي للكلمة. توجد كما هو موضح في المبحث (٣, ٥) أنواع أخرى من المسافات (مثل مانهاتن وكانبرا)، ويمكن استخدامها لأغراض إحصائية. في تحليل التناظر نستخدم مسافة تسمى كاي تربيع ممتدة من معادلة اختبار كاي تربيع (Dodge 2008: 68; Greenacre 2007: 26-9). مسافة بُعد كاي تربيع هي شكل موزون للمسافة الإقليدية يعطي أهمية أكبر نسبيًا لفئات ذات عدد أقل من الحالات. تُحسب مسافة كاي تربيع على النحو الآتي:

مسافة كاي تربيع (أ، ب) = الجذر التربيعي لـ: $(س_ب - س_أ)^2 \div \text{تعريف متوسط س} + (ص_ب - ص_أ)^2 \div \text{تعريف متوسط ص} + \dots$

حيث (س_أ) هو الإحداثي الأول للنقطة أ، و(س_ب) هو الإحداثي الأول للنقطة ب، و(ص_ب) هو الإحداثي الثاني للنقطة ب، إلخ. يمكننا الاستمرار في إضافة فئات كلمات مختلفة في مثالنا لتحديد موضع الأهداف والانتقال من مساحة ثنائية الأبعاد إلى مساحة متعددة الأبعاد.

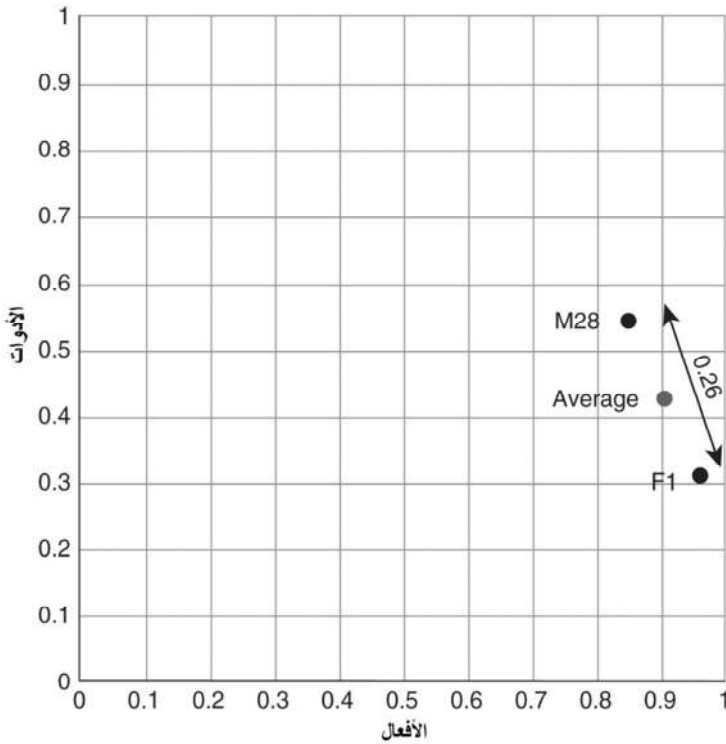


شكل (٦، ٧). ملفات تعريف المتحدث (الصف): المسافة الإقليدية.

في مثالنا تكون مسافة كاي تربيع على النحو الآتي:

$$(٦، ١٥) \quad 0.26 = \text{الناتج} = \sqrt{\frac{(0.77-0.87)^2}{0.82} + \frac{(0.23-0.13)^2}{0.18}} = \text{مسافة كاي تربيع (M28، F1)}$$

تكون مسافة كاي تربيع بين نقطتين أكبر دائماً من المسافة الإقليدية البسيطة كما هو موضح في الشكل (٦, ٨) الذي يعيد رسم البيانات من الجدول (٦, ٣) بمسافة كاي تربيع^(١٦). هناك العديد من الأسباب الرياضية لاستخدام كاي تربيع بدلاً من المسافة الإقليدية التي تتجاوز نطاق هذه المقدمة (Benzécri 1992: 44-58; Greenacre 2007). أفضل طريقة لفهم استخدام مسافة كاي تربيع في تحليل التناظر هي إدراك علاقتها باختبار كاي تربيع. يمكن استخدام اختبار كاي تربيع بوصفه البديل لتحليل التناظر لجدول النتائج المتقاطعة نفسها^(١٧).



شكل (٦, ٨). مسافة كاي تربيع للمف تعريف المتحدث (الصف).

(١٦) كما يتضح من الشكل (٦, ٧) فإن المسافة الإقليدية المكافئة هي 0.14.

(١٧) عندما لا يمكن استخدام اختبار كاي تربيع إلا مع الجدولة المتقاطعة كما هو في مثالنا فإن تحليل التباين ممكن مع متغيرات متعددة تُعرض في جداول متعددة الاتجاهات. للاطلاع على تحليل التباين المتعدد انظر غرينيكر (Greenacre 2007: 137ff).

تنطوي الخطوة الأخيرة الأكثر تعقيداً من تحليل التناظر على اختزال الأبعاد. يتضمن الجدول (٦, ٢) فئة الأفعال وفئة الأدوات التي يمكن عرضها في مخطط تشتت ثنائي الأبعاد. أما الجدول (٦, ١) فيتضمن تسع فئات يستحيل عرضها في مخطط التشتت البسيط. ولهذا السبب يتضمن تحليل التناظر إجراء اختزال الأبعاد، وهو يشبه إلى حد ما تحليل العوامل من ناحية إجرائه الذي نوقش في المبحث (٤, ٥). تُستبدل الفئات الفردية (الأدوات، والأفعال، إلخ) بعد ذلك في مخطط التشتت في الشكل (٦, ٨) بعوامل تجمع بين فئات متعددة، وتشرح أكبر قدر ممكن من التنوع في البيانات. يشابه مخطط التناظر الناتج -كما رأينا في الشكل (٦, ٦)- مع مخطط التشتت، إذ يحتوي على المحورين السيني والصادي ويمثلان العوامل دون الفئات المفردة. للإشارة إلى العلاقة بين الفئات الفردية والعوامل يعرض مخطط التناظر أيضاً الفئات الفردية. يوضح مخطط التناظر أيضاً مقدار التنوع الأصلي في البيانات التي يوضحها كل عامل. في الشكل (٦, ٦) يوضح العامل (1) نسبة 41 يوضح العامل (2) نسبة 32.2 من التنوع، وهو ما يعني أن مخطط التناظر يوضح ما نسبته 73.2 من التنوع، وهو مقدار معقول جداً، وتمثل المعاوضة في عملية اختزال الأبعاد في أن النسبة 26.8 من التنوع تضيع في الترجمة، وسنحتاج إلى مزيد من الأبعاد (العوامل) لتمثيل المزيد من التنوع، ولكن الصورة الإجمالية ستزداد تعقيداً. ستكون النتيجة الثنائية الأبعاد في بحثنا هنا مرضية لأنها تبين لنا بوضوح الاتجاهات الرئيسية في البيانات مع شرح مقدار كبير من التنوع يُشكّل ما نسبته أكثر من ٧٠٪. يمكنك عدّ مخطط التناظر خريطة تسهيل واختزال لأهم ملامح التضاريس التي تساعدنا على التنقل في جميع أنحاء المدينة أو عبر الريف وإيجاد طريقنا إلى الوجهة التي نقصدها.

عندما ننظر إلى مخطط التناظر في الشكل (٦, ٦) نلاحظ تجميعات من كلتا الفئتين للكلمات من عينات المتحدث. تساعدنا الفئات اللغوية على تفسير العوامل ومعانيها اللغوية. فعلى سبيل المثال، يمتد العامل (1) من فئات الكلمات النحوية الأخرى إلى الأفعال وحروف الجر وأدوات العطف، وتتمتع الفئتان الأخيرتان أيضاً بقيمة أعلى من العامل (2) الذي يمتد من الضمائر إلى الأدوات. يمكننا أيضاً ملاحظة المسافات (كاي تربيع) القريبة بين حروف الجر وأدوات العطف وكذلك بين الأسماء والصفات والظروف والأدوات. ويمكننا أيضاً أن نرى تجمع العينات المستمدة من خطاب المتحدثين نفسه (F6, F1)، (M28, M7) مترابطة نسبياً، ويمكننا قياس مسافات كاي تربيع بين العينات. تُعرض في المخطط عينات المتحدث مرتبطة بتفضيلاته لفئات الكلمات المختلفة. يشير المخطط بعدئذ إلى وجود دليل قوي في البيانات على أساليب متميزة للتحدث لدى هؤلاء المتحدثين الأربعة.

التقرير الإحصائي: تحليل التناظر

١- ما يجب الإفصاح عنه

من الأفضل تلخيص تحليل التناظر بمخطط التناظر الذي يستند إلى الجدولة المتقاطعة. ينبغي الإفصاح عن كل من المخطط والجدول، وينبغي أن نلاحظ النسبة الإجمالية للتنوع التي أوضحها العاملان والتفسير اللغوي. ينبغي أيضاً أن نحاول بعد ذلك تحديد تجميعات ذات مغزى من المتحدثين في المخطط مع التعليق على علاقتهم. لاحظ أنه لا يمكن تفسير مسافة كاي تربيع إلا من ضمن فئات الصف والعمود وليس بينها. يُمكننا في الشكل (٦، ٦) قياس مسافة كاي تربيع بدقة بين نقاط بيانات المتحدث الفردية، ولا يمكننا ربط نقاط المتحدث مباشرة بفئات الكلمات الفردية بقياس المسافة في المخطط لأن فئات الصفوف والأعمدة تعمل على مستويات مختلفة.

٢- مثال الإفصاح في التقرير

• حُلِّلت البيانات في الجدول (٦، ١) باستخدام تحليل التناظر. يُعرض مخطط التناظر الناتج في الشكل (٦، ٦) بوجه عام، ويُفسَّر مخطط التناظر 73.2% من التنوع في البيانات. يمكننا أن نرى أربع تجميعات رئيسية من عينات المتحدث.

(٦، ٥) السياق اللغوي: نمذجة التأثير الممتزج

فكّر...

قبل قراءة هذا المبحث انظر إلى المثالين أدناه. هل من فرقٍ في المعنى بين العبارتين؟ هل يمكنك تخيل سياقات يمكنك فيها قول هاتين العبارتين؟

1. This is really good هذا حقاً جيد
2. This is very good هذا جيد جداً

من المبادئ الأساسية في اللغويات الاجتماعية بمنهجها التقليدي ما يسميه لا بوف مبدأ الموثوقية^(١٨) (Tagliamonte 2006: 12-13 Labov 1972: 27). ينص هذا المبدأ على أنه لكي يكون

(١٨) قارن مع مبدأ ليتش (الموثوقية الكاملة) (Leech 1992: 112).

التحليل مفيداً يتعين علينا إيجاد جميع السياقات وتحليلها التي ترد فيها جميع أشكال المتغيرات الاجتماعية واللغوية المختلفة. في المبحث (٢, ٦) ناقشنا فكرة المتغير اللغوي الاجتماعي اللابوفي والقيود المحتملة. نأخذ المتغيرات اللغوية الاجتماعية اللابوفية نقطة انطلاق وننظر في الطرق التي يمكن من خلالها تحليل هذه المتغيرات. فعلى سبيل المثال، سوف نأخذ التباين بين العبارتين المنطوقتين في مهمة (فكر). هذا المثال مستوحى من إيتو وتايلامونتي (Ito & Tagliamonte 2003). لا يوجد فرق بين العبارتين من النظرة الأولى. من جهة المحتوى الرئيسي (يُطلق عليه أحياناً المعنى الافتراضي) فهاتان العبارتان يعنيان الشيء نفسه، وهو ما يعد معياراً مهماً لتحديد المتغير اللغوي الاجتماعي اللابوفي (المبحث ٢, ٦). باتباع مبدأ الموثوقية فإن خطوتنا التالية هي العثور على جميع الأمثلة في المجموعة التي (يمكن أن) يرد فيها هذا النوع من التنوع بوصفه متغيراً يحمل المعنى الآتي: (هذا جيد that is good)، ويسمى هذا النوع من حصر المتغيرات ذات المعنى المشترك تطويق حدود المتغير *circumscribing the variable context*. علينا أن نبحث في البيانات عن جميع الصفات المتدرجة من جهة المعنى بين القوة والخفة التي يمكن تحويلها بمقويات مثل *really* و *very* (Ito and Tagliamonte 2003:263-4). نريد في دراسة هذه الحالة أن نركز فقط في التباين بين المقويين *really* و *very*، وهما المقويان الأكثر شيوعاً في الكلام، وقد اخترنا فقط مجموعة فرعية من جميع الصفات في المدونة، وهي تلك الصفات المحوّرة بـ *really* أو *very* (أي: ١٩٥٢ حالة من ٥٨١٢٧ حالة ظهور للصفة في مدونة BNC64 التي تتضمن مليوناً ونصف المليون كلمة من الخطاب البريطاني (Brezina & Meyerhoff 2014)).

هذا يجعل رسم حدود المتغير أكثر تحديداً مع تضيق نطاق سؤال البحث الآتي: ما العوامل التي تؤثر في استخدام *really* في مقابل *very*؟

تتمثل الخطوة التحليلية الأولى في استخراج البيانات وتصنيف جميع الأمثلة في مجموعة البيانات (الشكل ٩, ٦) بالنظر في كل السياقات اللغوية التي ترد فيها *really* و *very*، وتسمى هذه الخطوة العوامل الداخلية *internal factors*. أما الخطوة الثانية فتتمثل بالنظر في المتحدث والمتغيرات المتعلقة بالموقف مثل العمر، والجنس، والنوع، إلخ، وتسمى العوامل الخارجية *external factors*. تُعرف العوامل الداخلية والخارجية باسم متغيرات المتنبئ *predictor variables*.

م	المحدث	النوع	السن	الفئة	التركيب النحوي	الناتج	يعود	العقدة	يسار
1	M31	أ-مذكر	77	A_DE	ب-خبرية	B_very	to your mum's apple p	very partial	on the weekend # You
2	F17	ب-مؤنث	40	C_C1	ب-خبرية	B_very	. # Computer studies p	very difficult	not found very difficult;
3	M30	أ-مذكر	70	D_AB	ب-خبرية	B_very	! # Yellowish you mea	very funny	the Police Station ! # H
4	F32	ب-مؤنث	75	B_C2	ب-خبرية	B_very	! # he didn't try, wel	very bright	every body # knew that
5	F17	ب-مؤنث	40	C_C1	ب-خبرية	B_very	and helpful. # If anyon	very punctual	my point of view i have
6	M30	أ-مذكر	70	D_AB	ب-خبرية	B_very	, you know ? # They m	very difficult	n't know whether we o
7	F19	ب-مؤنث	41	A_DE	أ-وصفية	B_very	cough down your ches	very bad	with not moving that th
8	F15	ب-مؤنث	37	A_DE	ب-خبرية	B_very	, is it safe ?, no its not	very safe	that film# isn't it? #is
9	F1	ب-مؤنث	14	B_C2	ب-خبرية	A_really	. #! like shooting, sho	really good	do, choo, do do do do
11	F4	ب-مؤنث	20	C1	ب-خبرية	A_really	In the area. # So my ta	really new	just tell everyone I'm n
12	M22	أ-مذكر	51	D_AB	أ-وصفية	B_very	point there Lynda # E	very good	me with bells on my an]
13	F21	ب-مؤنث	46	C_C1	أ-وصفية	B_very	people # at the time ,	very professional	, I said we didn't even
14	F9	ب-مؤنث	30	B_C2	ب-خبرية	B_very	to cut, I thought I did	very difficult	was very good cutting b
15	F11	ب-مؤنث	34	B_C2	ب-خبرية	A_really	, let me switch this thi	really interesting	we're gonna practice #
16	F24	ب-مؤنث	54	B_C2	أ-وصفية	B_very	character, I just find it	very peculiar	Right I'll leave that the;
17	M12	أ-مذكر	33	A_DE	ب-خبرية	B_very	#Don't know, I don't	very true	#That's true, that's tru
18	F11	ب-مؤنث	34	B_C2	ب-خبرية	B_very	that's what I find of h	very practical	, either way she # amaz

شكل (٩، ٦). نموذج من مجموعة البيانات اللغوية الاجتماعية: العوامل الداخلية والخارجية.

يوضح الشكل (٩، ٦) أول ١٨ سطرًا من مجموعة البيانات اللغوية الاجتماعية استنادًا إلى مدونة BNC64. تحتوي مجموعة البيانات بالكامل على ١٦٤٦ إدخالًا صحيحًا (١٩٥٢ سطرًا مطروحة منها الإدخالات الخطأ). نظرًا لأن المتغيرات الداخلية (الموقع النحوي في حالتنا) ينبغي ترميزها يدويًا، نحتاج إلى تحديد متغيرات المتنبي ذات الصلة قبل ترميز مجموعة البيانات بأكملها لتجنب إعادة الترميز. في حال وجود متغير تقييمي judgement variable واحد أو أكثر فمن الأفضل ترميز عينة عشوائية من البيانات ترميزًا مزدوجًا والإفصاح عن إحصائية ملائمة المقدّر البيني (راجع المبحث ٥، ٣).

تحتوي مجموعة البيانات في الشكل (٩، ٦) على متغير نحوي داخلي واحد وأربعة متغيرات تنبؤية خارجية (الجنس: ذكر أو أنثى، والعمر، والطبقة)، فمتغير الناتج هنا هو متغير ثنائي يضم very و really، ويظهر سطر الكشف السياقي الكامل في أعمدة (يمين والعقدة ويسار)، ولا يستخدم للتحليل الإحصائي، لكنه سيستخدم للترميز. لاحظ أن المتغيرات الفئوية في مجموعة البيانات مرمزة ببساطة: (A_) و (B_) و (C_) إلخ، وذلك لضمان التفسير المباشر للنتائج مقابل قيم المنطلق المرمزة بالبادئة A_ (المبحث ٤، ٤).

يُطلق على التقنية الإحصائية نمذجة التأثير الممتزج (سنوضح تطبيق هذه التقنية مع مجموعة البيانات عند عرض الشكل ٩, ٦). نماذج التأثير الممتزج mixed-effects models هي مجموعة من التقنيات الإحصائية المتطورة للمتغيرات المتعددة. سنركز هنا في استخدام واحد محدد فقط لنمذجة التأثير الممتزج يشبه إلى حد ما الانحدار اللوجستي (المبحث ٤, ٤).

التأثيرات العشوائية:				
اسم المجموعات	التباين	الانحراف المعياري		
المتحدث (الحصر)	0.9902	0.9951		
عدد الملاحظات: ١٦٤٦ / المجموعات: ٦٠ متحدثاً				
التأثيرات الثابتة:				
التقييم	خطأ الانحراف المعياري	قيمة z	نسبة الاحتمال (> z)	
(الحصر)	-0.89807	0.55256	-1.625	0.1041
الجنس ب_أنثى	-0.18425	0.32193	-0.572	0.5671
الفئة AB	0.13502	1.19296	0.113	0.9099
الفئة ب_ C2	0.19664	0.46684	0.421	0.6736
الفئة ج_ C1	0.07452	0.45810	0.163	0.8708
الفئة C1	-0.63974	1.13243	-0.565	0.5721
الفئة د_ AB	0.72511	0.46981	1.543	0.1227
السن	0.03913	0.00940	4.163	3.14e-05 ***
التركيب النحوي ب_خبري	0.33611	0.15168	2.216	0.0267 *
رموز الدلالة: (***) يعني القيمة (0.001) و(**) يعني القيمة (0.01) و(*) يعني القيمة (0.05) و(.) يعني القيمة (0.1) ولا شيء يعني القيمة (1)				

شكل (١٠, ٦). المخرجات لنمذجة التأثير الممتزج.

تصنع نمذجة التأثير الممتزج نموذجاً من البيانات التي تتنبأ على نحو أفضل بالمتغير (اللغوي) الناتج مع مراعاة كل من التأثيرات الثابتة والعشوائية، ومن ثم نمذجة التأثير الممتزج (Gries 2013a; Johnson 2009). تعدّ التأثيرات الثابتة fixed effects متغيرات مستقلة تمثل الهدف

المباشر للبحث الاجتماعى اللغوى مثل الجنس أو الطبقة الاجتماعية أو السن أو النوع. تمثل التأثيرات العشوائية random effects مصادر التنوع الأخرى التى ينبغى مراعاتها فى النموذج مثل تفضيلات المتحدث الفردية، وهى ليست جزءاً من سؤال البحث. يمكن لنمذجة التأثير الممتزج تحليل الموقف اللغوى المعقد وحساب كل من التنوع المنهجى الناجم عن عوامل خارجية وداخلية مختلفة تُصنّف بـ التأثيرات الثابتة، وكذلك التنوع الفردى بسبب تفضيلات المتحدثين وما إلى ذلك مما يُصنّف بـ التأثيرات العشوائية.

بسبب تشابه هذا النوع من النمذجة مع الانحدار اللوجستى الموضح بالتفصيل فى البحث (٤، ٤) نحيل القراء على الفصل الرابع لغرض الاطلاع على شرح مفصل للمصطلحات والمبادئ الأساسية. نسلط الضوء فيما يأتى على الاختلافات الرئيسية بين الانحدار اللوجستى ونمذجة التأثير الممتزج (الانحدار اللوجستى الممتزج التأثيرات) ونضرب مثلاً على استخدام هذه التقنية. عندما تُجرى تقنية نمذجة التأثير الممتزج نحصل على ناتج مثل ذلك المعروض فى الشكل (١٠، ٦).

يتكون تفسير نمذجة التأثير الممتزج من جزأين رئيسيين؛ الأول: تقييم خصائص النموذج، والثانى: تفسير أهمية التأثيرات الثابتة. فى مثالنا يُعدّ النموذج ذا دلالة إحصائية ويظهر تأثير العمر والموقع النحوى على استخدام really على عكس very (متغير المنطلق baseline فى تحليلنا موسوم بالبادئة A_). يفضل المتحدثون الأكبر سناً very والسياقات النحوية ذات الصلة الخبرية (مثل you're very lucky)، أما التنبؤات الأخرى فهى غير ذات دلالة. تتماشى هذه النتيجة مع دراسة إيتو وتايلامونتي (Ito and Tagliamonte 2003) التى تستند إلى مدونة مختلفة (أصغر) أفادت بالتحليل نتائج تفصيلية أكثر.

يكمن الفرق العملى الرئيس بين الانحدار اللوجستى البسيط وبين نمذجة التأثير الممتزج فى قدرة الثانى على مراعاة التنوع الفردى بين المتحدثين. هذه ميزة مرغوبة إلى حد كبير، ويجعل نمذجة التأثير الممتزج متفوقاً على الانحدار اللوجستى البسيط VARBRUL فى اللغويات الاجتماعية (Johnson 2009). لسد الفجوة المصطلحية بين اللغويات الاجتماعية ولغويات المدونة الحاسوبية سنشير فى البحث التالى إلى الفروقات بين هذين التخصصين من ناحية البيانات اللغوية الاجتماعية.

قاموس صغير للمصطلحات اللغوية الاجتماعية

في لغويات المدونة الحاسوبية واللغويات الاجتماعية التنوعية تُستخدم مصطلحات مختلفة تعني الشيء نفسه أو تُستخدم المصطلحات نفسها لتعني مفاهيم مختلفة (Johnson 2009) وهذا كله قد يُوقع في اللبس. في الآتي قاموس صغير يوضح هذا التداخل واللبس.

المصطلح التنوعي	المفهوم	المقابل في لغويات المدونة الحاسوبية
العامل factor	نوع المتحدثين (مثل الذكور أو الإناث، والصغار أو الكبار) أو نوع السياق (مثل الموقع النحوي) الذي يفضل استخدام شكل معين للمتغير اللغوي الاجتماعي	مستوى متغير المتنبئ level of a predictor variable
مجموعة العامل factor group	متغير يستخدم لشرح التنوع اللغوي (الاجتماعي)	المتغير التنبؤي أو العامل predictor variable or factor
أهمية العامل factor weight	يعبر هذا الرقم عن أهمية تأثير عامل بوصفه احتمالاً؛ عادة ما يُعبر عن المصطلح البديل (معامل) بوحدات احتمال اللوج	المعامل coefficient
مبدأ الموثوقية principle of accountability	مبدأ علمي عام، ينص على أننا بحاجة إلى حساب جميع الحالات ذات الصلة في البيانات؛ وهو عكس (الالتقاطية)	الاستقراء الكامل total accountability
المتغير اللغوي الاجتماعي sociolinguistic variable	المتغير اللغوي الذي نحن مهتمون به؛ يحتوي المتغير اللغوي الاجتماعي على العديد من المتغيرات (في العادة اثنان) التي تعبر عن المعنى نفسه	متغير الناتج outcome variable
الكلمة الفعلية token	حدوث خاصية لغوية رئيسية. أما (الكلمة النوعية) في لسانيات المدونة فهي أي كلمة جارية في المدونة	الحالة، الملاحظة case, observation
تحليل الانحدار اللوجستي البسيط VARBRUL	إجراء إحصائي نموذجي في علم اللغة الاجتماعي التبايني	(شكل من أشكال) الانحدار اللوجستي (a form of) logistic regression
شكل من أشكال المتغير اللغوي الاجتماعي variant of a sociolinguistic variable	أحد الأشكال/الصيغ اللغوية التي تتنافس في سياق معين ولها المعنى نفسه	مستوى متغير ناتج level of an outcome variable

التقرير الإحصائي: نمذجة التأثير الممتزج

١- ما يجب الإفصاح عنه

من المهم عند استخدام نمذجة التأثير الممتزج الإفصاح عن نوع النمذجة وتفاصيل النموذج. سينصب التركيز في تأثير المتنبئات الفردية.

٢- مثال الإفصاح في التقرير

• استخدم الانحدار اللوجستي ذو التأثير الممتزج مع المتحدثين الفرديين بوصفه تأثيرًا عشوائيًا، واستخدم الجنس والطبقة والعمر والموقع النحوي بوصفها تأثيرات ثابتة. وقد أظهر النموذج الذي كان دالًا ($p < .001$)، تأثيرًا مهمًا للعمر والموقع النحوي، للمتحدثون الكبار والسياقات الخبرية بفضل المقوي very.

(٦، ٦) التطبيق: من هذا الشخص من البيت الأبيض؟

لدي صديق لم يرغب في الكشف عن هويته وجنسه ويرسل إلي كل عام لغزًا لغويًا بدلًا من بطاقة عيد الميلاد. كان اللغز هذا العام لغويًا اجتماعيًا في رابط بريد إلكتروني يحتوي على ملف يتضمن خطابًا مكتوبًا كما هو في الآتي:

From: [Redacted]
Sent: 19 December 2015 17:13
To: Brezina, Vaclav
Subject: Happy birthday

Message Speech_transcript.zip

Sociolinguistic Riddle

I am a man or a woman.
 I am young or old.
 I work in Washington DC and
 speak for the President.
 Who am I?
 (Find out without Google.)

لغز اجتماعي لغوي

أرجل أنا أم امرأة؟
 صغير أم كبير؟
 أعمل في واشنطن العاصمة
 وأتحدث باسم الرئيس.
 من أنا؟
 (حاول الإجابة دون الاستعانة بجوجل)

Attachment: Speech_sample.zip

The good news is that the filibuster has been broken on national service. The Senate has decided that there is no need for a second cloture vote. Mitchell and Dole just announced that national service will be the first order of business on Tuesday and we are fully confident that the Senate bill will now pass. So good news breaking out all over. [...]

[الترجمة] ملحق: ملف مضغوط_ عينة كلامية

الخبر السار هو أن الماطلة السياسية بخصوص الأمن القومي قد تُحُلِّي عنها. قرر مجلس الشيوخ أنه لا توجد حاجة لإجراء تصويت ثانوي لإقفال المناقشة. أعلن ميتشيل ودول للتو أن الخدمة الوطنية ستكون أول طلب للعمل يوم الثلاثاء، ونحن على ثقة تامة بأن مشروع قانون مجلس الشيوخ سيوافق عليه الآن. أخبار جيدة جدًا تحدث في كل مكان. [...]

قضيتُ وقتًا في حل اللغز بعد أيام قليلة من تلقي البريد الإلكتروني. هناك طريقتان أساسيتان في اللغويات القضائية تعتمدان على مقدار الأدلة المتاحة، ولدي بحث في هذا السياق يشبه عملَ محقق لغوي قضائي. إذا كانت كمية الأدلة صغيرة وتتكون من بضع جمل أو فقرات فإن القراءة الدقيقة لسمات استخدام اللغة الخاصة shibboleths فيها تكون كافية. أما إذا كان هناك كثير من البيانات المتاحة كما كان الحال بالنسبة لي في عينة دراستي التي تتكون من نحو مئتي ألف كلمة فلا بد من المنهج الإحصائي. كان الجزء الثاني من اللغز واضحًا ومطابقًا لنوع اللغة في العينة، إذ أشارت الجملة الآتية "أعمل في واشنطن العاصمة وأتحدث باسم الرئيس" إلى أن عينة الخطاب تأتي من السكرتير الصحفي للبيت الأبيض. لحسن الحظ، استطعت استخدام مدونة مقارنة للمؤتمرات الصحفية للبيت الأبيض (Barlow 2013) لمحاولة مطابقة العينة في الملحق التي أسميتها النسخة س مع بيانات مدونة البيت الأبيض. في الجدول (٤، ٦) ملحة عامة عن المدونة التي استخدمتها.

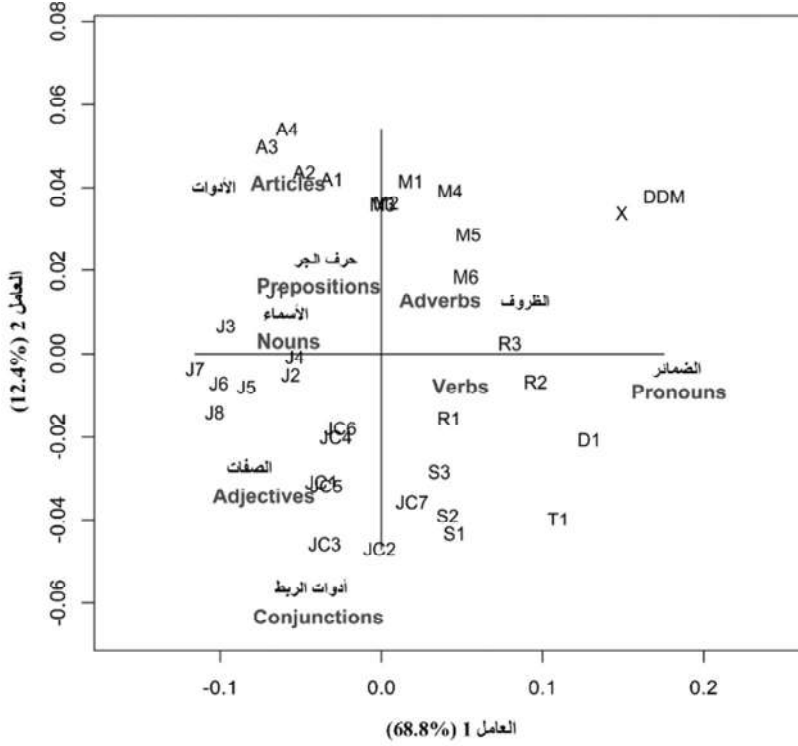
كانت الخطوة الأولى هي تحديد ما إذا كانت (النسخة س) من إنتاج متحدث أو متحدثة. في المدونة أنثيان (دي دي مايرز ودانا بيرينو Dee Dee Myers and Dana Perino) وستة ذكور. تُسهم كل أنثى بعينة واحدة في حين توجد ٣٣ عينة كلام من مجموعة الذكور. لا يمكن استخدام الكلمات التقليدية التي تميز بين الجنسين مثل lovely التي تفضلها النساء في الخطاب البريطاني غير الرسمي؛ لأنه في المدونة بأكملها أكثر من ستة ملايين كلمة. لا توجد سوى عشرين حالة ظهور لكلمة lovely ولم يكن ذلك مفاجئًا لأننا نتعامل مع سجل منطوق رسمي ومحدد للغاية للغة الإنجليزية الأمريكية. نحن بحاجة هنا إلى النظر في المزيد من الخصائص العامة مثل استخدام الضمائر الشخصية التي يمكن افتراضها لتمييز جنس المتحدث (Argamon et al. 2003; Rayson et al. 1997). للمقارنة بين مجموعتين من المتحدثين، يُستخدم اختبار-ت عادةً. على الرغم من إغراء استخدام جميع البيانات الموجودة في المدونة (٣٣ عينة كلام) فذلك غير ممكن. يتطلب افتراض اختبار-ت اختيار كل نص

اختياراً مستقلاً عن النص الآخر لتصميم عينات غير متباينة، وسيستقضى هذا الافتراض إذا ضمنا نصوصاً متعددة لكل متحدث. إذا اخترنا العينة في نص واحد لكل متحدث فسوف ينتهي بنا الأمر إلى تسعة نصوص، اثنان منها للإناث وسبعة للذكور. هذه عينة صغيرة للغاية وغير منتظمة إذ عدد الذكور فيها أكبر من عدد الإناث. توصي المعرفة الإحصائية التقليدية (على سبيل المثال: Siegel 1956) بعدم استخدام اختبار-ت في مثل هذه الحالات. يُشير بحث جديد (de Winter 2013) إلى أن اختبار-ت فعال حتى مع وجود عينات صغيرة جداً وغير منتظمة طالما لم يُنتقض افتراض المساواة في التباينات بشدة. يمكننا الاعتناء بهذا الافتراض باستخدام نسخة ويلتش الفعالة أكثر من اختبار-ت.

جدول (٤، ٦). مدونة البيت الأبيض.

اسم الملف	عدد الكلمات الفعلية	عدد العينات (٢٠٠٠٠٠ كلمة)	أعوام التوظيف	تاريخ الميلاد	السكرتير الصحفي
DDM	٠,٢ مليون	1	1993-4	1961	دي دي مايرز
M1, M2, M3, M4, M5, M6	١,٢ مليون	6	1995-8	1954	مايك مكوري
A1, A2, A3, A4	٠,٨ مليون	4	2001-3	1960	آري فليشر
S1, S2, S3	٠,٦ مليون	3	2003-6	1968	سكوت مكيلان
T1	٠,٢ مليون	1	2006-7	1955	توني سنو
D1	٠,٢ مليون	1	2007-9	1972	دانا بيرينو
JC1, JC2, JC3, JC4, JC5, JC6, JC7	١,٤ مليون	7	2011-14	1965	جاي كارني
J1, J2, J3, J4, J5, J6, J7, J8	١,٦ مليون	8	2014-16	1975	كوش إيرنست

يُنتج اختبار-ت نتيجة ذات دلالة إحصائية [ت (3.87) = 4.47، وقيمة الاحتمال <0.05] لكشف الفرق بين الأمراء الصحفيين من الذكور والإناث في البيت الأبيض عند النظر في استخدام الضمائر الشخصية. كان متوسط مجموعة الذكور 14232.86 والانحراف المعياري 1672.08، في حين كان متوسط مجموعة الإناث 18065.50 والانحراف المعياري 820.95. كانت قيمة النسخة س (18,044) وهي الأقرب بكثير إلى متوسط مجموعة الإناث. يمكن افتراض أن النسخة س تنتمي إلى متحدث وليس متحدثاً. في المدونة هناك متحدثان: دي دي مايرز ودانا بيرينو؛ فأيهما هي؟



شكل (١١, ٦). تحليل التناظر لاستخدام أسماء الصحافة بالبيت الأبيض فئات الكلمات.

للإجابة عن هذا السؤال استخدم تحليل التناظر مع جميع عينات الكلام في المدونة بالإضافة إلى النسخة س (1 + 34: عدد العينات، والرقم ١ يشير إلى المتحدثة التي يتطلب تحديد اسمها). نظر التحليل في نسب فئات الكلمات المختلفة في هذه العينات بوصفها المتغيرات اللغوية المتكررة والمستقلة للموضوع (الشكل ١١, ٦).

ضمّن تحليل التناظر عينات الكلام الفردية بوضوح من الأسماء الصحفيين بالبيت الأبيض. فعلى سبيل المثال، تتجمع جميع عينات كلام آري فليشر (A1-A4) في أعلى اليسار، في حين تتجمع عينات كلام سكوت مكيلان (S1-S3) في أسفل اليمين لتكون أقرب إلى المتوسط من عينات كلام توني سنو (T1) ودانا بيرينو (D1). ماذا عن النسخة س الغامضة؟ في الشكل البياني تتجمع على مقربة من عينة دي دي مايرز وبعيداً عن العينات الأخرى. يمكن الافتراض بأن النسخة س تنتمي إلى

السكرتير الصحفي للبيت الأبيض. يتميز خطاب دي دي مايرز -أول امرأة تتولى هذا المنصب- بالاستخدام المتكرر للضمائر، والاستخدام غير المتكرر نسبياً للأسماء والصفات وأدوات العطف. لا يمكن تحديد هذه التنوعات الأسلوبية إلا بأخذ عينات الكلام الكبيرة وتحليلها كمياً.

بعد البحث والتحقق أرسلت لصديقي البريد الإلكتروني الآتي:

From: Brezina, Vaclav

Sent: 1 January 2016 00:30

To: [REDACTED]

Subject: Re: Happy birthday

هل هي دي دي مايرز (لا حاجة لقيمة احتمال)؟ سأبحث في جوجل (؛

Was it Dee Dee Myers (no p-value needed)? I'll go and google it (؛

بالنسبة لأولئك الذين لا يزالون غير مقتنعين يمكنكم أن تتحدوا حذوي، وتحاولوا البحث عن الجملة الأولى من النسخة س على الإنترنت:

the good news is that the filibuster has been broken on national service

الخبر السار هو أن الماطلة السياسية بخصوص الأمن القومي قد نُحِّلَتْ عنها

(٦، ٧) تمارين

١- ما التخصصات اللغوية التي تستخدم مفهوم (الأسلوب) بوصفه أحد مفاهيمها الأساسية؟ كيف يمكن استخدامه؟

٢- ما حالات التنوع الآتية التي تطابق تعريف لابوف للمتغير اللغوي الاجتماعي ويمكن بعدئذ بحثه باستخدام طرق لابوف؟ علل جوابك.

(أ) حذف حرف h في سياقات اجتماعية مختلفة: أي التنوع بين نطق كلمة hair على سبيل

المثال: /heə/ أو /eə/.

(ب) التنوع بين إستراتيجيات تسمية المشروبات الغازية في الولايات المتحدة، على سبيل

المثال soda أو pop أو coke، (<http://popvssoda.com>)

(ج) التنوع بين البناء للمجهول والبناء للمعلوم، على سبيل المثال I did it مقابل It was done.

(د) التنوع بين الأفعال التي تعبر عن التفضيلات بدرجات مختلفة من القوة، مثل: adore

love و like و don't mind.

(هـ) استخدام أسلوب الاحتراز في الخطاب والتنوع بين الكلام الاحترازي وغير الاحترازي، مثل: I sort of agree مقابل I agree.

٣- انظر إلى الطرق الآتية عن نتائج اختبار-ت، وتحليل التباين ف F ، واختبار مان-ويتني U . حدد المعلومات الخطأ أو المفقودة.

(أ) ت = 2.77 قيمة الاحتمال = 0.007 د = 0.69

(ب) ف (56) = 5.95، قيمة الاحتمال = 0.00201359 <

(ج) $U = 705$ ، $p = 1.3$

٤- استخدم المعادلات في هذا الفصل لحساب إحصائيات الاختبار (وليس قيمة الاحتمال) لما يأتي:

(أ) استخدام lovely في خطاب الذكور والإناث [اختبار-ت ومان-ويتني U]

مجموعة ذكور (ع1 = 10): 0.91، 1.4، 2.18، 6.21، 2.63، 1.2، 0، 1.06، 6.49، 5.43

مجموعة إناث (ع2 = 10): 8.84، 1.09، 12.47، 1.65، 3.93، 1.1، 4.11، 21.21، 2.51، 0.47

(ب) استخدام immit في خطاب المتحدثين من الجنوب والوسط والشمال [تحليل التباين الأحادي].

جهة الجنوب (ع1 = 10): 4.19، 29.29، 5، 30.43، 6.09، 12.77، 25.93، 0.61، 28.08، 15.94

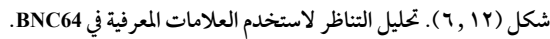
الجهات الوسطى (ع2 = 10): 9.68، 3.65، 1.2، 0، 2.07، 2.26، 5.18، 0، 0

جهة الشمال (ع3 = 10): 9.09، 9.09، 0، 7.38، 5.77، 0، 4.47، 0، 3.23

٥- احسب كوهن-د لمجموعة البيانات (أ) في التمرين الرابع.

٦- استخدم أداة مقارنة المجموعة للتحقق من نتائج التمرين الرابع والخامس.

٧- فُسر مخطط التناظر في الشكل (١٢، ٦) الذي يستند إلى مدونة BNC64 وهي مدونة للخطاب البريطاني غير الرسمي. يُمثل ١٦ متحدًا و١٦ محدثة على الشكل البياني وفقًا لاستخدامهم لأنواع مختلفة من دلالات درجات اليقينية (certainly، maybe، وperhaps، وpossibly، إلخ). يُميز كل متحد ومحدثة وفقًا لجنسه: ($\underline{M}1a$ و $\underline{F}1a$)، ورقمه ($\underline{F}1a$) وفرع العينة ($\underline{F}1a$ ، $\underline{F}1b$)، ولكل متحد وعينتان.



معيار معلومات أكايكي	معيار معلومات	الاحتمالية اللوغارتمية	الانحراف	درجة حرية المتبقيات
2412.3	2456.9	-1198.1	2396.3	1943
المتبقيات الموزونة	بازي			
scaled residuals				
الحد الأدنى	الربع ١	الوسيط	الربع ٣	الحد الأقصى
-1.9037	-0.6202	-0.3895	0.7452	2.4673
التأثيرات العشوائية	الانحراف	التباين		
random effects	المعياري			
اسم المجموعات	1.233	1.52		
النص (الحصر)				

عدد الملاحظات: التأثيرات الثابتة:	1951	المجموعات:	Text, 743
(الخصر)	التقييم	خطأ المعيار	قيمة ز
0.1783	0.2364	0.754	0.450635
0.1567	0.1515	1.034	0.300982
-0.9825	0.2717	-3.616	0.000299 ***
-0.2894	0.2569	-1.126	0.259977
-0.7682	0.2920	-2.631	0.008511 **
-1.1864	0.2153	-5.511	3.58e-08 ***
-0.8857	0.2030	-4.363	1.29e-05 ***
رموز الدلالة: (***) يعني القيمة (0.001) و(**) يعني القيمة (0.01) و(*) يعني القيمة (0.05) و(.) يعني القيمة (0.1) ولا شيء يعني القيمة (1)			

٩- هل يتلفظ الرجال أكثر بالكلمات النابية من النساء؟ استخدم أداة البحث والمقارنة بمدونة BNC64 (corpora.iancs.ac.uk/bnc64) لاختبار الفرضيات اللغوية الاجتماعية المختلفة عن وجود الكلمات النابية في خطاب بريطاني غير رسمي. اعتنِ بمختلف التدابير الإحصائية وتفسيرها. املاً الجدول (٥, ٦) وأول اعتناءً خاصاً لمختلف الإجراءات الإحصائية وتفسيرها.

جدول (٥, ٦). القسّم والجنس: مدونة BNC64.

الكلمات النابية	النتيجة دالة إحصائياً؟	اختلاف ذو مغزى؟
١-		
٢-		
٣-		
٤-		

تذكر الآتي

- يمكن توظيف الاختلاف اللغوي الاجتماعي بطرق مختلفة باستخدام المتغير اللغوي الاجتماعي اللابوفي الذي يحافظ على المعنى الأساسي، أو باتباع المنهج الوظيفي والنظر في توزيع الخصائص اللغوية في مجموعات المتحدثين.

- يُستخدم اختبار-ت وتحليل التباين (وكذلك مان-ويتني U وكروسكال-والاس غير البارامتريين) لبحث تأثير المتغيرات الاجتماعية المستقلة (الجنس والطبقة الاجتماعية) في استخدام خصائص لغوية مختلفة.
- تحليل التناظر هو تقنية تحليلية استكشافية تقارن استخدام متغيرات متعددة لمُتحدثين مختلفين، وتُختزل تلك المتغيرات في عاملين لغرض استنتاج تمثيل بياني فعال يُسمى مخطط التناظر.
- نموذج التأثير الممتزج هي مجموعة من التقنيات الإحصائية المعقدة التي يمكنها حساب متغيرات متعددة في الوقت نفسه عن تأثير التنوع الفردي بين المتحدثين.

قراءات إضافية

- Brezina, V. & Meyerhoff, M. (2014). Significant or random? A critical review of sociolinguistic generalisations based on large corpora. *International Journal of Corpus Linguistics*, 19(1), 1-28.
- Cardinal, R. N. & Aitken, M. R. (2013). *ANOVA for the behavioral sciences researcher*. Hove: Psychology Press.
- Corder, G. W. & Foreman, D. I. (2009). *Nonparametric statistics for non-statisticians: a step-by-step approach*. Hoboken, NJ: Wiley.
- Coupland, N. (2007). *Style: language variation and identity*. Cambridge University Press.
- Greenacre, M. (2007). *Correspondence analysis in practice*. Boca Raton, FL: Chapman & Hall/CRC.
- Johnson, D. E. (2009). Getting off the GoldVarb standard: introducing Rbrul for mixed-effects variable rule analysis. *Language and Linguistics Compass*, 3(1), 359-83.
- Labov, W. (2010). *Principles of linguistic change*, vol. 3: *Cognitive and cultural factors*. Oxford: Wiley-Black well.

أدوات لانكستر الإحصائية

- ١- يمكن استنساخ جميع التحليلات المبينة في هذا الفصل باستخدام أدوات لانكستر الإحصائية. تشمل الأدوات المتاحة على وجه الخصوص ما يأتي:
 - أداة مقارنة المجموعة.
 - أداة تحليل التناظر.
 - نموذج التأثير الممتزج للانحدار اللوجستي.
 - مدونة BNC64.
- ٢- يقدم الموقع أيضًا موادًا إضافية للطلاب والمعلمين.

التغير عبر الزمن: البيانات التعااقبية (التاريخية)

(١, ٧) ما الذي يتناوله هذا الفصل؟

يناقش هذا الفصل الإجراءات الإحصائية التي يمكن استخدامها لدراسة البيانات التاريخية أو التعااقبية عبر الزمن، وهي البيانات التي توضح تطور المتغيرات اللغوية مع مرور الوقت. تُحدّد خصائص معيّنة للدراسات التعااقبية أولاً، ثم تُطبّق تقنيات توفر تمثيلات فعالة للتغير التعااقبي. يركز الفصل على المقارنة الإحصائية لفترتين زمنيتين باستخدام إجراء يسمى التمهيد bootstrapping. يلي ذلك مناقشة التطبيق التعااقبي للتحليل التجميعي الذي نوقش سابقاً في الفصل الخامس. سنقدم نوعاً خاصاً من تقنية التجميع يُسمى تحليل التجميع المجاور ويراعي الترتيب التعااقبي للبيانات. يقف الفصل بعد ذلك على منهج الكشف الإحصائي عن القمم والقعور في البيانات التعااقبية، وما يمتد منها من تحليل يسمى تحليل التقلّب الاستعمالي Usage Fluctuation Analysis. تطبّق تقنية القمم والقعور نموذج الانحدار غير الخطي على البيانات لتحديد النقاط الزمنية القصوى أو القيم المتطرفة التي وقع فيها تغير كبير في الخطاب. يتتبع تحليل التقلّب الاستعمالي تطور مُصاحبات الكلمة المعنية مع مرور الوقت، وباستخدام تقنية القمم والقعور يحدد النقاط التي حدث فيها تغير كبير في استخدام كلمة ما.

سنجيب في هذا الفصل عن الأسئلة الخمسة الآتية:

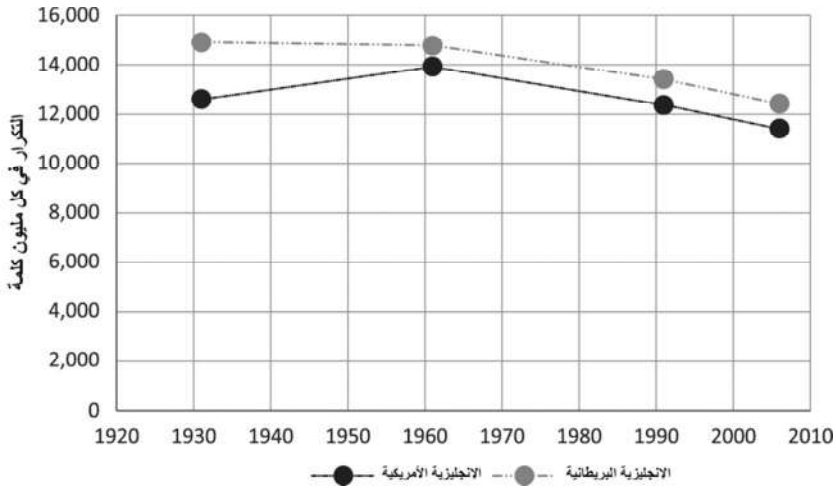
- كيف يمكننا قياس تطور اللغة وتمثيله مع مرور الوقت؟ (المبحث ٢, ٧).
- كيف يمكننا تقييم ما إذا كان قد حدث تغير مع مرور الوقت؟ (المبحث ٣, ٧).
- كيف يمكننا تجميع الفترات الزمنية إحصائياً؟ (المبحث ٤, ٧).

- كيف يمكننا نمذجة التغير في الخطاب وتمثيله؟ (المبحث ٥، ٧).
- كيف يمكن استخدام التقنيات التي نوقشت في هذا الفصل في البحث؟ (المبحث ٦، ٧).

(٧، ٢) قياس الزمن وتمثيله بوصفه متغيراً

فكّر...

انظر إلى الشكل (٧، ١) قبل قراءة هذا المبحث. بماذا يفيد حول تطور الأفعال الناقصة في اللغة الإنجليزية في القرن العشرين؟



شكل (٧، ١). الأفعال الناقصة في مدونات عائلة براون.

تُعدّ فكرة التغير مع مرور الوقت أساسية للتحليلات في هذا الفصل، ويعد الوقت متغيراً مقياسياً مستمراً من منظور إحصائي إذ يمكننا قياس الوقت على مدى قرون، أو عقود، أو سنوات، أو أشهر، أو أسابيع، أو أيام، أو ساعات، أو دقائق، أو ثوان، أو ميلي ثانية. تسمى الدراسة التي تتضمن وقتاً متغيراً دراسة تعاقبية diachronic أو طولية longitudinal. يُمكن قياس الوقت بالنسبة لمراحل الحياة البشرية (أي: الأعمار)، وهو متغير مهم في الدراسات اللغوية الاجتماعية (الفصل السادس) ودراسات اكتساب اللغة. لأن أغلب الدراسات التعاقبية تركز على التغير في اللغة فينبغي

ألا ننسى الجانب الآخر من تغير اللغة الذي يتعلق بالثبات مع مرور الزمن لكونه لا يقل أهمية عن التغير. تشبه اللغة منزلًا كبيرًا من الطوب المتعدد الطوابق يخضع لتجديد مستمر في طوابق مختلفة، ولاستبدال بعض الطوب لا بد من بقاء بعضه الآخر في الهيكل للحفاظ على المنزل قائمًا. يركز بيكر Baker (٢٠١١) في دراسته التغير في اللغة الإنجليزية البريطانية على مدار القرن العشرين في الكلمات التي تغيرت وتيرة تكرارها أو معانيها أو كلاهما، وكذلك في الكلمات التي ظلت ثابتة مستقرة. يسمي بيكر هذه الكلمات التي تظل ثابتة في استخدامها كلمات الإيقاف lockwords ويعطي money و times مثالين لها. من الناحية الإحصائية، يكمن التحدي الرئيسي هنا في كيفية تحديد الثبات في تكرارات الكلمات في مختلف المدونات، وبعبارة أخرى، كيفية تحديد كلمات الإيقاف؛ لأن النموذج العلمي والإحصائي برمته يركز في تحديد الاختلافات والتغير، ويتجاهل الثبات في أغلب الأحيان. ثم إن الاستنتاجات الواردة في الإحصائيات مصوغة بطريقة لا تجعلنا نقبل أبدًا الفرضية الصفريّة التي تنص بطبيعتها على أنه لا يوجد فرق أو تغير، وليس لدينا أدلة كافية لرفض الفرضية الصفريّة لأن عدم وجود أدلة لرفضها لا يعني أنها صحيحة. هذا يشبه سياق المتهم في المحكمة، فبراءته لا تعني بالضرورة عدم ارتكابه الجريمة لاحتمال عدم نجاح الادعاء في إثبات التهمة عليه (المبحث ٣، ١). نحتاج لتحديد الثبات عندئذ إلى الاعتماد على موارد الإحصاءات الوصفية وفهم جيد للبيانات من أجل تحديد العناصر الثابتة في استخدام اللغة (انظر: مناقشة كلمات الإيقاف في المبحث ٤، ٣).

نحتاج إلى مدونات تاريخية أو تعاقبية لإجراء هذا النوع من التحليل آليًا. تختبر المدونات التعاقبية diachronic corpora مراحل مختلفة من تطور اللغة أو الخطاب مع مرور الزمن. تشمل أمثلة المدونات التاريخية للغة الإنجليزية عائلة براون^(١)، ومدونة هلسنكي للنصوص الإنجليزية، ومدونة مراسلات اللغة الإنجليزية المبكرة CEEC، ومدونة اللغة الإنجليزية التاريخية الأمريكية Corpus of Historical American English (COHA)، ومدونة كتب اللغة الإنجليزية المبكرة على الإنترنت Early English Books Online (EEBO). تختلف هذه المدونات في أحجامها، فعائلة براون ومدونة هلسنكي

(١) صُممت مدونات عائلة براون في الأصل بوصفها مدونات تزامنية؛ لأن كلاً منها يتبع الإطار نفسه للعينة (انظر المبحث ٤، ١)، وأخذت كل عيناتها في نقاط زمنية مختلفة، وعليه فيمكننا استخدام عائلة براون بأكملها بوصفها مدونات تعاقبية. انظر الأمثلة في هذا المبحث وانظر بيكر أيضا (٢٠١١، ٢٠١٧).

ومدونة CEEC تتكون من بضعة ملايين من الكلمات على عكس مدونة EEBO التي بلغت مليارات الكلمات. وكما هو الحال مع أي بحث مدوني، نحتاج إلى إجراء تقييم نقدي لجودة وكمية الأدلة اللغوية المتاحة في المدونات التاريخية. تمثل البيانات التاريخية تحدياً إضافياً بسبب البعد التعاقبي الذي يجلب مصادر إضافية من التنوع. بالاعتماد على ليتش (Leech ٢٠١١) الذي يناقش تطور الأفعال الناقصة التي جرى تمثيل معلومة عنها في مهمة فكر^(١) ينبغي أن نراعي ثلاثة جوانب مهمة في اللغويات التاريخية القائمة على المدونة التي تحتاج إلى تقييم ثلاث ركائز فيها، الأولى: التمثيلية التعاقبية للمدونات، والثانية: تفسيرات بديلة للتطور اللغوي، والثالثة: تبدل معاني الصبغ اللغوية. ولننظر في هذه الثلاث واحدة تلو الأخرى.

أولاً، التمثيلية التعاقبية diachronic representativeness هي سمة من سمات المدونة التاريخية التي تعكس منهجياً المجموع اللغوي مع مرور الوقت. المدونة عينة مصممة من اللغة، وتحتوي عادة سمات مماثلة لسمات المجموع اللغوي الأكبر، وتكون في حالة من التقلب والتغير (المبحث ٤، ١). يتضمن هذا التغير ظهور كلمات جديدة أو خصائص نحوية جديدة، وقد تكون التمثيلية التعاقبية أهم من الحجم الكبير للمدونة. فعلى سبيل المثال، يوضح ليتش (٢٠١١) أنه لتتبع تطوير خاصية نحوية مثل الأفعال الناقصة، توفر عائلة مدونات براون التي تمثل مجموعة من الأوعية من اللغة المكتوبة صورة أدق عن تغير اللغة من مدونة أكبر بكثير قائمة على مصدر واحد مثل مجلة تايم TIME الأمريكية^(٢). من المحتمل جداً أن تعكس مدونة مجلة تايم نظراً لتمثيليتها المحدودة التغيرات المحلية التي تنطوي على ممارسات تحريرية مختلفة مع مرور الوقت. لا يمكن تعميم النتائج التي تستند إلى هذه المدونة دون مزيد من الأدلة في السماح بمناقشة التطور الشامل للغة الإنجليزية في القرن العشرين.

عند التفكير في التمثيلية التعاقبية نحتاج إلى التعامل مع بعض القيود المتأصلة في البيانات التاريخية. علينا أن ندرك أنه لم يُسجَل ويُحفظ من لغة الماضي سوى جزء صغير منها. المدونة التاريخية ليست عينة لغوية متوازنة من فترة تاريخية معينة؛ بل هي عدسة ضيقة توفر نظرة ثاقبة عن اللغة التي

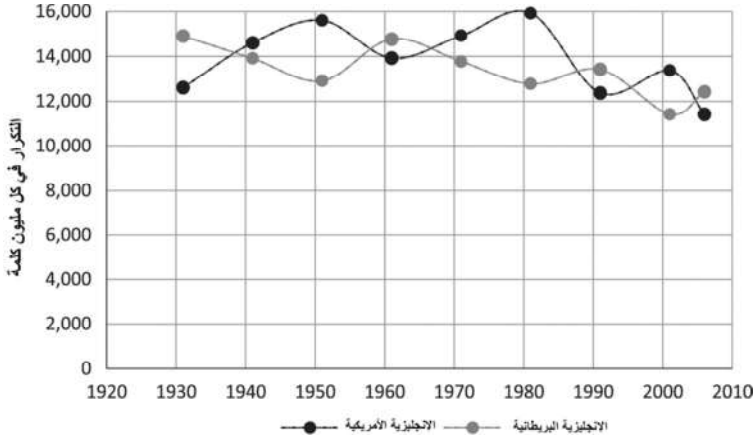
(٢) هذا المقال جزء من نقاش متعمق للغاية بين ليتش (٢٠٠٣، ٢٠١١) وميلر (Millar ٢٠٠٩) إذ طُرح عدد من الأسئلة المثيرة للاهتمام المتعلقة بتحليل المدونات وتغير اللغة.

(٣) تتكون مدونة مجلة تايم التي استخدمت للدراسة التعاقبية للغة الإنجليزية من 100 مليون كلمة من مجلة تايم وتتناول الفترات الزمنية بين 1923-2006، وهي متوفرة على الرابط الآتي: <http://corpus.byu.edu/time>.

حُفِظَت (McEnery & Baker 2017)، والمدونات التاريخية هي تقريباً كل مدونات اللغة المكتوبة، ولا يمكن الوصول إلى أي مدونة تاريخية لم تحط بالتسجيل أو التدوين. تعود التسجيلات الصوتية الأولى للغة المنطوقة إلى أواخر القرن التاسع عشر، وبعض الأنواع المكتوبة مثل الرسائل الشخصية تعكس الاستخدام شبه الرسمي وغير الرسمي للغة، ويمكن أن تخضع للتحليل اللغوي الاجتماعي التبايني الذي يُجرى عادةً مع البيانات المنطوقة (Nevalainen & Raumolin-Brunberg 2003: 26). والكتابة التي تُحَفَظ تمثل عادةً الكتب والنشرات والوثائق الرسمية التي كتبتها أقلية صغيرة من الأشخاص الذين عاشوا تلك الفترة، وعليه فإن عيناتنا اللغوية من البيانات التاريخية منحازة سواء من جهة الأنواع المقدمة أم التكوين الاجتماعي للمؤلفين (Nevalainen & Raumolin-Brunberg 2003). ومع الالتفات إلى هذه الإشكاليات في المدونات التاريخية تشير نيفالينين (Nevalainen 1999) إلى الهدف من التحليل اللغوي التاريخي، وهو السعي إلى تحقيق الفائدة وإن كانت البيانات رديئة أو منحازة.

ثانيًا، نحتاج إلى التفكير في تفسيرات بديلة للتطور اللغوي بناءً على الأدلة المتوفرة، ولو كانت الأدلة غير كافية أو تميل نحو وعاء معين من النصوص أو أوعية معينة من المتحدثين والكتاب فإن الأنماط التي نلاحظها ستكون غامضة. كلما رجعنا بالتاريخ وجدنا أدلة أقل، فعلى سبيل المثال، تتألف جميع النصوص المحفوظة من فترة اللغة الإنجليزية القديمة (450-1100) مما لا يزيد على ثلاثة ملايين كلمة جارية^(٤) أما من القرن السابع عشر فصاعدًا فتشتمل المدونات على مليارات الكلمات، ومنها مدونة (EEBO v.3) التي استخدمها مكنري وبيكر (٢٠١٧) والتي تتكون من أكثر من مليار كلمة للقرن السابع عشر وحده. في هذه الحالة لا ينبغي أن ننحرف إلى تقديم تفسيرنا بدلاً من الأدلة المفقودة، وإنما تقتضي الدقة فصل الأدلة (أي: البيانات) عن تفسيرها. بالعودة إلى مناقشة ليتش لننظر في مسألة نقطتي العينات التاريخية (1961 و1991) فيما يتعلق بقدرتنا على ملاحظة تغير اللغة بينها وتفسيره. في الشكل (١، ٧) ترتبط هاتان النقطتان التاريخيتان بخط مستقيم، وتؤكد نقطتا العينة الأخرى (1931 و2006) هذا الاتجاه. نحتاج إلى إدراك أن الخط المستقيم بين نقطتي البيانات هو بالفعل جزء من تفسير البيانات، الذي قد يكون أو لا يكون دقيقًا. كلما قلت نقاط العينات التاريخية التي لدينا اقترب هذا التفسير إلى التخمين. انظر في التفسير البديل في الشكل (٢، ٧) لنقاط البيانات نفسها من الشكل (١، ٧).

(٤) يعتمد عدد الكلمات هذا على المدونة الكاملة للإنجليزية القديمة The Complete Corpus of Old English!

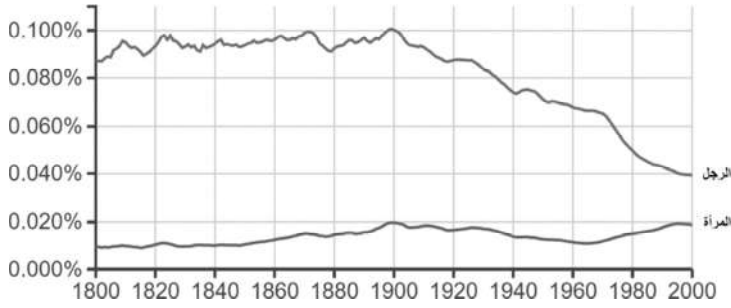


شكل (٧, ٢). تفسير بديل لنماذج في مدونات عائلة براون.

يمكن تخفيف المشكلة بإضافة المزيد من نقاط البيانات التي تتيح لنا تتبع مسار تغير اللغة بدقة. عندما يكون الحصول على المزيد من نقاط البيانات بهذه الطريقة غير ممكن ينبغي لنا ألا نمانع من تقبل تفسيرات بديلة. لتمثيل بعض التغيرات اللغوية، يكون الخط المستقيم (النموذج الأبسط) مناسباً. بالنسبة للعمليات التعاقبية الأخرى تعد النماذج الأكثر تعقيداً (المنحنيات المعقدة) أكثر ملاءمة. على سبيل المثال، عادةً ما تتبع التغيرات الصوتية والتغيرات النحوية منحني على شكل حرف (S) (S) (Nevalainen 2003: 53-5 & Raumolin-Brunberg 2003: 53-5 & Blythe & Croft 2012) في حين أن العديد من التغيرات المعجمية والخطابية تتبدل في صورة قمم وقعور (المبحث ٥, ٧).

ثالثاً، يعد تعدد معاني الصيغ اللغوية (الاشتراك اللفظي التعاقبي diachronic polysemy) ظاهرة ينبغي مراعاتها بعناية عند النظر إلى تطور اللغة. نحتاج إلى تقديم سرد عن التطور الدلالي، وليس مجرد لمحة عامة عن التكرارات المتغيرة للصيغ اللغوية. على سبيل المثال، يوضح ليتش (٢٠١١) أنه في كل من الإنجليزية البريطانية والأمريكية يُستخدم الفعل ناقص may أكثر بكثير بالمعنى المعرفي الدال على الاحتمالية كما في she may be right، ويُستخدم استخداماً أقل بكثير بالمعنى الوجوبي deontic الدال على التأدب في الاستئذان كما في please may I finish إذ يمكن أن تستخدم can عادة بدلاً من may. عند النظر إلى التطور التاريخي للأفعال الناقصة ينبغي أيضاً الإفصاح عن هذا التطور الدلالي لتقديم الصورة الكاملة.

تُميّز الركائز الثلاث التي نوقشت أعلاه (وهي: التمثيلية التعاقبية للمدونات، والتفسيرات البديلة للتطور اللغوي، وتبدل معاني الصيغ اللغوية) لغويات المدونة الحاسوبية عن الطرق الكمية الضعيفة لغوياً لـ (منهجيات البيانات الضخمة) في اللغة مثل الدراسة الرقمية الثقافية (Michel et al. 2011). عادةً ما تقدم منهجيات البيانات الضخمة تفسيرات شاملة للتطور التاريخي للكلمات والمفاهيم دون اهتمام يُذكر أو دون اهتمام أصلاً بالأساسيات والمبادئ اللغوية المهمة.

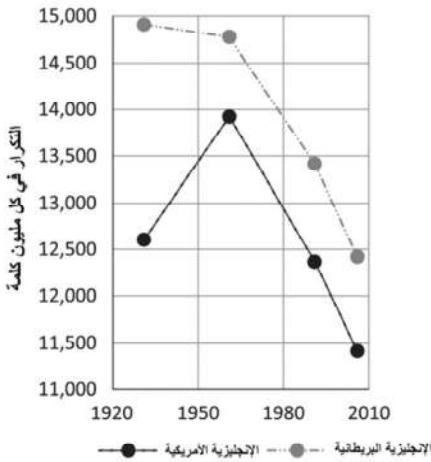
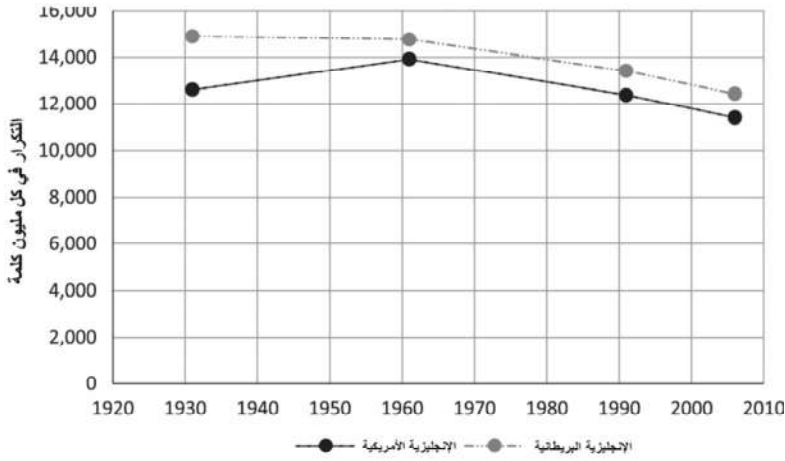


شكل (٣، ٧). مشاهد نغرافية n-gram من جوجل لكلمة (الرجل) وكلمة (المرأة).

من السهل جداً إنتاج رسم بياني للتغير اللغوي الاجتماعي الواضح لاستخدام كلمتي (الرجل) man و(المرأة) woman، ولكن من الصعب تفسير هذا الدليل إذا لم تتمكن من الوصول إلى الأمثلة الفعلية لاستخدام هذه الصيغ اللغوية والمعلومات الخاصة بالتكوين المتغير للمدونة. يُمكننا في الشكل (٣، ٧) أن نرى التكرارات المتغيرة لصيغة (الرجل) man في مقابل تكرارات صيغة (المرأة) woman في الفترة من 1800 إلى 2000. يمكننا أن نرى أن صيغة man تتناقص مع مرور الوقت، أما كلمة (المرأة) woman فتزداد، ويمكن تفسير ذلك على أنه سد لفجوة عدم المساواة بين الجنسين في الخطاب. قد يكون هذا التطور مرتبطاً بالتركيب المتغير للمدونة والتغيرات الثقافية والدلالية أو مجموعة من العوامل الأخرى المختلفة. نحن بعفوية لا نعرف، ويمكننا أن نضع مبدأً عاماً ينص على أن التحليل الذي لا يتعامل مع البيانات التعاقبية بطريقة تضمن الشفافية والثبات وقابلية المقارنة مع مرور الزمن فهو تحليل لا معنى له.

يمكننا أن نرى كيف تُضحي الدراسة الرقمية الثقافية والنُهج الماثلة بالتمثيلية والحساسية تجاه تبدل المعنى لتحصيل كمية كبيرة من الأدلة (ومن هنا تأتي منهجيات البيانات الضخمة)؛ ومع ذلك، هذا ليس جيداً أبداً (Leech 2011; Pechenick et al. 2015; McEnery & Baker 2017: section 1.3).

لنلق نظرة على بعض خيارات تمثيل البيانات باستخدام البيانات غير الرسمية. رأينا في الأشكال (٧، ١) و (٧، ٢) و (٧، ٣) تغير اللغة في شكل خط بياني line graph وهو عرض يسير يُمثّل المتغير الزمني على المحور السيني، ويُمثّل تكرارات المتغيرات اللغوية على المحور الصادي. تساعدنا الرسوم البيانية الخطية على تفسير أنماط التغير في المدونات. لاحظ أن الخط الفعلي في الشكل البياني ينتج بتوصيل نقاط البيانات توصيلاً دقيقاً، ولهذا السبب أعتد أولياً على هيئة نقط وخطوط بينها في الشكلين (٧، ١) و (٧، ٢).



شكل (٧، ٤). الأفعال الناقصة في مدونات عائلة براون: (القمة) الأصلية و(القاع) معاد قياسه.

تبرز نقاط البيانات بوصفها نقاط الأدلة، وعند التعامل مع الرسوم البيانية الخطية ينبغي أن ندرك أن التغير أو التقلب في البيانات التاريخية يمكن أن يبدو أقل أو أكثر حسب مقياس المحاور ونقطة البداية (الأصل) للرسم البياني. ولتوضيح ذلك، تمثل الرسوم البيانية في الصورة الأولى والصورة الثانية من الشكل (٤، ٧) البيانات نفسها مع التفسير نفسه، والفرق الوحيد بينهما يكمن في مقياس المحاور. لاحظ أيضًا أنه في الصورة الثانية لا يبدأ المحور الصادي من 0 ولكن من 11,000 ويصل إلى 15,000 بدلاً من 16,000.

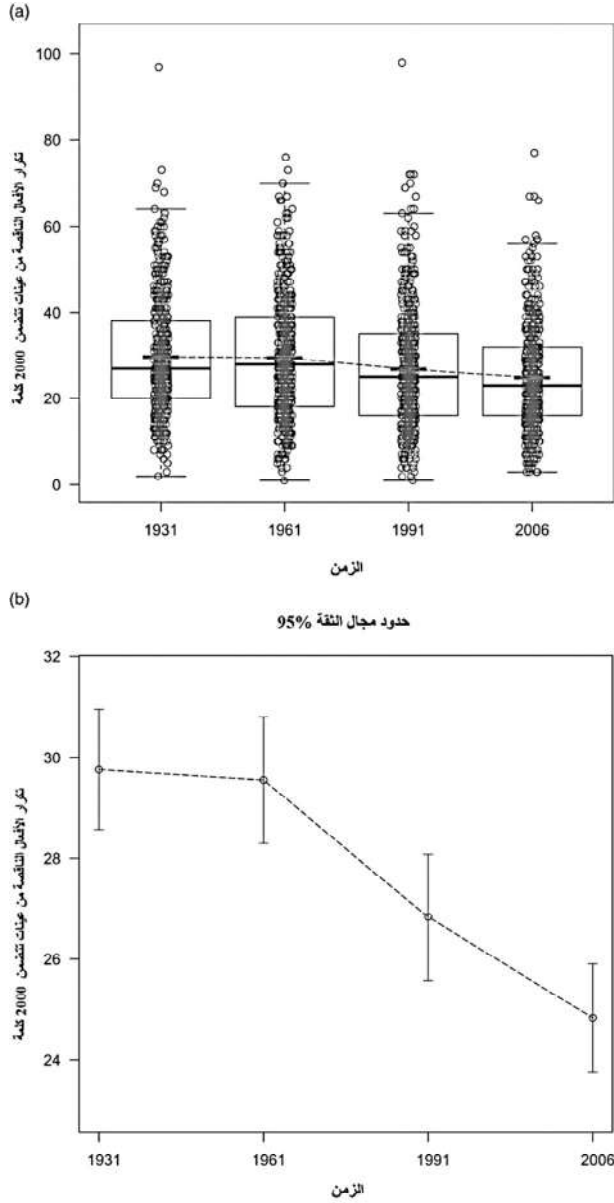
الهدف من المقارنة بين الصورتين هو ضرورة استعمال المقياس نفسه تمامًا عند مقارنة رسوم بيانية متعددة، ويساعد اعتماد الصورة الثانية على فهم التمثيل البياني وتفسيره تفسيرًا واضحًا. يوضح كليفلاند Cleveland (1994: 66ff.) أن الرسوم البيانية المصممة المثل يبلغ متوسط منحدرها ٤٥ درجة كيف لا تكون منبسطة أو متذبذبة أكثر من اللازم. علينا أيضًا أن نتحلّى بالموضوعية وألا نُزايِد في الشكل البياني على ما هو متمثل فعلاً في البيانات.

تشمل الخيارات الأخرى لتمثيل البيانات التعااقبية مخطط الرسم المربع، وأعمدة الخطأ، وخطوط المؤشر، ومخطط الشموع. نوقشت مخططات الرسم المربع وأعمدة الخطأ بالتفصيل في المبحث (٥، ١)، وهي معينة على توسيع النظر داخل كل نقطة بيانات في العينات التعااقبية، ويمكن استخدام أعمدة الخطأ لعرض مجالات الثقة ٩٥٪ عن القيم المتوسطة لكل فترة تاريخية. يعرض الشكل (٥، ٧) البيانات نفسها الممثلة في الخط الرقيق للإنجليزية البريطانية في الشكل (٤، ٧)، وبدلاً من تمثيل كل فترة تاريخية بقيمة واحدة من المتوسط يُظهر هذا الشكل التنوع الداخلي بين النصوص في كل فترة.

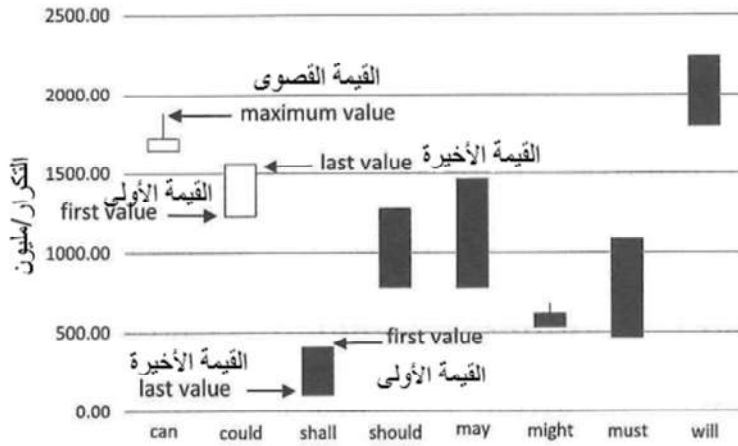
عند التعامل مع عدد كبير من المتغيرات التي تتطور مع مرور الوقت يمكننا استخدام خطوط المؤشر ملخصات لاتجاهات فردية متعددة. خط المؤشر sparkline هو رسم بياني صغير بحجم كلمة واحدة يمكن دمجها بسلاسة في النص (Tuft 2006: 46-63). بالنسبة للبيانات التعااقبية تكون خطوط المؤشر الخطية مفيدة جداً بوجه عام. يقدم المثال التالي خط مؤشر يوضح تطور الفعل الناقص must في القرن السابع عشر ويُمثل مئة نقطة عينات مختلفة:



يتميز استخدام must في القرن السابع عشر بمقدار كبير من التقلب يمكن أن يوفر خط المؤشر معلومات عن الحد الأدنى والحد الأقصى للقيمة (نقاط مصمتة فوق الخط وتحت) ليكون شكلاً غنياً بالمعلومات يشغل مساحة أصغر بكثير جداً من الوصف الكلامي. من بدائل خطوط المؤشر المتعددة مخطط الشموع الذي يُلخص تطور المتغير اللغوي في صورة مربع و(فتائل) تشبه الشموع. على الرغم من أنها تشبه سطحياً مخطط الرسم المربع يعمل مخطط الشموع بمبادئ مختلفة.



شكل (٥, ٧). الأفعال الناقصة في اللغة الإنجليزية البريطانية: (أ) مخطط رسم المربع (ب) أعمد الخطأ بمجال الثقة ٩٥٪.



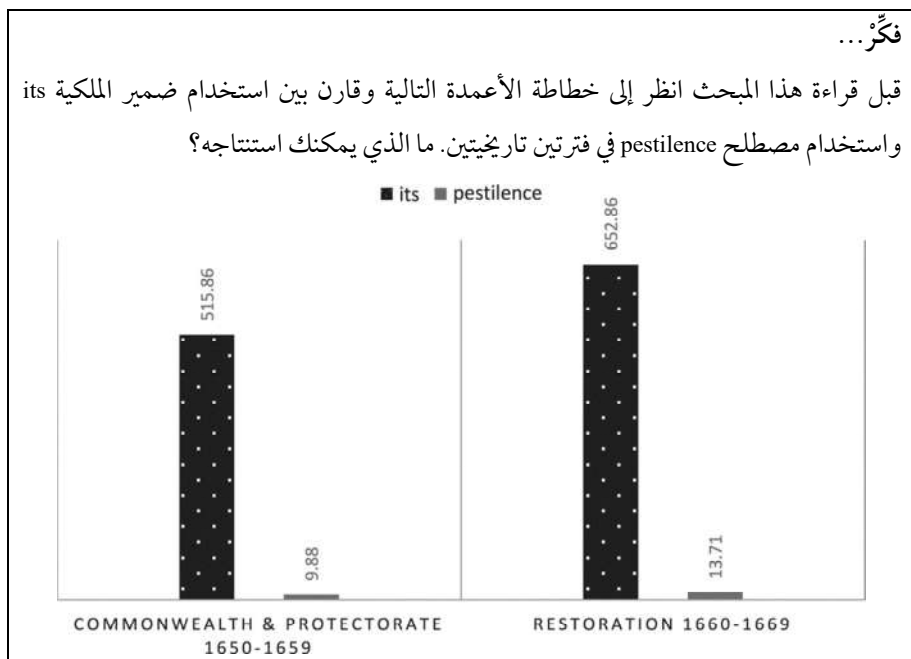
شكل (٦، ٧). مخطط الشموع: تطور الأفعال الناقصة الفردية بين ١٩٣١-٢٠٠٦م.

[ترجمة]: يستطيع can، قد يكون could، سوف shall، ينبغي should، قد amay، قد might، يجب must، سوف will

يُستخدم مخطط الشموع في الأغلب في التقارير المالية لتتبع تطور أسعار الأسهم مع مرور الوقت. يوضح الشكل (٦، ٧) مثالاً على مخطط شموع يشير إلى تطور أفعال ناقصة مختلفة في الفترة ١٩٣١-٢٠٠٦. يُظهر كل (مخطط شموع) النقطة الأولية في السلسلة الزمنية والنقطة النهائية والحد الأدنى والحد الأقصى للقيمة. يُشار إلى القيم الأولية والنهائية بالمربع، وفي حالة الاختلاف عن القيم الأولية والنهائية يُشار إلى الحد الأدنى والحد الأقصى للقيم بفتائل تبرز خارج المربع. ويظهر لون المربع المعبأ مقابل غير المعبأ اتجاه التغير، إذ يدل المربع المعبأ على التناقص ويدل المربع غير المعبأ على التزايد. في الشكل (٦، ٧) يمكننا أن نرى مباشرة أنه في هذه الفترة قد زاد تكرار فعلين ناقصين من أصل ثمانية كما في المربع غير المعبأ، أما في البقية فقد قل التكرار. وما عدا حالتين تكون فيهما القيمة القصوى خارج الصندوق كانت الزيادة والنقصان سلسة وبدون تقلبات في التكرار.

يُمثل تحليل بيانات اللغة التاريخية تحدياً له متطلبات خاصة مرتبطة بالبعد التاريخي. يُعدّ التمثيل الفعال باستخدام المخططات الخطية، ومخططات رسم المربع، وأعمدة الخطأ، وخطوط المؤشر، ومخططات الشموع نقطة انطلاق مفيدة لأي بحث تاريخي.

(٧, ٣) إيجاد الاختلافات وتفسيرها: تغيّر النسبة المئوية واختبار التمهيد



أهم سؤال ينبغي أن نتناوله عند مقارنة مدونتين مختلفتين هو ما إذا كانت تكرارات المتغيرات ذات الصلة أكبر أو أصغر بشكل ملحوظ في المدونة الأولى مقارنة بالمدونة الثانية. عند التعامل مع المقارنات التعاقبية نحتاج إلى تقييم الاختلافات الملحوظة في التكرارات وارتباطها بالتغير في الخطاب أو النص مع مرور الوقت أو ارتباطها بمصادر أخرى للتنوع. أوضحت غابلاسوفا وآخرون (Gablasova et al. 2017a) أن عينات المدونات من الوعاء نفسه من اللغة تختلف في الأغلب اختلافاً كبيراً، ويتطلب هذا الاختلاف الحاجة إلى توخي الحذر الشديد بشأن كيفية تفسير الاختلافات. ينبغي توخي المزيد من الحذر عند التعامل مع المدونات التاريخية، وذلك بأخذ عينات سانحة (راجع مشكلة البيانات الرديئة التي نوقشت في المبحث ٧, ٢). لنلق نظرة على بعض الأمثلة من مدونة EEBO v.3 والطرق الممكنة لتحليل بياناتها وتفسيرها. يسرد الجدول (٧, ١) أربعة متغيرات لغوية (its و must و time(s) و pestilence)، ويقدم مقارنة بين نقطتي العينات في مدونة

EEBO، تعكس فترتين تاريخيتين في القرن السابع عشر. الأولى: الكومنولث والمحميات (الحكومات الخارجة عن الحكم الملكي كويلز وكورنول إلخ) والثانية: الاسترداد. اعترى النصف الثاني من القرن السابع عشر اضطراب سياسي كبير والعديد من التغيرات الاجتماعية التي أثرت على الخطاب واللغة نفسها.

جدول (١، ٧). مقارنة فترتين في مدونة EEBO: الكومنولث والمحمية Commonwealth & Protectorate والاسترداد Restoration.

المتغير اللغوي	المدونة الأولى الكومنولث والمحمية (١٦٥٠-٩) حجم المدونة: 168912439	المدونة الثانية الاسترداد (١٦٦٠-٩) حجم المدونة: 111998646	زيادة/ نقصان النسبة المئوية
its	*515.86	652.86	+27
must	1,173.02	1,135.67	-3
time(s)	1,445.57	1,355.84	-6
pestilence	9.88	13.71	+39

*التكرارات النسبية لكل مليون

لنتمكن من رؤية الفرق بين نقطتي العينات الزمنية حُسبت زيادة/ نقصان النسبة المئوية البسيطة، وهما إحصائية دالة على عدد النقاط المئوية التي تزيد أو تنقص بها قيمة المتغير اللغوي المعين بين فترتين زمنيةتين، وتُحسب باستخدام المعادلة الآتية:

$$(١، ٧) \quad \text{زيادة/ نقصان النسبة المئوية} = \frac{\text{التكرار النسبي في المدونة الثانية} - \text{التكرار النسبي في المدونة الأولى}}{\text{التكرار النسبي في المدونة الأولى}} \times 100$$

حيث المدونة الأولى هي مدونة من فترة سابقة والمدونة الثانية هي مدونة من فترة لاحقة. يُشار في الجدول (١، ٧) إلى زيادة أو نقصان النسبة المئوية للمتغيرات اللغوية الأربعة في العمود الأخير. يوضح الحساب التالي كيفية استخراج النسبة المئوية لزيادة أو نقصان النسبة المئوية للعنصر الأول في الجدول لضمير الملكية its.

$$(٢، ٧) \quad \text{زيادة/ نقصان النسبة المئوية للضمير} = \frac{515.86 - 652.86}{515.86} \times 100 = 26.6\%$$

ويوضح الجدول أيضًا حالتين من حالات الزيادة وحالتين من حالات النقصان. في حالتين (its و pestilence) تبدو الزيادة كبيرة إلى حد ما (٢٧٪ و ٣٩٪ على التوالي). لتكون قادرًا على تحديد كون حالات الزيادة والنقصان عائدة إلى الصدفة (أو خطأ في أخذ العينات) وتحديد أيها ذات دلالة إحصائية يمكننا استخدام اختبار إحصائي تقيمي. على الرغم من أن اختبار كاي تربيع والاحتمالية اللوغارتمية يُستخدمان في الأغلب في هذه الحالة فهما ليسا مناسبين لهذا النوع من المقارنة (Lijffijt et al. 2012, 2016; Brezina & Meyerhoff 2014)، والبديل عنهما يكون باستخدام اختبار يراعي توزيع المتغيرات اللغوية في النصوص الفردية في كلتا المدونتين (انظر المبحث ٤، ١ عن تصميم عينة النص أو المتحدث الفردي).

corpus	Text A Freq: 10	Text B Freq: 5	Text C Freq: 15	Text D Freq: 1	Text E Freq: 20	mean = 10.2
resampling cycle 1	Text A Freq: 10	Text A Freq: 10	Text B Freq: 5	Text D Freq: 1	Text E Freq: 20	mean ₁ = 9.2
resampling cycle 2	Text A Freq: 10	Text A Freq: 10	Text A Freq: 10	Text B Freq: 5	Text B Freq: 5	mean ₂ = 8.0
[...]						
resampling cycle n	Text C Freq: 15	Text D Freq: 1	Text D Freq: 1	Text E Freq: 20	Text E Freq: 20	mean _n = 11.4

ترجمة الشكل: النص Text، التكرار Freq، المدونة Corpus، المتوسط Mean، تدوير إعادة تدوير العينة resampling cycle. شكل (٧، ٧). التمهيد: وصف المفهوم.

تشمل الخيارات التي لدينا للاختبار اختبار-ت واختبار مان-ويتني U (المبحث ٣، ٦) واختبار التمهيد bootstrap test. تعد أساليب التمهيد Bootstrapping إضافة جديدة نسبياً إلى مربع الأدوات الإحصائي لأنها تتطلب قوة حسابية كافية لم تكن متوفرة سابقاً (Efron 1979)، وتقدم طرقاً قوية لتقييم التباين وتقدير الدلالة الإحصائية (Chernick & LaBudde 1994 Efron & Tibshirani 2014). التمهيد Bootstrapping هو عملية إعادة أخذ عينات متعددة، وتحدث عادةً آلاف المرات مع استبدال البيانات، وهذا يعني أننا نأخذ عينة عشوائية من النصوص من مدونة بحيث يمكن أن يرد

كل نص عدة مرات في العينة لأننا نستبدلها (أي: نعيدها إلى المجموع مرة أخرى) بمجرد أخذها. في كل دورة من دورات إعادة أخذ العينة نلاحظ انخفاض قيمة الإحصاء المعني (على سبيل المثال، متوسط المتغير اللغوي). هذا يعطي نظرة ثاقبة على مقدار التنوع في البيانات ويعطينا الثقة للتعميم بناءً على هذه العينة. لنأخذ، على سبيل المثال، مدونة تتكون من خمسة نصوص A و B و C و D و E ولكل نص تكرار معين للمتغير اللغوي المعني (الشكل ٧، ٧).

يشير (التمهيد bootstrapping) إلى التعبير الاصطلاحي: (ينشد المرء نفسه من خلال مهبط الغنائم (pull oneself up by one's bootstraps)، ويدل في موضوعنا على طبيعة المساعدة الذاتية لخوارزمية التمهيد (Efron 1979: 465) التي تستخدم البيانات الموجودة في العينة نفسها عدة مرات لإنشاء تقدير للقيمة المعنية في المجموع اللغوي.

سنركز في اختبار التمهيد المقترح في دراسة ليجفيجت وآخرين (Lijffijt 2016) لكونه اختباراً غير بارامترى للدلالة الإحصائية ولا يوجد فيه افتراض عن توزيع المتغير اللغوي في المجموع اللغوي، ويمكن استخدامه حتى مع توزيعات منحرفة للغاية. يقارن هذا الاختبار اثنين من المدونات ويحسب قيمة الاحتمال المرتبط بالمقارنة. يعمل الاختبار بمبدأ التمهيد بعد إعادة أخذ عينات متعددة مع الاستبدال للمدونتين الأولى والثانية. يقدم ليجفيجت المعادلة التالية للاختبار:

$$p = \frac{(2 + \text{عدد دورات التمهيد}) \times (p_1 \text{ أو } 1 - p_1, \text{ أيها أصغر})}{1 + \text{عدد دورات التمهيد}} \quad (٣, ٧)$$

وتدل p_1 على ناتج: $\frac{\text{مجموع قيم } H \text{ لجميع دورات التمهيد}}{\text{عدد دورات التمهيد}}$

ويمكن أن تكون (H) القيمة (1) أو (0.5) أو (0) ودلالة كل قيمة هي على النحو الآتي:

$H = 1$ إذا كانت القيمة المعنية في المدونة المدوّرة الأولى أكبر من القيمة المعنية في المدونة المدوّرة الثانية

$H = 0.5$ إذا كانت القيمة المعنية في المدونة المدوّرة الثانية تساوي القيمة المعنية في المدونة المدوّرة الأولى

$H = 0$ إذا كانت القيمة المعنية في المدونة المدوّرة الثانية أكبر من القيمة المعنية في المدونة المدوّرة الأولى

الفكرة الكامنة وراء الاختبار سهلة للغاية. نحن نقارن بعدد كبير من تدوير التمهيد بين

المدونة الأولى المعاد أخذ عيناتها وبين المدونة الثانية المعاد أخذ عيناتها، ثم نبحث عن اختلاف ثابت

بين المدونات المعاد أخذ عيناتها. يؤدي ذلك إلى انخفاض قيمة الاحتمال (الدلالة الإحصائية). تظهر قيمة الاحتمال المنخفضة إذا كانت المدونة الأولى في جميع الحالات أو معظمها إما أكبر أو أصغر من المدونة الثانية. لكي يُظهر الاختبار نتائج موثوقة يلزم السماح بتدويرات تمهيد كافية تزيد على الـ 1000 مرة.

في الواقع العملي وعند إجراء اختبار التمهيد نحتاج إلى تتبع توزيع المتغير اللغوي في النصوص الفردية ومعيّرة التكرارات؛ أي: نحن بحاجة إلى الحصول على تكرارات نسبية (المبحث ٢، ٣). يرد في الشكل (٧، ٨) مثال على مجموعة البيانات هذه التي تعرض الحالات العشر الأولى فقط من بين آلاف الحالات. تُدخل هذه البيانات أولاً في المعادلة (٧، ٣)، والنتيجة هي قيمة الاحتمال التي تظهر مؤشراً على الدلالة الإحصائية للمقارنة.

يعرض الجدول (٧، ٢) نتائج قيم الاحتمال لاختبار التمهيد للمقارنة من الجدول (٧، ١) بعد أن تتبعنا تطور أربعة متغيرات لغوية في القرن السابع عشر. يمكننا الآن أن نستنتج أن ثلاثة من المقارنات ذات دلالة إحصائية، وهذا يعني أن قيمة الاحتمال أصغر من 0.05. يمكننا إضافةً إلى قيمة الاحتمال سرّ حجم التغير (أي: حجم التأثير)، ويمكن القيام بذلك بحساب قياس كوهن-د أو كوهن-د الفعّال (المبحثان ٦، ٣ و ٨، ٤) وتوفير مجالات الثقة لهذا المقياس. يمكن أيضاً تقدير كوهن-د أو أحجام التأثير الأخرى باستخدام تقنية التمهيد^(٥).

ID	1650_59	1660_69
1	0	0
2	662.01	0
3	191.36	0
4	1625.28	0
5	475.62	0
6	338.29	0
7	326.26	377.42
8	110.05	3062.48
9	0	1059.2
10	0	1236.4

شكل (٧، ٨). مثال على مجموعة بيانات اختبار التمهيد لـ: its في مدونة EEBO.

(٥) لاحظ أن التأثيرات الواردة في الجدول (٧، ٢) أصغر بكثير من التأثيرات القائمة على النسبة المئوية الإجمالية للتغير الواردة في الجدول (٧، ١) لأن تقنية التمهيد على عكس النسبة المئوية للتغير الكلي، وتراعي الاختلاف الفردي.

جدول (٢، ٧). نتائج التمهيد لمقارنة فترتين في مدونة EEBO.

المتغير اللغوي	قيمة الاحتمال بناء على اختبار التمهيد (عينات من ١٠ آلاف كلمة)	دال؟ (<0.05)	حجم التأثير: [مجال الثقة] كوهن-د
its	0.0001	نعم	تأثير أقل: [0.07, 0.15] 0.11
must	0.0061	نعم	تأثير أقل: [0.02, 0.1] 0.06
time(s)	0.8070	لا	أقل تأثير، مجال الثقة يشمل الصفر: [0.04, -0.05] 0.01
pestilence	0.0001	نعم	أقل تأثير [0.07, 0.15] 0.11

أخيرًا نحتاج إلى تحديد الاختلافات الدالة إحصائيًا بين المدونتين اللتين تمثلان الفترتين (١٩٥٠-١٩٥٩) و(١٦٦٠-١٦٦٩). يمكن القيام بذلك من خلال التقييم النقدي للمدونة من جهة تمثيليتها وقابليتها للمقارنة أولاً، وللتائج المتعلقة بمصادر البيانات الأخرى أو الأبحاث السابقة أو إحداهما ثانيًا، وللتائج المتعاقبة فيما بعد إن وجدت ثالثًا. على سبيل المثال، يقيم الجدول (٣، ٧) باختصار هذه الجوانب الثلاثة لنتائج التحليلات الإحصائية بالجدول (٢، ٧).

التقرير الإحصائي: اختبار التمهيد bootstrapping

١- ما يجب الإفصاح عنه

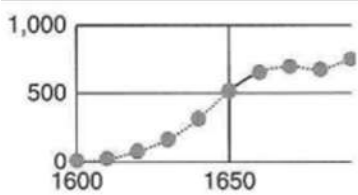
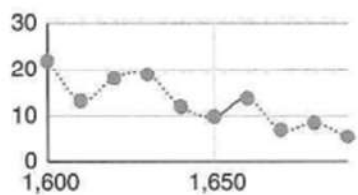
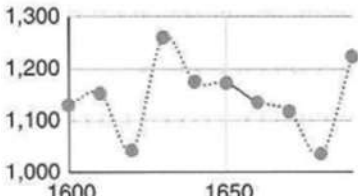
النتائج الوحيدة لاختبار التمهيد هو قيمة الاحتمال التقريبية التي تعتمد على عدد العينات. ينبغي الإفصاح عن قيمة الاحتمال وعدد العينات، كما ينبغي الإفصاح عن قيمة الاحتمال الأكبر من 0.001 بوصفها قيمة دقيقة وعن حجم التأثير (كوهن-د) ومجال الثقة لحجم التأثير.

٢- مثال الإفصاح في التقرير

• استخدم اختبار التمهيد (Lijffijt et al. 2016) لمقارنة المدونات التي تمثل نقطتي أخذ العينات التعااقبيتين (بين الأعوام ١٦٥٠-١٦٥٩ و ١٦٦٠-١٦٦٩). كان الاختلاف في استخدام الضمير its

ذا دلالة إحصائية وبقيمة احتمال أكبر من 0.001 < وبـ ١٠ آلاف تدويرٍ تمهيديٍّ. كان حجم التأثير (كوهن-د) صغيراً جداً: 0.11 و مجال الثقة ٩٥٪ [0.07, 0.15].

جدول (٣، ٧). التقييم النهائي للنتائج: its و must و pestilence.

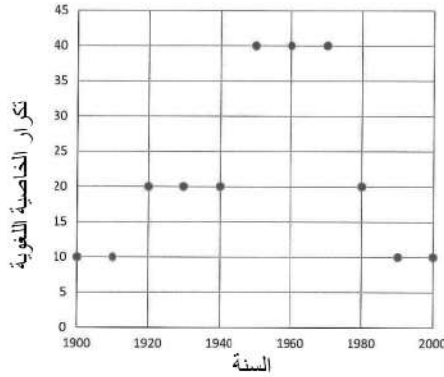
المتغير اللغوي	(1) تمثيل المدونة وقابليتها للمقارنة	(2) مصادر البيانات الأخرى	(3) صورة تعاقبية أكبر
its	تأخذ مدونة EEBO عيناتها من الكتب المطبوعة. لذلك	أُبلغ عن زيادة في استخدام its في القرن السابع عشر في الأدبيات المستندة إلى مدونات أخرى (على سبيل المثال Nevalainen & Raumolin-Brunberg (2003)	
must	فهي تعكس ممارسة النشر في فترة تاريخية معينة (على سبيل المثال أنشئت بعض النصوص في وقت أبكر من طباعتها، أو طبعت في مراحل لاحقة مع	لم يُفصح عن التغير في تكرار الأفعال الناقصة في القرن السابع عشر	
pestilence	تعديلات)؛ يحتاج التقلب في تكوين النوع أيضاً إلى النظر فيه	ربما تعكس الزيادة المؤقتة في ستينيات القرن السابع عشر Great Plague لـ ١٦٦٥-١٦٦٦ ينبغي البحث في المنافسة مع مصطلحات التسمية الأخرى مثل plague و black death	

(٧, ٤) تجميع الفترات الزمنية: تحليل التجميع المجاور

فكّر...

قبل قراءة هذا المبحث فكر في السؤال التالي:

كيف يمكنك تجميع نقاط البيانات في الشكل (٧, ٩) وفقاً لتشابهها؟

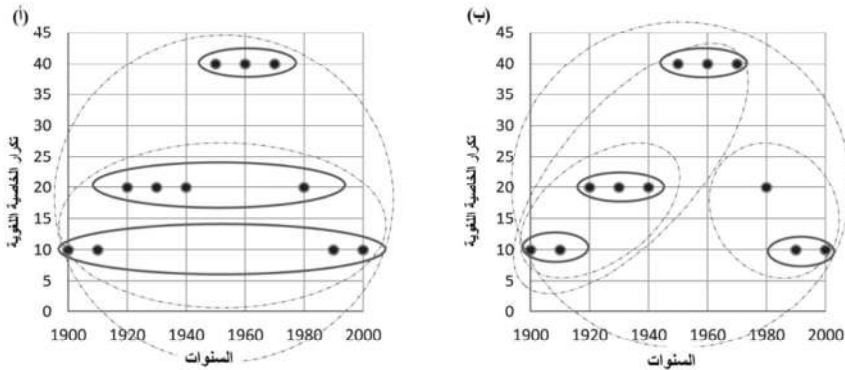


شكل (٧, ٩). نقاط البيانات مع مرور الوقت: مثال مبتكر.

شُرحت المبادئ العامة لتحليل التجميع في المبحث (٥, ٣) الذي يقارن المسافات بين نقاط البيانات الفردية ويجمع نقاط البيانات الأقرب معاً لإنشاء نقطة بيانات جديدة. تُكرر هذه العملية حتى تنضم كل نقطة بيانات إلى تجميع كبير واحد. تساعدنا هذه العملية على تحديد بنية البيانات بناءً على تشابه النصوص والمتحدثين أو الأنواع على سبيل المثال. في التحليلات التاريخية نضيف بعداً إضافياً إلى تحليل التجميع ألا وهو الوقت. يوفر الوقت طبقة إضافية من البنية لأنه يرتب بياناتنا وفقاً للتسلسل الزمني للأحداث. يمكننا أن نقرر ما إذا كنا نريد تضمين التسلسل الزمني بوصفه مبدأً منظماً في تحليل التجميع. إذا كان سؤال البحث عاماً على النحو الآتي: (ما الفترات التاريخية المتشابهة في استخدام المتغيرات اللغوية المستهدفة؟) يمكننا تطبيق تحليل التجميع المتكامل الهرمي البسيط (المبحث ٥, ٣). أما إذا أردنا تحديد فترات زمنية أكبر متصلة بناءً على تشابه استخدام المتغير اللغوي المستهدف نحتاج إلى النظر في إجراء تجميعي يحترم ترتيب الوقت لنقاط البيانات الفردية. في هذه الحالة يمكننا استخدام إجراء اقترحه غريس وهيلبرت (Gries & Hilpert 2008, 2010) يسمى تجميع الجوار القائم على المتغيرة variability-based neighbor clustering VNC. يُجسب هذا التجميع بين نقاط البيانات المجاورة زمنياً (على سبيل المثال

السنوات الفردية التي تتبع بعضها بعضاً) ثم تدمج أكثرها تشابهاً. يُكرر هذا حتى تُدمج جميع نقاط البيانات في تجميع واحد. يوضح الشكل (٧, ١٠) الفرق بين تجميع التكتلات الهرمية وتجميع الجوار القائم على المتغيرة. يمكن تجميع نقاط بيانات مجموعة البيانات المقدمة في مهمة (فكر) معاً بناءً على مجرد تشابهها بغض النظر عن الترتيب الزمني في الجزء الأيسر أو دمجها بوصفها نقاط بيانات مجاورة في الجزء الأيمن. فعلى سبيل المثال، يجمع إجراء تجميع التكتلات الهرمية معاً الأعوام ١٩٠٠ و ١٩١٠ و ١٩٩٠ و ٢٠٠٠ لأن لها تكرار المتغير اللغوي المستهدف نفسه (10). أما تجميع الجوار القائم على المتغيرة فتأخذ السنوات الأربع وتدمج منها فقط ١٩٠٠ و ١٩١٠ في مجموعة واحدة و ١٩٩٠ و ٢٠٠٠ في مجموعة أخرى وفقاً لمبدأ عمليات الدمج المجاورة. يمكن الاطلاع على المخطط التفرعي الناتج عن كلا الإجراءين في الشكل (٧, ١١) ويتضمن تمثيلات بيانية بديلة للتجميعات من الشكل (٧, ١٠).

في تجميع الجوار القائم على المتغيرة كما هو الحال في تحليل التجميع التكتل الهرمي، يلزم اتخاذ قراراتين رئيسيين عن الإعدادات المستخدمة من الناحية الإجرائية، وهما: قياس المسافة (أو التشابه) وطريقة دمج نقطتي بيانات (قاعدة الدمج). هناك مجموعة من قياسات المسافة التي يمكن استخدامها، وقد نوقشت المسافة الإقليدية أو مسافة مانهاتن أو كانبرا في المبحث (٣, ٥). يمكننا قياس التشابه بمعنى أوسع باستخدام معامل ارتباط بيرسون أو سيرمان (المبحث ٢, ٥)، ويكون هذا الخيار مفيداً إذا لم نكن مهتمين بالاختلافات الدقيقة بين نقاط البيانات، ولكننا نرغب في إيجاد الاتجاهات المتشابهة في مجموعة البيانات. عند استخدام الارتباط تقاس المسافة بناقص واحد (1-):
 $1-r$ أو $1-r_s$ لأن الارتباط يكون مرتفعاً وقريباً من (1) عندما يكون التشابه مرتفعاً.



شكل (٧, ١٠). مبدأ التجميع: (أ) تجميع التكتل الهرمي، و(ب) تجميع الجوار القائم على المتغيرة.

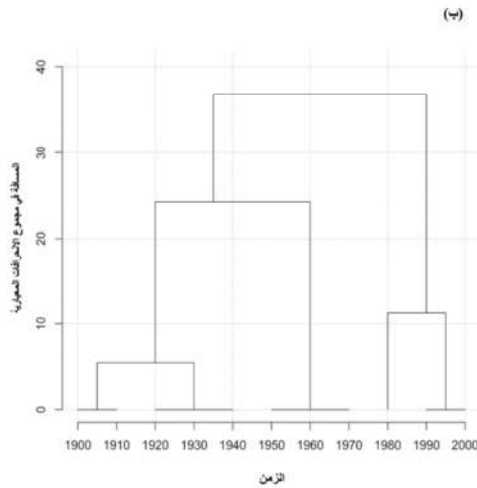
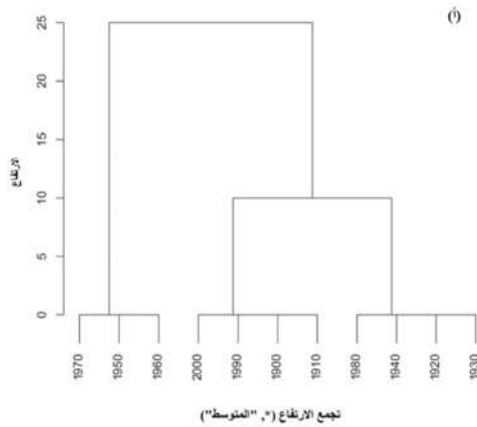
يمكننا أيضًا استخدام الانحراف المعياري أو معامل التنوع (المبحث ٤, ٢) مقياسًا للمسافة، ويُعبّر هذان المقياسان عن مسافة القيم الفردية من المتوسط. ويُعبّر معامل التنوع تعبيرًا محددًا عن هذه المسافة بوحدات قياسية، تجعل القيم المعروضة على المحور الصادي للمخطط التفرّعي قابلةً للمقارنة عبر المتغيرات اللغوية المختلفة. تعتمد قاعدة الدمج على ما نعدّه القيمة الجديدة لنقاط البيانات المدججة، وتشمل الخيارات الربط المنفرد single linkage، والربط التام complete linkage، والربط المتوسط average linkage، وطريقة وارد Ward's (المبحث ٣, ٥)، وينبغي أن يكون الدافع وراء القرار المتعلق بهذه الإعدادات هو هدف التحليل وهدف السؤال البحثي؛ لأنه يؤثر مباشرةً في نتيجة تحليل التجميع وشكل المخطط التفرّعي. تدعو الحاجة في لغويات المدونة الحاسوبية إلى مزيد من التحقق من موثوقية التقييم التجريبي في إجراءات الإعدادات المختلفة في التحليل. وباستخدام تجميع الجوار القائم على المتغيريّة استخدم غريس وهيلبرت (Gries & Hilpert 2008, 2010 و Hilpert & Gries 2009 و Hilpert 2011) الخيارات الآتية:

الأول: الانحراف المعياري، ومعامل التنوع، وارتباط بيرسون، ونسبة المتوسطات المصححة^(٦).

الثاني: الربط المتوسط.

ويمكن استخدام تجميع الجوار القائم على المتغيريّة لتصنيف نقاط البيانات الزمنية الفردية (مثل السنوات والعقود) المتشابهة في الاستخدام اللغوي، ويمكننا أن ننظر إلى التغير النحوي أو المعجمي وكيفية تشتته على مر الزمن. فعلى سبيل المثال، يمكننا أن نلاحظ ظهور ضمير الملكية its في القرن السابع عشر وتصنيف مراحل ظهوره.

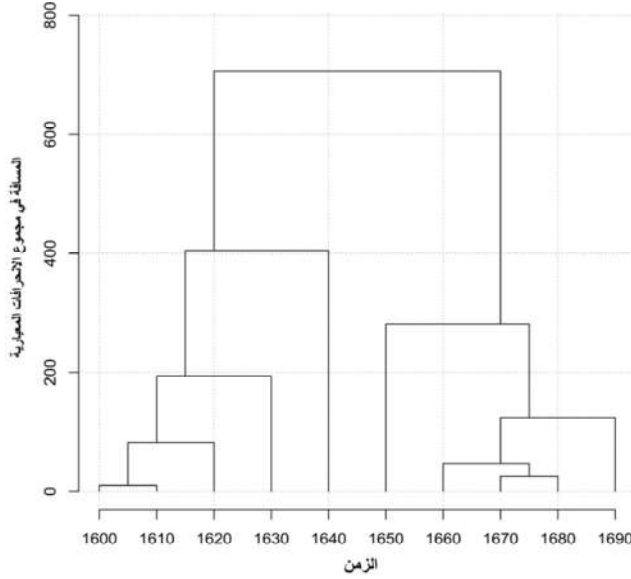
(٦) يمكن الاطلاع على تفاصيل نسبة المتوسطات المصححة، بما في ذلك المعادلة، من دراسة غريس وهيلبرت (2008: 73).



شكل (١١، ٧). المخططات التفرعية: (أ) تجميع التكتل الهرمي (ب) تجميع الجوار القائم على المتغيرة.

جدول (٤، ٧). التكرار النسبي (لكل مليون كلمة) لضمير الملكية its في القرن السابع عشر.

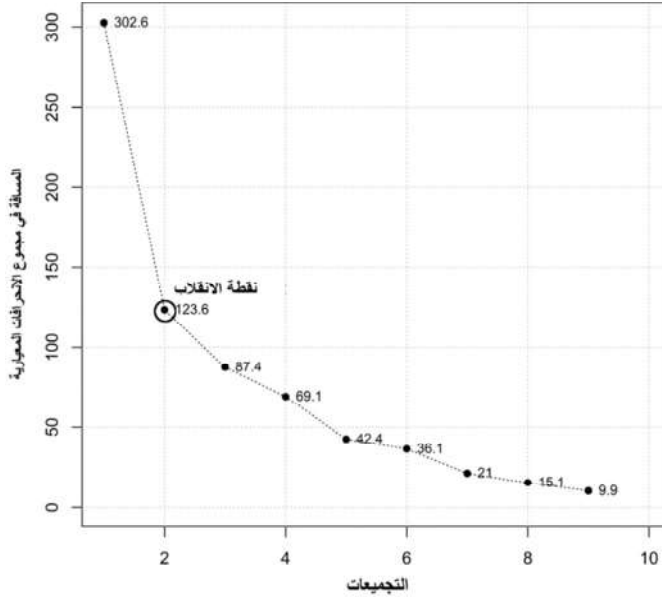
العقد	1600s	1610s	1620s	1630s	1640s	1650s	1660s	1670s	1680s	1690s
التكرار النسبي	9.39	23.4	77.71	161.93	309.79	515.86	652.86	694.93	673.6	751.37



شكل (١٢, ٧). مخطط تفرّعي لاستخدام ضمير الملكية its في القرن السابع عشر.

بدأ استخدام هذا الضمير بوصفه نسخة محايدة من his الذي كان يستخدم فيما سبق لكل من المذكور والمحايد، وتنافس هذا الضمير مع صياغات مثل thereof and of it (Nevalainen & Raumolin- 62: 2003). يوفر الجدول (٤, ٧) التكرارات النسبية للضمير its في القرن السابع عشر مقيسًا بالعقد.

يوضح المخطط التفرّعي المبني على الجدول (٤, ٧) انقسامًا واضحًا بين فترتين كبيرتين من مطلع القرن السابع عشر إلى أربعينياته ومن مطلع خمسينياته إلى تسعينياته (الشكل ١٢, ٧). إذا أردنا تقسيم هذه الفترات الكبيرة إلى أقسام أصغر فعلينا أن نقرر أين تكمن الاختلافات الرئيسية. في المخطط التفرّعي يُعبّر عن الفرق بين التجميعات على أنه الارتفاع الذي يوضحه الخط العمودي في المخطط. ولأن المقارنة بين الارتفاعات في الشكل (١٢, ٧) قد تكون صعبة، خاصةً مع نقاط بيانات متعددة، فإنه يمكن أن يساعدنا المخطط الركامي الشبيه إلى حد ما بالمخطط المستخدم على تحليل العوامل (المبحث ٤, ٥) في تقرير التقسيمات الفرعية للتجميعات الرئيسية. ويعرض المخطط الركامي عدد التجميعات على المحور السيني ويعرض الارتفاع (المسافة بين التجميعات) على المحور الصادي.



شكل (١٣، ٧). مخطط تفرعي لاستخدام ضمير الملكية its في القرن السابع عشر.

يشبه المخطط الركامي شكلاً مشابهاً لمنحدر جبل به أحجار صغيرة (نقاط بيانات) تغطي المنحدر، ومن هنا جاءت عبارة (المخطط الركامي). يشير الانحدار الحاد إلى اختلافات كبيرة بين مجموعات التجميعات، أما الشكل المستوي فيشير إلى اختلافات صغيرة بين مجموعات التجميع. في الشكل (١٣، ٧) الذي يعرض المخطط الركامي المتعلق بالمخطط التفرعي في الشكل (١٢، ٧) يمكننا أن نرى أنه بعد تجميعين كبيرين يبدأ المخطط الركامي في الاستواء، ونسمي هذا الموضع نقطة انقلاب. باستخدام مخطط ركامي يمكننا التأكد أن التقسيم الفرعي الرئيسي في المخطط التفرعي هو بين تجميعين يمتد الأول بين ١٦٠٠-١٦٤٠ والثاني بين ١٦٥٠-١٦٩٠.

التقرير الإحصائي: تجميع الجوار القائم على المتغيرة

١- ما يجب الإفصاح عنه

تجميع الجوار القائم على المتغيرة هو إلى حد كبير وسيلة بصرية استكشافية لإظهار التجزئة التعاقبية للبيانات. يُستخدم بارامتران لتحديد التجميع، وهما: قياس المسافة وقاعدة الدمج، مع ضرورة

إضافة نتائج التحليل والمخطط التفرّعي. والسؤال الرئيسي الذي ينبغي تناوله هو: ما عدد التجميعات ذات المغزى التي تمثل فترات تاريخية مستمرة يمكن ملاحظتها في المخطط؟ يمكن استخدام مخطط ركامي للمساعدة على تحديد إجابة هذا السؤال.

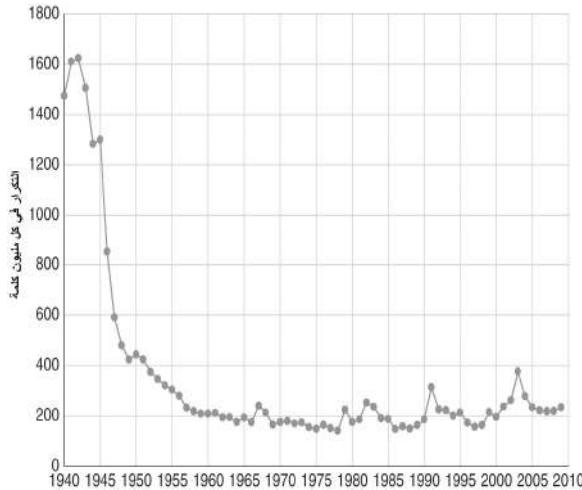
٢- مثال الإفصاح في التقرير

حُلّت البيانات باستخدام تقنية تجميع الجوار القائم على المتغيرة (Gries & Hilpert 2008) مع الانحراف المعياري (SD) بوصفه مقياس المسافة والربط المتوسط لقاعدة الدمج. يُظهر المخطط التفرّعي الناتج في الشكل (١٢، ٧). أشار المخطط الركامي إلى تجميعين رئيسيين من مطلع القرن السابع عشر إلى أربعينياته ومن مطلع خمسينياته إلى تسعينياته.

(٧، ٥) نمذجة التغيرات في الخطاب: القمم والقعور وتحليل الثقل الاستعمالي

فكّر...

قبل قراءة هذا المبحث انظر في تطور تكرارات كلمة war في جريدة التايمز بين عامي ١٩٤٠-٢٠٠٩، هذه الجريدة هي صحيفة يومية بريطانية، وهي جزء من صحافة (الجودة). هل يمكنك ربط القمم الملحوظة بالأحداث التاريخية في القرنين العشرين والحادي والعشرين؟



يتطور الخطاب تطورًا لا حدَّ له، فهناك فترات تبرز فيها مفاهيم أو كلمات معينة أو يُبتكر فيها كلمات، وفترات تبهت فيها هذه المفاهيم والكلمات وتبرز كلمات ومفاهيم أخرى. الخطاب هو سوق لغوية تتنافس فيها الكلمات والأفكار التي تعبر عنها هذه الكلمات على الظهور والاستخدام. كيف يمكن تحليل مثل هذا النظام الديناميكي عندما ننظر في تطوره في فترة من الزمن؟ قدم غابريلا توس ومارتشي Gabrielatos & Marchi (٢٠١٢) أسلوبًا لتحليل الخطاب التعاقبي الذي يسميانه (طريقة الموجة والقمة والقعر wave, peak and trough WPT)، ولكن يُعرف باسم (القمم والقعور). القمم والقعور طريقة لتحليل بيانات تعاقبية وتمثيلها تمثيلًا بيانيًا، ويُطبَّق فيها نموذج الانحدار غير الخطي، وتحديدًا النموذج التجميعي المعمم Generalized Additive Model GAM، ويُجرى هذا التطبيق على نقاط البيانات التي تُظهر تطور متغير لغوي مع مرور الوقت، من أجل تحديد القيم المتطرفة ونقاط الخروج عن الاتجاه العام ذات الدلالة الإحصائية. على عكس النوع الأكثر تقليدية من تحليل الانحدار الذي يستخدم خطأً مستقيمًا لإظهار الاتجاه الرئيسي في البيانات (المبحث ٢، ١)، تتناسب القمم والقعور مع منحنى البيانات، وتعكس تمثيلًا أفضل اتجاه الصعود والهبوط (اتجاه القمم والقعور) بخصوص تطور الخطاب التعاقبي. تتكون عملية إنتاج تحليل القمم والقعور من خطوتين إلزاميتين وخطوتين اختياريتين:

١- الإلزامية: الحصول على تكرارات (نسبية) من مدونة المتغيرات اللغوية المعنية لكل فترة من الفترات (مثل السنوات والعقود وما إلى ذلك) التي يشملها التحليل.

٢- اختيائية: حساب الاختلافات بين قيمتين متتاليتين بأخذ القيمة الثانية وطرحها على القيمة الأولى لنقاط البيانات الزمنية، ويحصل ذلك لتسليط الضوء على القيم العالية التي تسبقها القيم المنخفضة (أو العكس) والتي تشير إلى حدوث تغيير أكبر.

٣- اختيائية: تحويل القيم باستخدام لوغاريتم ثنائي ($\log_2$) لتقليل القيم المتطرفة. لا تكون هذه الخطوة ممكنة إلا إذا كانت كل القيم المحولة أرقامًا موجبة لأن اللوغاريتم غير معرف للأرقام السالبة. نظرًا لأن الخطوة الثانية تنتج عادة قيمًا سالبة أيضًا، فإن التحويل اللوغاريتمي ممكن مع بيانات الخطوة الأولى.

٤- الإلزامية: تهيئة نموذج انحدار غير خطي يُعرض على أنه منحنى في الشكل البياني، وحساب مجالي الثقة ٩٥٪ و ٩٩٪ التي تُعرض بوصفها مناطق مظلمة حول المنحنى، وتحديد القيم المتطرفة ذات الدلالة أو نقاط البيانات الواقعة خارج نطاق مجال الثقة.

لنلقِ نظرة -على سبيل المثال- على البيانات من مهمة (فَكَر) التي تتبّع ظهور مصطلح war المستخدم في صحيفة التايمز. فيما يأتي عرض لكل الخطوات الاختيارية والإلزامية لإجراء القمم والقعور:

الأولى: الإلزامية. وللتوضيح، يعرض الجدول (٧, ٥) نقاط البيانات الأولية العشر من مجموعة البيانات بأكملها التي تتناول المدة ١٩٤٠-٢٠٠٩. تُستخدم التكرارات النسبية (لمبحث ٢, ٣) لكلمة war لتفسير حقيقة أحجام العينات غير المتساوية في المدة الزمنية.

الثانية: اختيارية. تعرض الاختلافات بين كل نقطتين زمنيتين متتاليتين من الجدول (٧, ٥) في الجدول (٧, ٦). يمكننا أن نرى أن هذه الخطوة تضمن حصول القيم العالية (مثل: 1609.75) المسبوقة بقيم أقل (مثل 1473.69) على درجة أكبر (136.6) من القيم العالية (مثل 1623.78) المسبوقة بقيم عالية أخرى (مثل 1609.75). الدرجة الناتجة عن المقارنة الأخيرة هي 14.03 فقط.

جدول (٧, ٥). التكرارات النسبية (لكل مليون كلمة) لكلمة war.

المدة	1940	1941	1942	1943	1944	1945	1946	1947	1948	1949
التكرار النسبي لكلمة war	1473.69	1609.75	1623.78	1505.46	1283.21	1299.81	851.03	590.37	479.7	423.14

جدول (٧, ٦). الفروق بين التكرارات النسبية لكلمة war.

المدة	1940/1	1941/2	1942/3	1943/4	1944/5	1945/6	1946/7	1947/8	1948/9
القيمة ١ - القيمة ٢	136.06	14.03	-118.32	-222.25	16.6	-448.78	-260.66	-110.67	-56.56

الثالثة: اختيارية. يكون تحويل القيم باستخدام لوغاريتم ثنائي (لوج ٢) بالقيم الموجبة الواردة في الجدول (٧, ٥)، وقد مثّلت قيم لوج ٢ المحولة في الجدول (٧, ٧).

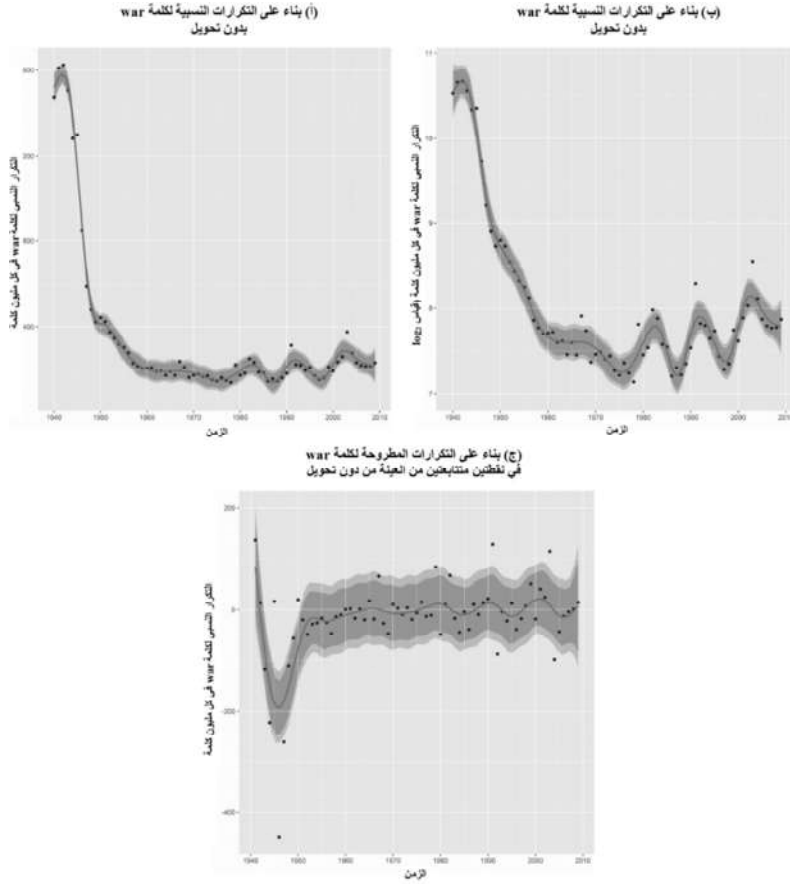
جدول (٧, ٧). التكرار النسبي ذو اللوغاريتم المحول (لكل مليون) لكلمة war.

المدة	1940	1941	1942	1943	1944	1945	1946	1947	1948	1949
لوج ٢ للتكرار النسبي لكلمة war	10.53	10.65	10.67	10.56	10.33	10.34	9.73	9.21	8.91	8.72

الرابعة: إلزامية. لا بد من تهيئة نموذج الانحدار غير الخطي مع عدة بيانات تعاقبية مختلفة من البيانات المذكورة سابقاً. لكي يشتغل النموذج بالطريقة الصحيحة سنحتاج إلى عدد كاف من نقاط أخذ العينات، أي سنعرض مجموعة البيانات الكاملة للرسوم البيانية للقمم والقعور بين عامي ١٩٤٠ و ٢٠٠٩.

عندما ننظر إلى النتيجة في (أ) و(ب) في الشكل (٧، ١٤) نرى بأنها تشبهان الخط البياني البسيط في مهمة (فكر). تُظهر الرسوم البيانية في هذا الشكل منحنى يناسب البيانات بالشكل الأمثل، ومجالات الثقة (٩٥٪) في منطقة أعمق، و(٩٩٪) في منطقة أعمق بحواف أفتح تساعدنا على تحديد القيم المتطرفة أو نقاط في الزمن يكون فيها تكرار war أعلى أو أقل بكثير مما توقعه النموذج (أو الخط). أصبحت بعض التباينات باستخدام تحويل اللوغاريتم في النتيجة (ب) أكثر وضوحاً، وأصبح الشكل البياني غير منحرف بشكل كبير بسبب التكرارات العالية جداً لمصطلح war إبان الحرب العالمية الثانية (١٩٤٠-١٩٤٥: المدة التي يتناولها الشكل البياني). يُساعدنا الحل في النتيجة (ب) على تحديد أربع قمم وقعر واحد من النقاط ذات القيم المهمة، ونلاحظ القمم في الأعوام ١٩٦٧ و ١٩٨٢ و ١٩٩١ و ٢٠٠٣ والقعر في عام ١٩٧٨. عندما ننظر إلى أسطر الكشف السياقي يمكننا ملاحظة أن القمم الموضحة ترتبط بالحرب العربية الإسرائيلية الثالثة (١٩٦٧)، وحرب فوكلاند (١٩٨٢)، وحرب الخليج (١٩٩١) وحرب العراق (٢٠٠٣).

من الصعب تفسير القعر في عام ١٩٧٨ لكن يمكن فهمه بطريقة غير مباشرة بالإشارة إلى السنوات المجاورة، وبخاصة في الأعوام ١٩٧٩-١٩٨٢ إذ كانت مرّات ذكر war أكثر بروزاً مما كانت عليه في عام ١٩٧٨. لنلتفت الآن إلى النتيجة النهائية في الجزء (ج) في الشكل البياني من الشكل (٧، ١٤) الذي يُبرز الاختلافات بين نقاط متتالية في الوقت المناسب. هنا يمكننا أن نلاحظ أربع قمم في الأعوام ١٩٤٥ و ١٩٥٠ و ١٩٩١ و ٢٠٠٣ وثلاثة قعور في الأعوام ١٩٤٦ و ١٩٩٢ و ٢٠٠٤. تمثل القمم نهاية الحرب العالمية الثانية في عام ١٩٤٥، والحرب الكورية والحرب الباردة في عام ١٩٥٠، وحرب الخليج في عام ١٩٩١، وحرب العراق في عام ٢٠٠٣. إذن القعور نقاطاً بها عدد أقل بكثير من مرات ذكر كلمة war بالنسبة إلى العام السابق. يمكن تفسيرها بالإشارة إلى القمم السابقة مباشرة في الآتي: نهاية الحرب العالمية الثانية (١٩٤٦)، وحرب الخليج (١٩٩٢) وحرب العراق (٢٠٠٤).



شكل (١٤، ٧). الشكل البياني للقمم والقصور الناتجة.

لنناقش تحليل التقلب الاستعمالي (Brezina et al. in prep.; Baker et al. 2017). يُعد هذا التحليل طريقة تبحث في تطور معنى كلمة معينة بتحليل متغير للمصاحبات المحيطة بالكلمة. تكمن صعوبة تحليل التقلب الاستعمالي في تفسير تطور تكرارات ذات صيغة لغوية معينة بدون السياق (انظر: مناقشة الدراسة الرقمية الثقافية والبيانات الضخمة في المبحث ٢، ٧)، لكنه يعمل على تحليل التطور في سياقات التصاحب تحليلًا منهجيًا ومنظمًا (المبحث ٢، ٣ للاطلاع على شرح التصاحب)، ويحدد النقاط التي يحدث فيها تغير كبير في المعنى، ويتكوّن من الخطوات التالية:

الأولى: تحديد مُصاحِبَات الكلمة المعنِيَّة (العقدة) عبر بيانات السلاسل الزمنية.

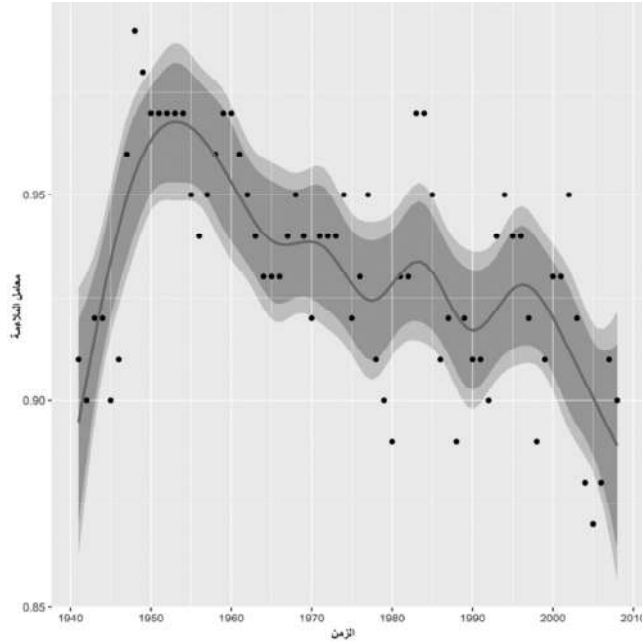
الثانية: التقدير المتكرر للفرق بين المصاحِبَات في أي نقطتين زمنيتين متصلتين.

الثالثة: استخدام تقنية القمم والقعور لتتبع النقاط التي تحدث فيها تغيُّرات كبيرة، وقد حُدِدت على أنها قعور مهمة لأنها أكبر نقاط للاختلاف في المعنى.

لنستخدم البيانات من مهمة (فَكَّر) مرة أخرى. بالنسبة لتحليل التقلُّب الاستعمالي فإنه بدلاً من النظر إلى تكرار كلمة war نحلل مُصاحِبَات كلمة war في صحيفة التايمز لكل مدة، وفي الآتي الإجراءات لذلك.

الثاني: أفضل طريقة لتقدير الفرق بين مديتين متتاليتين من جهة ملف تعريف المصاحبات الخاص بهما هي استخدام إحصائية ملاءمة المقدّر البيني مثل Gwet's AC₁ (المبحث ٥، ٣). في هذه الحالة، فبدلاً من السؤال عن مدى ثبات التصنيفين نتحقق من مدى ثبات مديتين تاريخيتين متتاليتين فيما يتعلق بالمصاحبات المستخدمة.

جدول (٨, ٧). ملف تعريف مصاحبات war.



شكل (١٥، ٧). نتائج تحليل التقلب الاستعمالي لكلمة war (حرب) بين عامي 1940 و2009 (قيمة المعلومات المتبادلة الدنيا ٦، ٥-٥، تكرار المتصاحب النسبي ١٠ وتكرار المتصاحب النسبي ١٠).

الثالث: تطبيق نموذج القمم والقعور، فبدلاً من التكرارات نستخدم مؤشر معامل المواءمة المشترك للملاءمة المُقدَّر البيني، ونبحث عن نقاط ذات قيم معامل المواءمة الدنيا بوصفها نقاط اختلاف للمعنى. يوضح الشكل (١٥، ٧) التحليل الناتج.

في الشكل (١٥، ٧) يمكننا تحديد عشرة قعور مهمة (أي: نقاط الاختلاف في المعنى) بالنظر إلى نقاط البيانات خارج مجال الثقة 99% (أي: خارج المنطقة الغامقة أو المظللة). نقاط الاختلاف هذه هي في الأعوام الآتية: ١٩٤٥، ١٩٤٦، ١٩٧٥، ١٩٧٩، ١٩٨٠، ١٩٨٦، ١٩٨٨، ١٩٩٢، ١٩٩٨، ٢٠٠٤، ٢٠٠٥. على سبيل المثال، في عام ٢٠٠٣ حُدِّدت ١٦ متصاحبة لكلمة war من بينها bush وgulf وIraq وweapons وterror وكلها ترتبط بالحرب في العراق، ومن بينها بقيت ١٤ متصاحبة (بما في ذلك المذكورة أعلاه) إلى عام ٢٠٠٤ لأن الحرب في العراق لا تزال دائرة.

ظهرت ١٢ متصاحبة جديدة ترتبط إلى حد كبير بالبرامج التلفزيونية المتعلقة بموضوع الحرب. وهي hitler و secrets [أي أسرار الحرب العالمية الثانية] و documentary و starring و UKTV، إلخ. وعلى نحو مماثل يمكننا البحث في نقاط أخرى من اختلاف المعنى، والهدف العام من تحليل التقلّب الاستعمالي هو اختزال نطاق التحليل إلى النقاط التي يحدث فيها تحول في الاستعمال، ويمكن بعد ذلك إجراء مزيد من البحث النوعي باستخدام الكشافات السياقية.

التقرير الإحصائي: القمم والقعور وتحليل التقلّب الاستعمالي

١- ما يجب الإفصاح عنه

بالنسبة إلى القمم والقعور التي تحتوي على خطوتين إلزاميتين وخطوتين اختياريتين نحتاج إلى الإفصاح عن جميع الخطوات التي أثبتت. بالنسبة لتحليل التقلّب الاستعمالي ينبغي الإفصاح عن تفاصيل الإجراء الخاص بتعريف المصاحبات باستخدام بارامترات المصاحبة الموصوفة في المبحث (٢، ٣) إضافة إلى اختيار إحصائية ملائمة المقدّر البيئي.

٢- مثال الإفصاح في التقرير

• استُخدمت تقنية القمم والقعور (Gabrielatos & Marchi 2012)، وقيس الفرق بين التكرارات النسبية لكلمة war بين نقاط زمنية متتالية في الأعوام ١٩٤٠-٢٠٠٩، وقد ساعد نموذج الانحدار غير الخطي (GAM) على تحديد أربع قمم للأعوام الآتية: ١٩٤٥ و ١٩٥٠ و ١٩٩١ و ٢٠٠٣، وثلاثة قعور للأعوام الآتية: ١٩٤٦ و ١٩٩٢ و ٢٠٠٤.

• جرى البحث في تطور مجموعات معنى كلمة war باستخدام إجراء تحليل التقلّب الاستعمالي. حُدّدت في المدة من ١٩٤٠-٢٠٠٩ (١٧١) متصاحبة لكلمة war (الحرب): قيمة المعلومات المتبادلة الدنيا ٣، يم5-يس5، تكرار المتصاحب النسبي ١٠ وتكرار المتصاحب النسبي (١٠)، وقورنت ملفات تعريف السنوات المتصلة باستخدام إحصائية اتفاق معامل AC₁.

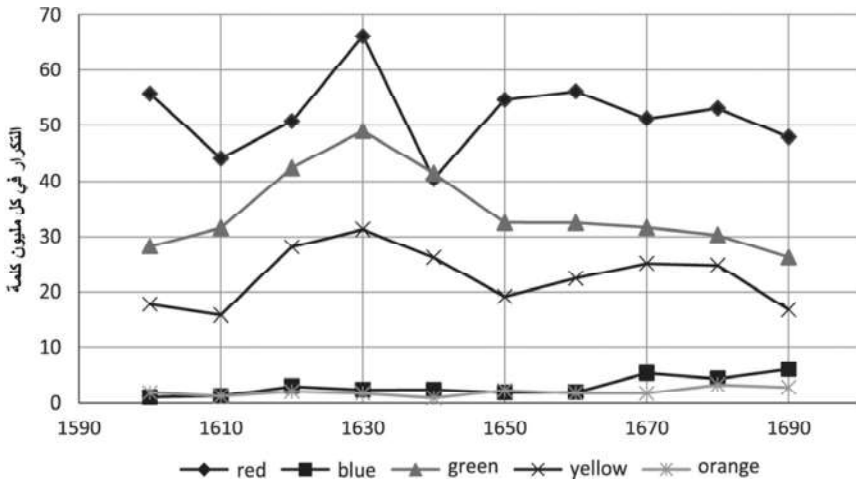
(٦، ٧) التطبيق: الألوان في القرن السابع عشر

أنا جالس على مكتبي، والجو في الخارج بائس وكئيب وممطر وكثيف بالضباب. من النافذة أستطيع أن أرى على التلّة قلعة لانكستر التي تعود إلى العصور الوسطى، وأستطيع أن أرى الدير

والوادي ونهر لُون، وللمباني والأشجار والناس لون باهت. يرفرف علم لانكستر الأحمر من القلعة مثل قطرة من الطلاء تركها فنان عن غير قصد. تعيدني أفكارني إلى التاريخ، وأخذت أتساءل كيف يُنظر إلى اللون في الماضي. هل كان يحمل الحس المعنوي والمادي نفسه؟ في وقت ما بعد الظهيرة المطيرة في لانكستر أبدأ بالبحث في مدونة EEBO. مليار كلمة من الكتابة المبكرة تقدم رؤية فريدة لاستخدام مصطلحات الألوان في القرن السابع عشر. أدون الملاحظات سريعاً، وأرسم الرسوم البيانية وقيم الاحتمال. إليكم نوعاً خاصاً من يوميات رحلتي إلى الماضي مع آلة الزمن في المدونة: استكشافاتي في يوم ١٢ نوفمبر ٢٠١٦.

السؤال الأول: ما الألوان التي كانت الأكثر شعبية في القرن السابع عشر؟

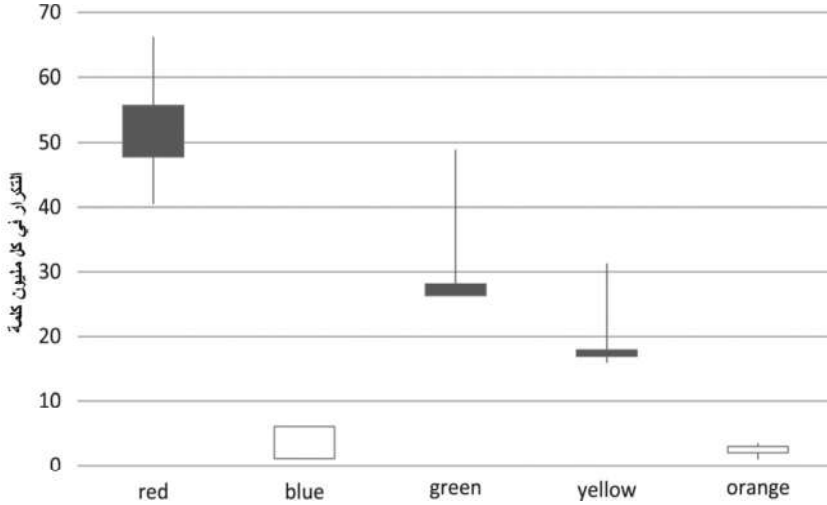
يوفر الخط البياني في الشكل (١٦، ٧) الإجابة. كانت دائماً هي الأحمر red والأخضر green والأصفر yellow.



شكل (١٦، ٧). تكرار مصطلحات الألوان في القرن السابع عشر.

السؤال الثاني: ما قصة مصطلحات اللون في القرن السابع عشر؟

يمكن الحصول على صورة موجزة من مخطط الشموع كما في الشكل (١٧، ٧). انخفض تكرار مصطلحات الألوان الثلاثة (red و green و yellow) انخفاضاً طفيفاً، في حين ازداد تكرار blue و orange. تبين الفتائل الطويلة في red و green و yellow تبدلات كبيرة في هذه المدة.



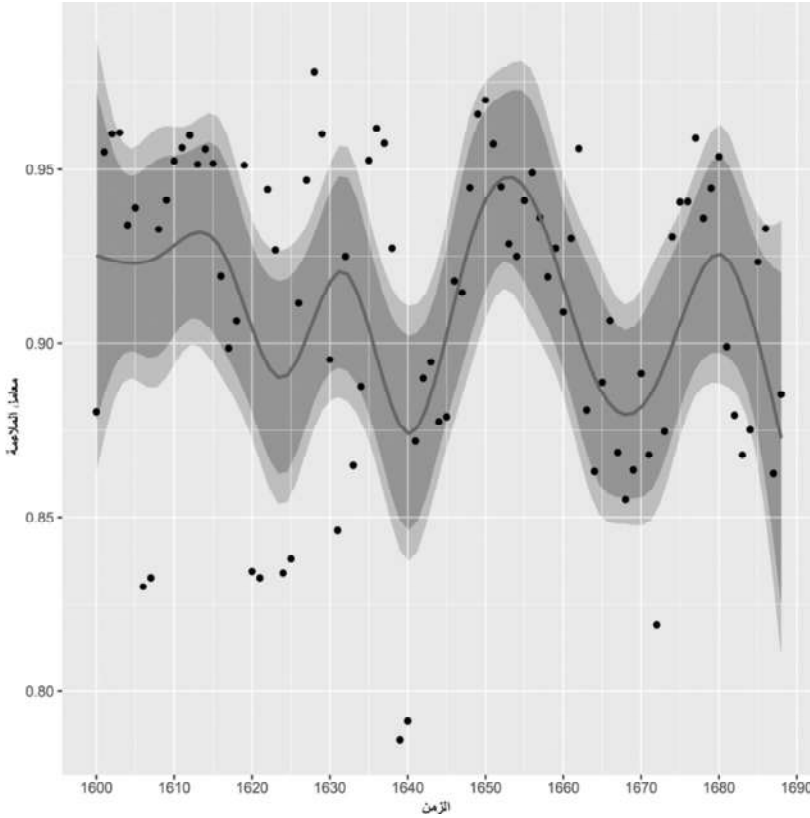
شكل (١٧، ٧). مخطط الشموع: الألوان في القرن السابع عشر.

السؤال الثالث: كان الانخفاض الأكبر في استخدام مصطلح red في عام 1640. هل هذا التغير في التكرار ذو دلالة إحصائية؟

أظهر اختبار التمهيد أن هذا التغير ذو دلالة بالفعل ($p < .001$).

السؤال الرابع: هل للون الأحمر red الأكثر شعبية الحس المعنوي ذاته على مدى القرن السابع عشر؟ لهذا اللون إحساس معنوي يتشكل من متصاحبات اسمية متجاوزة معه وثابتة، وهي على النحو الآتي: coral، و dragon، و flowers، و iron، و rose، و roses، و sea، و wine إضافة إلى المتصاحبات النعتية الآتية التي تعبر عن الألوان: black، و green، و hot، و scarlet، و white، و yellow.

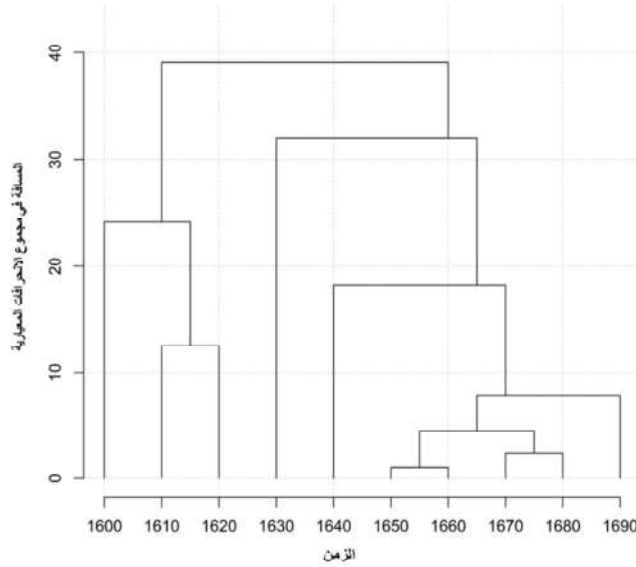
يُظهر تحليل التقلب الاستعمالي (الشكل ١٨، ٧) حدوث التحولات الرئيسية في استخدام الكلمات في عقود (أُخذت عشر سنوات في كل نقطة بيانات) للأعوام الآتية: ١٦٠٧، و ١٦٠٦، و ١٦٢٠، و ١٦٢١، و ١٦٢٤، و ١٦٢٥، و ١٦٣١، و ١٦٣٣، و ١٦٣٩، و ١٦٤٠، و ١٦٧٢، و ١٦٨٢، و ١٦٨٣، و ١٦٨٤. على سبيل المثال، يمكن تتبع القعر في عامي ١٦٣٩ و ١٦٤٠ لبحث التغييرات الاجتماعية والسياسية في إنجلترا عن هذه النقطة الزمنية. أكدت الأدبيات في هذه المدة على الارتباطات التوراتية باللون الأحمر: cup [of fornications], pottage, garments, pharaoh and his host, tail [of the red dragon = devil]



شكل (١٨، ٧). تحليل التقلب الاستعمالي لكلمة red (أحمر) بين عامي 1600 و1699 (قيمة المعلومات المتبادلة الدنيا ٣، ٥-٥، تكرار المتصاحب النسبي ١٠ وتكرار المتصاحب النسبي ١٠).

السؤال الخامس: بالنظر إلى المصطلح red طوال القرن السابع عشر، ما المدة التي يمكن تحديدها على أساس تكرار الاستخدام؟

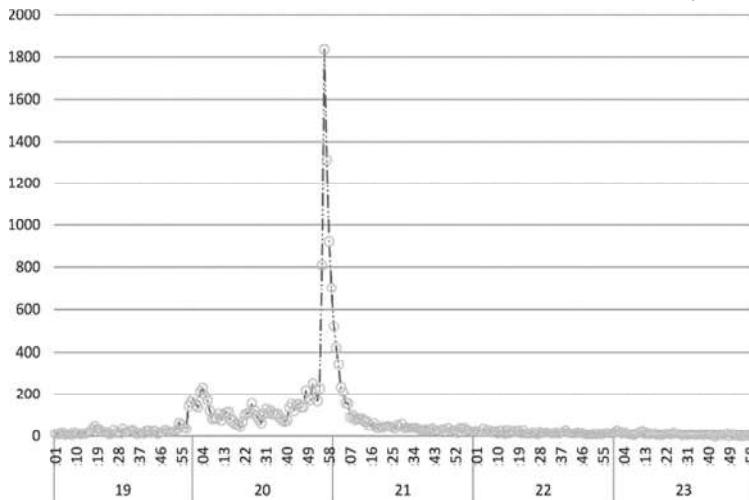
يمكن استخدام جميع الجوار القائم على المتغيرة لتجميع السنوات الفردية وفقاً لتكرار استخدام مصطلح red. يظهر المخطط التفريعي الناتج في الشكل (١٩، ٧) الذي يوضح الانقسام الرئيسي بين مطلع القرن السابع عشر وعشرينياته وبين مطلع ثلاثينياته وتسعينياته. خلاف ذلك، لا تدعم البيانات أي تجميع أكثر كثافة.



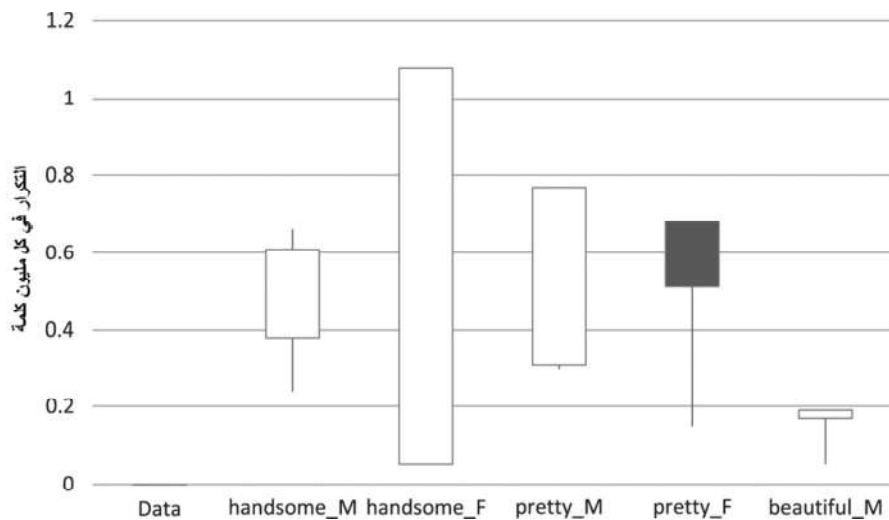
شكل (١٩, ٧). تجميع الجوار القائم على المتغيرة VNC: كلمة red (أحمر) في القرن السابع عشر.

(٧, ٧) تمارين

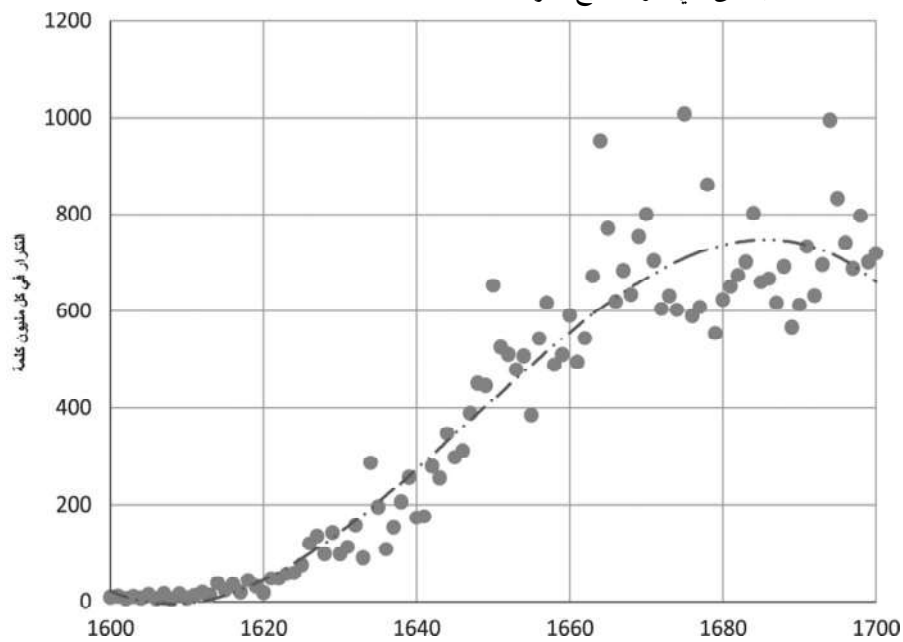
١- فسر الرسوم البيانية الثلاثة في الأشكال (٧, ٢٠, ٢١, ٢٢).



شكل (٢٠, ٧). عدد التغريدات المرتبطة برنامج X Factor البريطاني (١٦ نوفمبر ٢٠١٤م ٧-١١ مساء).



شكل (٧، ٢١). تطور تكرارات handsome (وسيم)، وpretty (مليح)، وbeautiful (جميل) متبوعة بالذكر M: أو بالأنثى F في القرن السابع عشر.



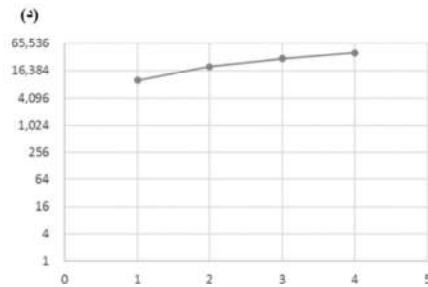
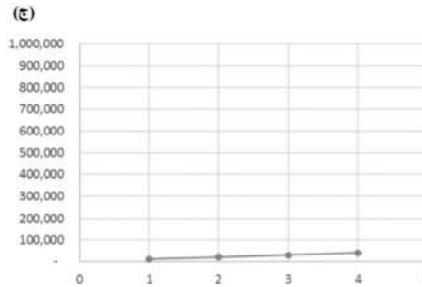
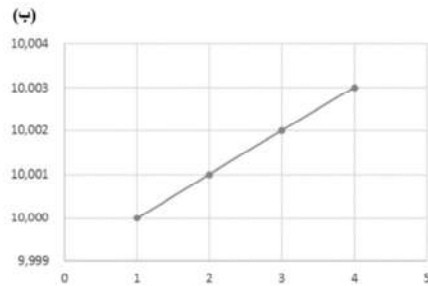
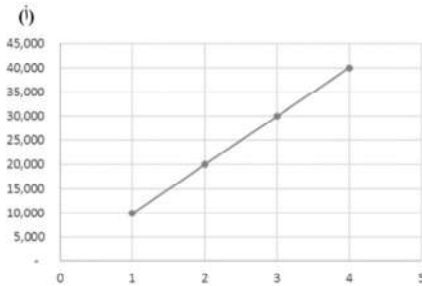
شكل (٧، ٢٢). تطور تكرارات ضمير الملكية its في القرن السابع عشر.

٢- املأ الفراغات في محل الخطوط في التوصيفات أدناه.

على مدار القرن العشرين، كانت تكرارات () shall ، () should ، () may ، () might ، () must ، () will ، () can ، () could .

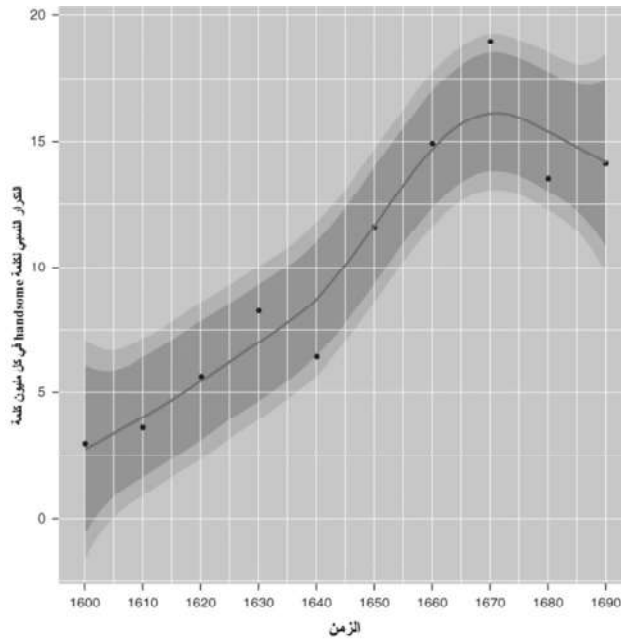
في القرن السابع عشر استخدمت صفة handsome () مع الإناث. بها أن صفة pretty () في هذا السياق كانت صفة beautiful () مع الإناث.

٣- انظر نظرة ناقدة في الاتجاهات في الشكل (٧، ٢٣). أي من هذه يمثل تغييرًا أكبر؟

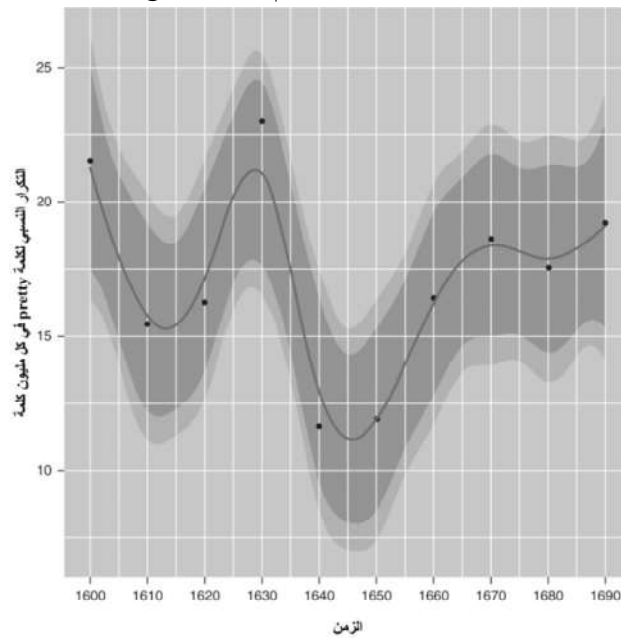


شكل (٧، ٢٣). أربعة سيناريوهات لتغير التكرار.

٤- فسر الرسوم البيانية للقمم والقعور في الشكلين (٧، ٢٤) و (٧، ٢٥) اللذين يوضحان تطور handsome و pretty في القرن السابع عشر.



شكل (٧, ٢٤). كلمة handsome (وسيم) في القرن السابع عشر.



شكل (٧, ٢٥). كلمة pretty (مليح) في القرن السابع عشر.

تذكر الآتي

- تتطلب التحليلات التاريخية التي تستخدم البيانات المتاحة والناقصة دراسة نقدية ما يأتي:
- التمثيلية التعااقبية للمدونات، والتفسيرات البديلة للتطور اللغوي، وتبذل معاني الصيغ اللغوية.
- تشتمل خيارات التمثيل على الآتي: الخطوط البيانية، ومخططات رسم المربع، وأعمدة الخطأ، وخطوط المؤشر، ومخططات الشموع.
- يُستخدم اختبار التمهيد لمقارنة مدونتين تمثلان نقاطاً زمنية مختلفة مع استخدام أسلوب إعادة أخذ العينات المتعددة لبيانات المدونة.
- (القمم والقعور) تقنية تُناسب الانحدار غير الخطي للبيانات التاريخية، وتنتج رسمًا بيانيًا يلقي الضوء على القيم المتطرفة المهمة في عملية التطور التاريخي للغة والخطاب.
- تحليل التقلب الاستعمالي هو إجراء معقد يجمع بين المقارنة التلقائية للتصاحب اللغوي في مدة تاريخية معينة وبين تقنية القمم والقعور.

قراءات إضافية

- Baker, P. (2011). Times may change, but we will always have money: diachronic variation in recent British English. *Journal of English Linguistics*, 59(1), 65-88.
- Brezina, V., McEnery, T. & Baker, H. (in prep.). *Usage fluctuation analysis: a new way of analysing shifts in historical discourse*.
- Hilpert, M. & Gries, S. Th. (2009). Assessing frequency changes in multistage diachronic corpora: applications for historical corpus linguistics and the study of language acquisition. *Literary and Linguistic Computing*, 24(4), 385-401.
- McEnery, T. & Baker, H. (2017). *Corpus linguistics and 17th-century prostitution*. London: Bloomsbury.
- Säily, T. (2014). Sociolinguistic variation in English derivational productivity: studies and methods in diachronic corpus linguistics. *Mémoires de la Société Néophilologique de Helsinki XCIV*, Helsinki. Available at: <https://helda.helsinki.fi/handle/10138/136128>.

أدوات لانكستر الإحصائية

- ١- يمكن استنساخ جميع التحليلات المبينة في هذا الفصل باستخدام أدوات لانكستر الإحصائية. تشمل الأدوات المتاحة لهذا الفصل:
 - اختبار التمهيد.
 - تجميعات الجوار.
 - القمم والقعور.
 - تحليل التقلب الاستعمالي.
- ٢- يقدم الموقع أيضًا مواد إضافية للطلاب والمعلمين.

الاختتام: عشرة مبادئ للتفكير الإحصائي والتحليل البعدي وأحجام التأثير

(١, ٨) ما الذي يتناوله هذا الفصل؟

يدور هذا الفصل الأخير حول استحضار كل ما ذكر على مستويات مختلفة، فهو أولاً يجمع المعرفة الإحصائية التي نوقشت في هذا الكتاب، ويسلط الضوء على عشرة مبادئ أساسية للتفكير الإحصائي المطبق على المدونات الحاسوبية. يقدم الفصل تقنية إحصائية تسمى التحليل البعدي meta analyses، وهي طريقة لجمع نتائج دراسات متعددة والجمع بينها منهجياً، لأجل فهم أفضل للنتائج. يمكن للتحليل البعدي أن يجمع بين نتائج دراسات متعددة في صيغة رياضية واحدة على عكس المعتاد في الدراسات التقليدية. وعلى الرغم من أن التحليل البعدي الشكلي أصبح الآن شائعاً إلى حد ما في عدد من التخصصات مثل علم النفس واكتساب اللغة الثانية والعلوم الطبية وما إلى ذلك فتطبيقه في لغويات المدونة الحاسوبية يمثل مشكلة بسبب الافتقار العام إلى إجراء قياسات حجم التأثير. يُشجع هذا الفصل على توحيد إجراءات قياسات أحجام التأثير في أبحاث المدونة، ويوضح كيف يمكن تطبيقه، ويستعرض قياسات حجم التأثير الشائعة ويوفر دليلاً لتفسيرها. سنحجب في هذا الفصل عن الأسئلة الثلاثة الآتية:

- ما المبادئ الرئيسية للتفكير الإحصائي في لغويات المدونة الحاسوبية؟ (المبحث ٢, ٨).
- كيف يمكن أن نجمع بين النتائج من دراسات متعددة؟ (المبحث ٣, ٨).
- كيف ينبغي تفسير أحجام التأثير؟ (المبحث ٤, ٨).

(٨, ٢) عشرة مبادئ للتفكير الإحصائي

فكّر...

قارن بين الجدولين (٨, ١) و(٨, ٢). ما عدد الاختلافات في القيم التي يمكنك العثور عليها؟ يوضح الجدول (٨, ١) القيم الواردة في مجموعة البيانات الأصلية المخزنة في جدول بيانات. أنشئ الجدول (٨, ٢) الذي ظهر في تقرير البحث النهائي بنسخ الأرقام وتقريبها يدوياً من الجدول (٨, ١).

جدول (٨, ١). استخدام الزمن الماضي والحاضر في سجلات لغوية مختلفة (مجموعة البيانات الأصلية).

السجل اللغوي	الزمن_الماضي	الزمن_المضارع
الأخبار - تقرير صحفي	4.490227273	4.424090909
الأخبار - الافتتاحية	2.452692308	5.735
التراجم	4.97025974	3.502987013
الكتابة - الأكاديمية	2.249625	3.776375
الرواية - العام	6.704482759	4.190689655
أدب - الغموض	8.12	3.141666667
القصص الرومانسية	7.690689655	3.076896552
الفكاهة	4.601111111	4.916666667

جدول (٨, ٢). استخدام الزمن الماضي والمضارع في سجلات لغوية مختلفة (تقرير بحثي).

السجل اللغوي	الأخبار والتقارير	الأخبار الافتتاحية	التراجم	الكتابة الأكاديمية	الرواية العامة	أدب الغموض	القصص الرومانسية	الفكاهة
الزمن_الماضي	4.4	2.5	4.9	2.2	6.7	8.1	7.7	4.6
الزمن_المضارع	4.4	5.7	3.5	3.7	4.2	3.1	7.1	4.9

يُعدّ الإحصاء مجالاً يساعدنا على التعبير عن المعلومات الكمية بدقة، وهذا مطلوب لإنتاج نتائج الأبحاث التي يمكن أن تشكل أساس المعرفة العلمية المتينة، ويمكن أن تؤدي إلى التقدم في هذا المجال، ولتليخيص أهم جوانب التفكير الإحصائي المطبقة على لغويات المدونة الحاسوبية، نذكر في الآتي المبادئ العشرة لذلك التفكير.

المبدأ الأول: الانتباه إلى التفاصيل وإيلاء الاهتمام عند النظر في نواتج أدوات المدونة، وإدخال البيانات في جداول البيانات، ونسخ البيانات إلى تقارير البحث وكذلك إلى امتدادات أخرى من معالجة البيانات المنخفضة المستوى

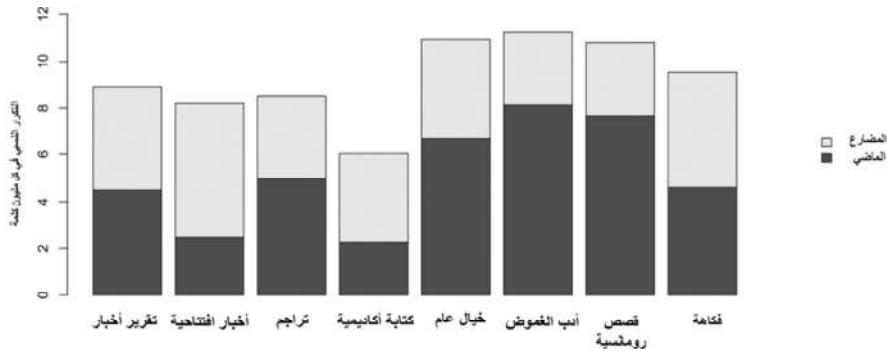
الشیطان، كما يقول المثل، يكمن دائماً في التفاصيل. على الرغم من ارتكاز الكتب الإحصائية المتخصصة على التقنيات الإحصائية وتفسيرات قيم الاحتمال وما إلى ذلك فالعمليات المنخفضة المستوى مثل هيكلية البيانات اللغوية الخام وتحديد الحزم الإحصائية المناسبة لها تظل عادةً مهمشة. لهذه العمليات القدر نفسه من الأهمية في الحصول على نتائج موثوقة، فعلم الإحصاء كما نعلم مجال لا يتعلق بتحليل البيانات فحسب، ولكن يتعلق أيضاً بالتجميع المنهجي والموثوق للبيانات (المبحث ٢، ١). عند المقارنة بعناية يتضح أن الجدولين (١، ٨) و (٢، ٨) من مهمة (فكر) يختلفان في أربعة مداخل، وهو ما يدل على الأخطاء المطبعية والأخطاء النموذجية الناجمة عن التقريب غير الصحيح. يمكن أن تصل مثل هذه الأخطاء بسهولة إلى التقارير البحثية النهائية. في هذه الحالة تحدث الأخطاء في ٢٥٪ من القيم في خانات الجدول! علينا أن نتذكر أنه إذا ارتكبنا خطأ في القياس أو ارتكبنا خطأ في أثناء نسخ البيانات من أداة إلى أخرى فإن التحليل النهائي سيشوبه عدم الدقة الأولية بغض النظر عن مدى تعقيد التقنية الإحصائية التي نستخدمها. يشير ليك و بينغ Leek & Peng (2015: 612) إلى أن "قيم الاحتمال هي مجرد قمة جبل الجليد"، ولذلك ينبغي إيلاء الاهتمام بسلسلة البيانات بأكملها، التي تتضمن جمع البيانات وتنظيفها وتحليل البيانات الاستكشافية والنمذجة الإحصائية والاستدلال الإحصائي. يمكن أن تساعد أدوات المدونة هذه العملية بدمج التحليلات الإحصائية لتضمين أكبر قدر ممكن من سلسلة البيانات. يمكن بذلك تقليل فرصة حدوث خطأ غير مقصود في نسخ البيانات. فعلى سبيل المثال، تتبع حزمة مدونة الجيل الجديد LancsBox # (Brezina et al. in prep.) هذه الفلسفة بتقديم وظائف (بحث وتحليل) قوية مع تحليلات إحصائية قابلة للتخصيص. النصيحة بسهولة هي: انتبه إلى التفاصيل وحدد النقاط في التحليل التي يمكن أن يحدث فيها أخطاء، وتحقق دائماً مرة ومرتين وثلاثاً من دقة القيم المستخدمة في التحليل.

المبدأ الثاني: الأساسيات أولاً: تعود على المدونة والإحصاءات الوصفية للأداء

من أهم المهّمات للمحلل تحقيق فهم عام وجيد للبيانات. يتطلب هذا عادةً القراءة المتأنية لدليل المدونة، والتعود على تكوين المدونة، وفحص أسطر الكشاف السياقي للاطلاع على الأمثلة الفعلية لاستخدام اللغة وراء الأرقام التي حصلنا عليها، وإنتاج موجز عام ورسوم بيانية يسيرة تبين الاتجاهات الرئيسية في مجموعة البيانات، ويندرج هذا كلّ تحت الإحصاءات الوصفية. دوّن وصفاً موثقاً للبيانات، فالتعميم من العينة (المدونة) إلى المجموع اللغوي (اللغة نفسها) باستخدام الاختبارات الإحصائية المتطورة وقيم الاحتمال يفترق إلى أسس، ويجعل التفسير الحقيقي للنتائج صعباً إن لم يكن مستحيلاً. فعلى سبيل المثال، تؤخذ مجموعة بيانات الجدول (١، ٨) من مدونة BE06 وهي مدونة من مليون كلمة من اللغة الإنجليزية المكتوبة جُمعت وفقاً لإطار أخذ العينات من عائلة براون (Baker 2009). يقدم الشكل (١، ٨) المأخوذ من مقال يصف المدونة وتصميمها تفاصيل عن تمثيليتها وتوزيع الوعاء والسجل اللغوي. لاحظ أن المدونة تتضمن أوعية وسجلات لغوية مكتوبة تقليدية، لكنها لا تختبر أوعية وسجلات جديدة على الإنترنت، مثل منشورات المدونات الشخصية، والتغريدات، ومنتديات النقاش عبر الإنترنت، إلخ. يعد هذا النوع من المعلومات ضرورياً لأي تعميمات تستند إلى مجموعة المستندات. بالتركيز في استخدام الزمن الماضي في الأوعية والسجلات اللغوية المختلفة من الجدول (١، ٨) يمكننا أن نرى أن الكتابة الأكاديمية تُسجل أقل تكراراً للزمن الماضي، أما أدب الغموض فيسجل أعلى معدل تكرار للأزمنة الماضية. يتضح هذا النموذج على الفور من خطاطة الأعمدة المتكدسة في الشكل (٢، ٨). انظر: الجدول (٣، ٨) التي يتضمن أمثلة على استخدام الزمن الماضي في هذين الوعاءين والسجلين اللغويين مع سياق الكشافات. إذن، فالشرط الأساسي للتحليل الناجح للمدونة هو وجود معرفة شاملة بمجموعة البيانات التي يستند إليها التحليل.

فئات النص	عدد النصوص في كل فئة
أ	(صحفي: تحقيق صحفي)
ب	(صحفي: افتتاحية)
ج	(صحافة: تقييقات نقدية)
د	(ديني)
هـ	(مهارات ومهن وهوايات)
و	(الثقافة الشعبية)
ز	(الأدب البحث، والتراجم، والمقالات)
ح	(وثائق حكومية متنوعة، تقارير مؤسسية، تقارير صناعية، جدول جامعي، مجلة دار الصناعة)
ط	(كتابات مستنيرة وعلمية)
ي	(كتابات مستنيرة وعلمية)
ك	(الغموض والرواية البوليسية)
ل	(الخيال العلمي)
م	(المغامرات والرواية الغربية)
ن	(القصص الرومانسية والغرامية)
س	(الفكاهة)

شكل (١، ٨). نظرة عامة على الأنواع في BE06 (Baker 2009).



شكل (٢، ٨). الزمن الماضي في أنواع الكتابة المختلفة من مدونة BE06.

جدول (٨,٣). أمثلة على استخدام الزمن الماضي في الكتابة الأكاديمية وأدب الغموض.

الكتابة الأكاديمية		
self-reported weight and height to calculate body	used	to measure anxiety and depression, and we
لقياس القلق والاكتئاب، واستخدمنا الوزن والطول المبلغ عنهما ذاتياً لحساب ...		
used to investigate the relationship between weight	was	control during feeding. Next, a Pearson's correlation
السيطرة في أثناء التغذية. بعد ذلك، استُخدمت علاقة بيرسون للتحقيق في العلاقة بين الوزن و		
re-created for various audiences in the second	were	analyses how practices of attention and description
يحلل كيف أُعيد إنشاء ممارسات الاهتمام والوصف لمختلف الجماهير في		
also found in dominant hand key grip	were	= 0.806). Secondary outcomes Statistically significant differences
= 0.806). عُثر على فروق ذات دلالة إحصائية في		
obtained from Sigma-Aldrich, Fisher Scientific or Shell	were	initial purities in excess of 99% and
تزيد على ٩٩٪، وحُصل عليها من سيجما-ألدريش أو فيشر أو شيل		
the model that was practiced in this	was	the nurse should further problems develop. This
ينبغي أن تحل الممرضة المشكلات التي تنشأ. كان هذا هو النموذج الذي تمارسه في هذا		
multiple pains in the knee-pain group (both	reported	and more younger responders than older responders
وأفاد المستجيبون الأصغر سناً من المستجيبين الأكبر سناً عن آلام متعددة في مجموعة ألم الركبة (كل من		
scientists generating transcriptomic data for a variety	included	the formulation of a working group that
تشكيل مجموعة العمل التي تضمنت العلماء الذين يولدون بيانات الترنسكريبتوم لمجموعة متنوعة		
worse in the knee-pain group than in	was	to 100 (best health). Physical function restriction
إلى ١٠٠ (صحة أفضل). كان تقييد الوظيفة البدنية أسوأ في مجموعة آلام الركبة منه في		
social housing and first time buyer homes.	supplied	a real need for significantly more State
حاجة حقيقية لمزيد كبير من الإسكان الاجتماعي الموفر من الدولة والمساكن المشتراة لأول مرة.		
أدب الغموض		
and June Carter came in, that bitch,	turned	the guys give a start and I
بدأ الفتيان والتفت وجاءت جون كارتر، تلك العاهرة		
her eyes, willing the image of their	closed	stared straight ahead and for several minutes
حدقت النظر إلى الأمام مباشرة لعدة دقائق وأغمضت عينيها، تريد		
after a long silence. "There's nothing creepy	said	the kid wanting to bury me," Mike
قال مايك بعد صمت طويل: "الطفل الذي يريد أن يدفني...". "لا يوجد شيء مخيف		
his glasses and marked a couple of	retrieved	to make his case strongly enough. He
ليجعل قضيته مناسبة جداً. استعاد نظارته وميز زوجين من		

تابع جدول (٣، ٨).

in an armchair, a book on his	slumped	head around the door and saw him
يتجول قرب الباب ورآه مسترخياً على كرسي بذراعين، وفي (يده) كتاب		
out. Christ, if you were crazy already,	blasted	as he went. A loud klaxon immediately
حيث ذهب. دوى صوت بوق بصوت عالٍ. المسيح، إذا كنت مجنوناً بالفعل،		
about to die. His mouth was so	was	his own heartbeat, and wondering if he
ضربات قلبه، ويتساءل عما إذا كان على وشك الموت. كان فمه		
bitterly. It was easy to behave when	thought	Sharples had been a model prisoner, Tony
كان شاربلز سجيناً مثاليًا، فكر توني بمرارة. كان من السهل أن تتصرف عندما		
he was being examined through the peephole.	guessed	from just behind the door, and Slider
من خلف الباب مباشرة، وخمن سلايدر أنه كان يُراقب من ثقب الباب.		
devised a method of avoiding taking his	had	who were deranged but smart. Recently, Allen
الذين كانوا مختلين ولكن أذكاء. في الآونة الأخيرة، ابتكر آلن طريقة لتجنب أخذ هذه		

المبدأ الثالث: الوضوح إذ ينبغي أن يكون استخدام الإجراءات الإحصائية واضحاً وشفافاً وله دوافع جيدة للوضوح في إحصائيات المدونة عدد من الجوانب. أولاً: عند استخدام التقنيات الإحصائية ينبغي أن نوضح سبب اختيار إجراء إحصائي معين. عادة ما يكون الدافع وراء اختيار إجراءات إحصائية معينة هو وجود نوع معين من بيانات المدونة، وتصميم بحث معين (المبحث ٤، ١). يمكن -على سبيل المثال- دراسة الفروق بين الأوعية والسجلات اللغوية الفردية في الجدول (١، ٨) باستخدام تحليل التباين الأحادي (المبحث ٣، ٦) فنأخذ في الحسبان متغيراً لغوياً واحداً كالزمن الماضي مثلاً. بدلاً من ذلك، يمكن استكشاف البيانات باستخدام التحليل المتعدد الأبعاد (المبحث ٤، ٥) إذ يُنظر في عدد كبير من المتغيرات اللغوية في الوقت نفسه. يناقش هذا الكتاب مجالات مختلفة من التحليل القائم على المدونة التي تتوافق مع التقنيات الإحصائية الأنسب. في العديد من الحالات تكون المنهجيات الأخرى صالحة على قدم المساواة في كل حالة على حدة، وينبغي توضيح الدافع وراء اختيار تقنية دون أخرى. ثانياً: ينبغي الإفصاح عن الإحصاءات بوضوح، ويتضمن ذلك عرضاً واضحاً للبيانات في الرسوم البيانية والجداول، وكذلك الإفصاح الموحد عن الاختبارات والإجراءات الإحصائية ليتسنى للقارئ تفسير النتائج بسهولة وإعادة إنتاج الدراسة. قُدمت في كل الفصول السابقة مباحث (التقرير الإحصائي) التي تضمّنت أساليب الإفصاح الإحصائي.

المبدأ الرابع: البيانات وإيلاء اهتمام خاص بوجودتها وبجودة إجراءات البحث

يعتمد نجاح أي بحث على جودة البيانات وفعالية الإجراءات التحليلية، وتُفحص جودة بيانات المدونة في أبحاث المدونة فحصاً نادراً. المقولة: "لقد وجدتُ ذلك في المدونة، إذن هو صحيح" نابعة من الشعور بأن للمدونات قوة سحرية مع أنها ليست كذلك. حتى المدونات الراسخة مثل المدونة الوطنية البريطانية BNC لا تخلو من الأخطاء وعدم الاتساق والتحيز المحتمل. ينبغي التعامل مع جميع المدونات بأسلوب نقدي، فعلى سبيل المثال يُوضح غابلاسوفا وجود قدر كبير من التنوع في المدونات المصممة لتمثيل النوع نفسه من اللغة، ويدعو إلى ضرورة التحقق التجريبي من بيانات المدونة. يمكن أن تقدم المدونة إجابات مختلفة إلى حد كبير عند بحثها باستخدام أدوات مختلفة. يشير بريزينا وتيمبرلي (Brezina & Timperley 2017) إلى أن الإجابات عن السؤال الذي يتعلق بحجم المدونة الوطنية البريطانية تختلف بنسبة تصل إلى ١٧٪ حسب الأداة التي نستخدمها. في الحساب ومعالجة البيانات يُستخدم الاختصار المعروف GIGO (ويعني مُدخلات خطأ تؤدي إلى مُخرجات خطأ garbage in, garbage out) لتذكيرنا بأنه ينبغي التعامل مع البيانات بأسلوب نقدي وإلا فموثوقية التحليلات غير مضمونة.

المبدأ الخامس: حجم التأثير: حساب حجم التأثير الذي لوحظ في البيانات والإفصاح عنه وتفسيره

من المهم دائماً التفكير في التأثير العملي للملاحظات بالنسبة للغة التي ننتجها باستخدام المدونات. لمساعدتنا على التعبير عن هذا الجانب من النتائج ينبغي استخدام قياسات حجم التأثير. يمكن تعريف حجم التأثير بأنه "مقدار أي شيء قد يكون ذا أهمية" للباحثين (Cumming 2012: 38). في أبحاث المدونة الحاسوبية ينصبّ التركيز في قياس التأثيرات اللغوية وتحديد كميتها مثل تكرارات الكلمات أو العبارات في المدونات أو الاختلافات بين هذه التكرارات في المدونات الفرعية المختلفة. تتضمن قياسات حجم التأثير مجموعة واسعة من القياسات الوصفية (Cumming 2012: 39 و Kirk 2005: 2). وانظر: البحث ٤، ٨ لمزيد من المناقشة).

جدول (٤، ٨). أمثلة على نصوص من الكتابة الأكاديمية وأدب الغموض.

أدب الغموض (BE06-L10)	الكتابة الأكاديمية (BE06-J61)
"I <u>couldn't</u> believe it, I never seen anyone famous, not, like, in real life. I <u>gave</u> 'em my best service, and in those days, I <u>was</u> hot, had some moves." Foley nearly said, "You still do." But bit down and <u>wondered</u> where the hell his partner <u>had</u> got to. Probably gone for a bourbon, Shiner back. He'd return, smelling of mints, like that <u>was</u> a disguise. He <u>asked</u>, "You talk to him, to Mr. Cash?" "Not at first. I <u>was</u> getting them vittles, drinks, making sure they were comfortable and after, I dunno, an hour, Johnny said, "Take a pew little lady, get a load off." She <u>rubbed</u> here eyes, then. "He <u>had</u> these amazing boots, all scuffed but, like, real expensive, snakeskin or something, and he <u>used</u> his boot to hook a chair, pull it up beside him." She <u>touched</u> her face, self conscious, said, "I <u>didn't</u> have the scar then, still <u>had</u> some dreams. Jesus". "لم أستطع أن أصدق ذلك. لم أر أبداً أي شخص مشهور وجهاً لوجه. لقد قدمت لها أفضل خدمة أستطيعها، وفي تلك الأيام كنت أشعر بالحر فمشيت بعض الخطوات". قال فولي: "ما زلت تفعل"، لكنني عضضت على أسناني، وتساءلت أين ذهب شريكه اللعين. ربما ذهب لتناول البوربون. عاد شاينز، لقد عاد متتكبراً، ورائحته كالنعناع. سألت "أنت تتحدث معه، أعني تتحدث مع السيد كاش؟" قال جوني: "ليس من قبل، كنت أشتري له طعاماً وشراباً، وأؤكد من أنه مرتاح، وبعد ساعة لست متأكداً. قال جوني: "خذي كرسيّاً سيدتي الصغيرة، وأنزلي أمتعتك". "لقد كانت ترتدي هذا الحذاء المدهش، كان بالياً، ولكنه يبدو غالي الثمن، ومن جلد الثعبان أو نحو ذلك. كانت تستخدم حذاءها لشد الكرسي وتحريكه إلى جانبها." لمست وجهها وقالت: "لا توجد ندبة على وجهي. يا إلهي، يبدو أنني مازلت أحلم".	Immigrants contribute significantly to the overall economic performance of their host economies. It is therefore not surprising that a large literature is concerned with the earnings mobility of the foreign born population, both in isolation, as well as in comparison with those who are native born. But immigrants have not only an immediate effect on wealth accumulation and earnings and skill composition. They transmit their earnings status, as well as socio-economic and cultural characteristics to the next generation. The economic adjustment process within the immigrant's own generation has long been recognised as an important step in understanding the economic effects of immigration. For understanding the long term consequences of immigration, assessment of intergenerational mobility in immigrant communities is perhaps equally important. يسهم المهاجرون إسهاماً كبيراً في الأداء الاقتصادي العام لاقتصاداتهم. لذلك ليس من المستغرب أن تعني دراسات كبيرة بحركة أرباح السكان المولودين بالخارج، سواءً على حدة أم بالمقارنة مع أولئك الذين يولدون في بلادهم. لكن المهاجرين ليس لهم تأثير مباشر على تراكم الثروة والأرباح وتكوين المهارات فقط، فهم ينقلون حالة أرباحهم وكذلك الخصائص الاجتماعية والاقتصادية والثقافية إلى الجيل القادم. لطالما اعترف بعملية التكيف الاقتصادي داخل جيل المهاجرين بوصفها خطوة مهمة في فهم الآثار الاقتصادية للهجرة. لفهم العواقب طويلة الأجل للهجرة قد يكون تقييم الحراك بين الأجيال في مجتمعات المهاجرين بالقدر نفسه من الأهمية.

لنراجع سريعاً مثلاً مختصراً. عند المقارنة بين مجموعتين، أي: مدونتين فرعيتين، يُستخدم في الأغلب مقياس حجم التأثير (كوهن-د) للفرق بين استخدام الزمن الماضي في الكتابة الأكاديمية وأدب الغموض (انظر: الجدول ١، ٨ للأشكال في BE06). قيم الناتج هي على النحو الآتي: (3.26-)

ومجال الثقة ٩٥٪ [-2.61, -3.9]، والتأثير كبير للغاية، ويشير الرقم السالب إلى أن التكرار في مدونة أدب الغموض أكبر من التكرار في مدونة الكتابة الأكاديمية. يُسلط الجدول (٤، ٨) الضوء على الأزمنة الماضية، وقد نوقش تفسير أحجام التأثير بالتفصيل في المبحث (٤، ٨).

المبدأ السادس: متابعة أفضل الممارسات في المجال مع المراجعة النقدية للممارسة الإحصائية في هذا المجال واتباع النماذج الجيدة

من المهم إدراك كيفية تطبيق المبادئ الإحصائية العامة في تحليل المدونات الحاسوبية التي شُرحت في الفصل الأول، واتباع أفضل الممارسات التحليلية لها. يسهم هذا الكتاب في المراجعة العامة لأفضل الممارسات، ويقدم اقتراحات محددة لاستخدام الإحصائيات في بحوث لغويات المدونة الحاسوبية. إذا كنت مبتدئاً في هذا المجال فاتّبع الممارسة الثابتة للتحليل الإحصائي. ابحث عن مقالة منشورة في مجلة ذات سمعة طيبة تتناول سؤالاً بحثياً شبيهاً بسؤالك البحثي، وحاول تطبيق الإجراءات والقياسات الإحصائية نفسها على بياناتك، ثم فكر فيما إذا كانت هذه الطريقة مفيدة للإجابة عن سؤال بحثك أم لا. ينبغي ألا تتوقف عند هذا الحد، فعندما تتحسن كفاءتك في التقنيات الإحصائية فيلزمك تقييم الممارسة الحالية بأسلوب نقدي وابتكاري، وقد يكون ذلك بالاطلاع على استخدام الإحصاءات في تخصصات أخرى. يتطلب هذا منك قراءة بحوث خارج مجالك، ومن المهم في هذا المقام أن ندرك أن استخدامات الإحصاء في مجالات النخبة العلمية ليست كلها متبعة لأفضل الممارسات. فعلى سبيل المثال، كان اختبار كاي تربيع أو اختبار الاحتمالية اللوغارتمية (اختبار نسبة الاحتمال) من الممارسات المتفق على قبولها في لغويات المدونة الحاسوبية لأي مقارنة تقريباً بين مدونتين أو أكثر، ومع ذلك انتُقدت مؤخراً واقتُرحت مقاييس بديلة لإجراء مقارنات أكثر جدوى (Brezina & Meyerhoff 2014; Lijffijt et al. 2016).

المبدأ السابع: الرسومات: تمثيل البيانات لتحديد الأنماط

الصورة تساوي ألف كلمة كما يقال في كثير من الأحيان. يمكن أن يساعدنا التمثيل الفعال على اكتشاف أنماط وعلاقات مهمة في البيانات. المبدأ الأساسي للتمثيل الجيد سير، ويعتمد على التركيز في عرض البيانات بدلاً من مجرد إضافة (الروتوش البصرية). اجعل العرض غنياً بالمعلومات

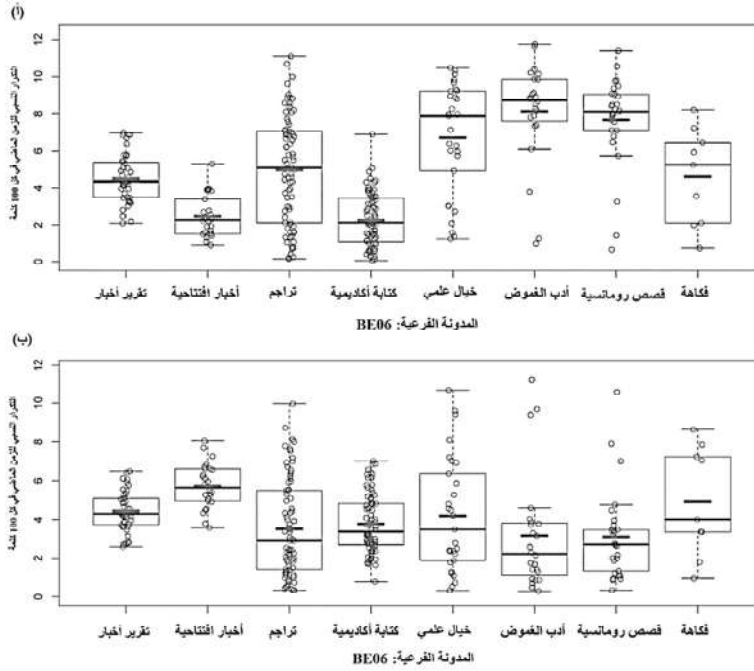
للسماح بملاحظة العلاقات بين المتغيرات (Tufte 2006). يمكن إعادة تقديم الشكل (٢، ٨) كمخططٍ رسم مربع، وهما أكثر ثراءً من الناحية المعلوماتية، ويكون واحدًا للزمن الماضي، والآخر للزمن المضارع (الشكل ٣، ٨). تُظهر مخططات رسم المربع النزعات المركزية (أي: المتوسطات والوسائط) في كل مدونة فرعية وكذلك النصوص الفردية المعروضة على شكل دوائر. لا توجد نقطة بيانات مفقودة مما يوفر لنا الصورة الكاملة، ويسمح لنا بمشاهدة توزيع القيم الفردية بما في ذلك القيم المتطرفة ما عدا الاتجاه العام.

من المهم أن تضع في حسابك أن التمثيل لا يناسب جميع الحالات، ففي بعض الأحيان تكون الجداول أو العبارات السهلة التي تصف البيانات أكثر فاعلية، وكما أوضح توفت نقلًا عن أد رينهاردت Ad Reinhardt في قوله: "إن كانت الصورة لا تعني ألف كلمة فهي صورة بائسة" (Tufte 1997: 119).

المبدأ الثامن: تسليط الضوء على أوجه التشابه والاختلاف: تقديم تفسير متوازن لاستخدام اللغة لا عجب عندما نبحث عن الاختلافات بين المدونات أو بين الاستخدامات المختلفة لكلمة أو مركب أو عبارة أن نجد عادةً القليل من الصورة الاختلافية. كان معظم التفكير المنهجي متحيزًا في البحث إلى أوجه الاختلاف وإلى تجاهل أوجه التشابه، وذلك ليس في لغويات المدونة الحاسوبية فقط بل في كل التخصصات أيضًا.

لا يُبلَّغ في الأغلب عن أوجه التشابه والتأثيرات الصفيرية كالتأثيرات الصغيرة أو غير الدالة أو المعنوية إحصائيًا على نحو كافٍ في الدراسات (انظر: مناقشة تحيز النشر في المبحث ٣، ٨). يأخذنا التركيز في الاختلافات في لغويات المدونة الحاسوبية إلى الإفصاح عن الكلمات المفتاحية وتجاهل كلمات الإيقاف والكلمات المستقرة عبر المدونات مثلًا (Baker 2011؛ انظر: المبحث ٤، ٣) أو التركيز على الاختلافات اللغوية الاجتماعية مثل الاختلافات بين الجنسين، وتجاهل أوجه التشابه العميقة في استخدام اللغة (Baker 2014؛ انظر أيضًا: الفصل السادس). أبرزنا حتى الآن في مثال مجموعة البيانات بالجدول (١، ٨) الاختلافات الأساسية بين الكتابة الأكاديمية وأدب الغموض. عند النظر إلى الشكل (٣، ٨) يمكننا أن نرى أيضًا نمطًا قويًا من أوجه التشابه بين افتتاحيات الأخبار والكتابة الأكاديمية عند النظر في استخدام الزمن الماضي أو بين أغلبية الأوعية أو السجلات اللغوية عند

النظر إلى استخدام الزمن المضارع. الرسالة المنهجية باختصار هي أننا نسعى دائماً إلى توفير تفسير متوازن للبيانات عن أوجه التشابه كما هو الحال عن أوجه الاختلاف.



شكل (٨، ٣). مخطط رسم المربع في أنواع مكتوبة مختلفة BE06: (أ) الزمن الماضي (ب) الزمن المضارع.

المبدأ التاسع: التداخل بين الإحصائيات واللغويات وتقديم تحليل إحصائي قوي يستند إلى النظرية اللغوية الاجتماعية

لننظر في السؤال المنهجي الآتي: أي من هذه الأمور أسوأ في أبحاث المدونة: اللغويات بدون إحصائيات أم الإحصائيات بدون اللغويات؟ كل خيار من هذه الخيارات الافتراضية يخفق في أحد الجوانب الأساسية. تفتقر اللغويات بدون إحصائيات إلى أدوات فعالة لتحليل كميات كبيرة من البيانات اللغوية، أما الإحصائيات بدون لغويات فيمكن أن تتحول بسهولة إلى تمرين ساذج في تحطيم الأرقام دون ارتباط بالواقع اللغوي والاجتماعي. علينا أن نتذكر أن الأبحاث اللغوية ليست

كلها كمية، ويمكن أن يجلب البحث النوعي رؤى مهمة في استخدام اللغة. يطرح اختيار النصوص والأمثلة للتحليل المتعمق السؤال عن مدى نموذجية هذه الأسباب ولماذا اخترناها (Baker et al. 2008). يمكن أن تُوفّر الطرق الإحصائية طريقة مبدئية لاختيار الأمثلة لمزيد من التحليل المتعمق، وأن تساعدنا على تجنب الالتقاطية cherry picking المتمثلة في اختيار الأمثلة التي تناسب أفكارنا المسبقة فقط. ينبغي أن نتذكر أن العلاقة المفيدة والمتبادلة بين الإحصائيات واللغويات هي وحدها التي يمكن أن توفر نتائج مفيدة، إذ تقدم الإحصائيات تقنيات لتقليل فجوات المشكلة وتركيز عنايتنا بالأنماط القياسية غير المعتادة لاستخدامات اللغة التي تحتاج إلى تفسير لغوي (راجع McEnery & Baker 2017).

المبدأ العاشر: المصطلحات المتخصصة واستخدام المصطلحات والتسميات الإحصائية التي تساعد على التعبير عن الأشياء بوضوح مع محاولة تجنب المصطلحات غير الضرورية

قد يجهل كثير من الناس علم الإحصاء الذي نحاول فيه عادةً تصوير العلاقات المعقدة في البيانات باستخدام التعبيرات الرياضية، وهو ما قد يتدمر منه كثير من الأشخاص. استبدلت الرموز الرياضية في هذا الكتاب قدر الإمكان بالمصطلحات العادية غير المتخصصة، فمثلاً يُعبّر عن $\sum_{i=1}^{10} i$ بمجموع كل الأعداد الصحيحة من 1 إلى 10، وفي الحل الرياضي سيعطيك كلا التفسيرين الرقم 55، ومن المهم استعمال المصطلحات السهلة غير المتخصصة عند الإفصاح عن نتائج التحليلات الإحصائية، وقد أجرينا هذا التبسيط في مباحث: (التقرير الإحصائي).

(٨، ٣) التحليل البعدي: التوليف الإحصائي لنتائج الأبحاث

فكّر...

قبل قراءة هذا المبحث فكر في الموقف الآتي:

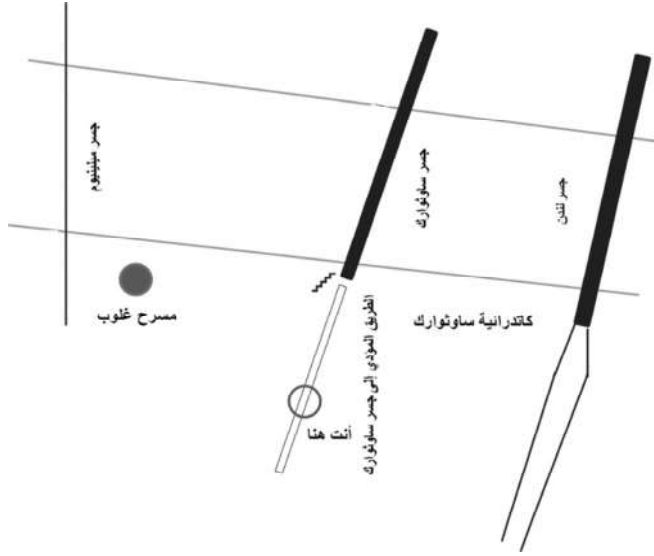
أنت في مدينة كبيرة مثل لندن وتبحث عن مسرح، وليكن مسرح شكسبير غلوب Shakespeare's Globe. ليس لديك خريطة، لذلك عليك أن تسأل المارة. يعرض الجدول (٨، ٥) خمس إجابات عن السؤال (أين يقع هذا المسرح؟). بعض الإجابات تختلف اختلافاً كبيراً. كذلك: أين تود الذهاب؟

جدول (٥، ٨). موجز الأسئلة.

الاتجاه	الإجابة	تعريف بالشخص
الشخص ١	إلى الأمام ويميناً 	لا أدري، ربما تحتاج إلى المضي قدماً ثم الانعطاف يميناً. شخص يبدو سائحاً.
الشخص ٢	إلى الأمام ويساراً 	هذا سهل للغاية. اتبع هذا الطريق ثم انعطف يساراً. حارس متجر محلي
الشخص ٣	يميناً وإلى الخلف 	انعطف يميناً ثم عد من حيث أتيت. أنا متأكد تماماً. رجل يرتدي زي صانع القبعات المجنون
الشخص ٤	إلى الأمام ويميناً 	آسف، فأنا في عجلة من أمري. امض قدماً وانعطف يميناً اليمين. رجل خارج من مبنى إداري
الشخص ٥	إلى الأمام ويساراً 	مسرح غلوب ليس بعيداً من هنا. لقد مررت به للتو. امش في هذا الطريق إلى الجسر، ثم انزل من الدرج، وخذ يسارك على طول النهر. شخص يتمشى مع كلب

يمكن فهم مثال الحالة في مهمة (فكر) على أنه تخيل موقعي يسعى إلى توضيح فكرة علمية. لمعرفة الإجابة عن سؤال بحثنا نلجأ إلى جمع البيانات وتقييمها. يحصل ذلك في العلوم مراراً وتكراراً لضمان موثوقية الإجابة. تسمى عملية تكرار الدراسة بالسؤال البحثي نفسه بـ (النتائج) ولكن بمجموعة بيانات مختلفة. يتبين النتائج في مثالنا بسؤال المارة المختلفين السؤال نفسه، ونحصل في العلوم على إجابات مختلفة من دراسات مختلفة، وتكون في الأغلب طفيفة جداً، والمهم من أجل تطوير مجالنا عموماً هو أن نكون قادرين على رؤية الصورة الأكبر، لذلك نحن بحاجة إلى وضع النتائج الفردية معاً وفهمها فهماً كلياً. تُسمى التقنية الإحصائية التي تفعل ذلك (التحليل البعدي) meta-analysis، وهو إجراء كمي للتوليف الإحصائي لنتائج البحث، ويعتمد على الجمع بين التأثيرات المبلغ عنها في العديد من الدراسات الفردية وبين حساب ملخص التأثير (Borenstein 2009).

يكمن المغزى من التحليل البعدي في جمع الأدلة المنشورة في شتى الدراسات لزيادة الثقة عن الاستنتاج التلخيصي الناتج، ويمكن تقدير هذه الزيادة باستخدام مجال الثقة. للعثور على مسرح غلوب في مثالنا نحتاج إلى تقييم الإجابات الفردية من حيث مدى اليقين والموثوقية، مع ضرورة فهم الفرق المفهومي المنفصل بين اليقين والموثوقية، ثم اتخاذ قرار بشأن المسار الذي ينبغي اعتماده. الطريق الذي سيأخذنا إلى غلوب في الحالة الموضحة في مهمة (فكر) هو إلى الأمام ثم إلى اليسار (الشكل ٤، ٨).



شكل (٤، ٨). البحث عن بشرح غلوب.

إذن، كيف يعمل التحليل البعدي على أرض الواقع؟ هناك ثلاث خطوات أساسية ينبغي اتباعها؛ الأولى: تحديد الدراسات ذات الصلة، والثانية: استخراج أو ترميز المعلومات ذات الصلة من الدراسات، والثالثة: التوليف الإحصائي.

الخطوة الأولى: تحديد الدراسات ذات الصلة

تحدد هذه الخطوة مجال البحث الذي نرغب في بحثه في الدراسة البعدية، ويتأتى ذلك بسؤال بحثي يُجاب عليه مراراً في دراسات متعددة باستخدام مدونات مختلفة. يمكن أن يكون سؤال البحث أعمّ أو أكثر تحديداً، ومن ثمّ يمكن تحديد درجة دقة الدراسة البعدية. فعلى سبيل المثال، سيكون السؤال العام للغاية هو: هل هناك تأثير للجنس في استخدام اللغة؟ بتوجيه من هذا السؤال البحثي ستكون قادرين على تحديد المئات من الدراسات ذات الصلة، لكن ربما تكون الإجابة عامة للغاية وغير محددة، فتُظهر بعض الميزات اللغوية أنها قائمة على النوع الاجتماعي، في حين أن بعضها الآخر لا يُظهرها، وهذا يمكن تقديره على أنه ملخص عام جداً. أما إن حددنا مجال البحث باستخدام سؤال بحث معين مثل الآتي -مثلاً-: هل هناك تأثير للجنس على استخدام الضمائر؟ فيمكننا الاستفادة من شيء أكثر تحديداً بل ربما أكثر إثارة للاهتمام، وعدد الدراسات ذات الصلة سيكون أقل بكثير. من المهم

من الناحية العملية تحديد معايير الاشتغال تحديداً صريحاً للدراسات كالاستفسارات الآتية: ما المتغيرات اللغوية والمتغيرات المستقلة التي نبحث عنها، وما متطلبات تصميم البحث والإطار الزمني ذي الصلة للدراسات التي نشرت مثلاً منذ عام 1990 (3-161: Wilson 2009).

بمجرد تحديد مجال البحث نحتاج إلى البحث في جميع المجالات العلمية ذات الصلة، وفي قواعد البيانات المتاحة، كباحث جوجل العلمي، وملخصات المجالات اللغوية LLBA، والأطروحات في قاعدة بيانات EThOS و proQuest، وغيرها، وذلك بحثاً عن دراسات منشورة أو غير منشورة تُجيب عن سؤال البحث وتفي بمعايير الاشتغال. الهدف من ذلك هو جمع الأدلة المتوافرة والمتعلقة بسؤال البحث. يرجع سبب اهتمامنا بالدراسات غير المنشورة مثل أطروحات الدكتوراه إلى أهمية تقليل تأثير ما يسمى تحيز النشر *publication bias* (مثل Kepes et al. 2014) الذي يتأثر بمعامل التأثير، ولا يفصح عن تأثيره بـ: النتائج الصفريّة *null results*. ينطبق هذا على الدراسات المنشورة التي تكون غير دالة على تأثير قوي على الرغم من متانتها. إذن، فإبراز مثل هذه الدراسات أمر حاسم للتمكن من رؤية الصورة كاملة، إذ ما يفصح عنه عادةً هو النتائج التي لا تمثل سوى قمة جبل الجليد مع كتلة غير معروفة من التأثيرات الأصغر المخبأة تحت الماء (Rothstein & Hopewell 2009).^(١)

المشكلة المتعلقة بتحيز النشر هي جودة الدراسات المختارة للتحليل البعدي. لتجنب تحيز النشر عند اختيار الدراسات نريد أن نلقي بالشبكة على أوسع نطاق ممكن، ولكننا نريد أن نتأكد من أن ما تصيده الشبكة هو دراسات سليمة منهجياً (الأسماك بدلاً من حطام السفينة). في التحليل البعدي كما هو الحال في البحث الأولي ينبغي أن يكون مبدأ GIGO (مُدخلات خاطئة تؤدي إلى مخرجات خطأ، انظر: المبحث ٢، ٨) بمنزلة تذكير بأننا بحاجة إلى إنعام النظر في الأدلة المتاحة عن جودة الدراسات الأولية قبل استخدامها في تحليلاتنا. يناقش فالتاين Valentine (٢٠٠٩) عددًا من جوانب جودة الدراسات البحثية بما في ذلك الصلاحية الداخلية، أي: التحكم في المتغيرات ذات الصلة والقياس الدقيق، والصلاحية الخارجية؛ أي: تعميم الدراسة، وبناء الصلاحية، أي: قياس ما نريد قياسه، وصلاحية الاستنتاج الإحصائي، أي: الدقة الإحصائية. في مجال لغويات المدونة الحاسوبية الذي لا تزال فيه الإجراءات المنهجية في طور التطوير والتوحيد القياسي تعد جودة الدراسات التي تظهر في

(١) تجدر الإشارة إلى أن التحيز تجاه الأساليب الإحصائية ونتائجها موجود في كل الدراسات المنشورة وغير المنشورة إلى

حد ما، غير أنها في المنشورة أكبر.

التحليل البعدي مصدر قلق خاص. لن نرتاب في تلاعب المؤلفين بالبيانات عن عمد أو إخفاء النتائج غير المرغوب فيها كما قد يحدث مثلاً في الدراسات الصيدلانية أو استطلاعات الرأي على سبيل المثال. في أكثر الأحيان قد ينتهك المؤلفون عن قصد اختباراً إحصائياً أو يستخدمون تصميمًا بحثياً غير مناسب. نحيل القراء هنا على المبحث (٢, ٨) لإلقاء نظرة عامة على (عشرة مبادئ للتفكير الإحصائي) والمطبقة على أبحاث المدونة الحاسوبية بوصفها إرشادات للممارسة البحثية الصارمة.

الخطوة الثانية: استخراج المعلومات ذات الصلة من الدراسات (الترميز)

تتضمن الخطوة الثانية الحصول على البيانات ذات الصلة من الدراسات التي حددناها في الخطوة الأولى، وهذا يشمل القراءة المتأنية لمباحث الأساليب والنتائج من التقارير البحثية والمقالات والكتب وتدوين المعلومات عن المدونات (الفرعية) المستخدمة أولاً، والطريقة المنهجية ثانياً، وأحجام التأثير الملاحظة في الدراسة ثالثاً (Wilson 2009). نحتاج -على وجه الخصوص- إلى معرفة عدد الكلمات الفعلية في المدونات الفرعية المقارنة مع أهمية التركيز في الحالات التي نلاحظ على نوع النصوص سواء حوارية أم منطوقة أم مكتوبة عند تصميم البحث أو عن مرار ورود الخاصية اللغوية المعنية بالتحليل عند تصميم بحث الخواص اللغوية (المبحث ٤, ١). من المهم أيضاً تدوين نوع تصميم البحث والمتغيرات اللغوية المستقلة المدرجة والاختبارات الإحصائية المستخدمة وأي معلومات أخرى ذات صلة عن المنهجية التي قد تؤثر في النتيجة. نحتاج أخيراً إلى تدوين قياسات حجم التأثير القياسية (القابلة للمقارنة) أو المعلومات (مثل قيمة-ت أو قيمة F) التي يمكننا بها استقراء مقياس حجم التأثير (المبحث ٤, ٨). تجدر الإشارة إلى أنه ينبغي تضمين الأبحاث التي تستند إلى البيانات، بغض النظر عن عدد المجالات، أو المنشورات المتعددة التي أعدها المؤلفون أنفسهم، أو الأبحاث المتكررة التي أجراها مؤلفون مختلفون استناداً إلى مجموعة بيانات مكرورة تسهم بمعلومة واحدة فقط في التحليل البعدي^(٢). وفيما يأتي ورقة ترميز ممكنة يمكننا استخدامها لتسجيل المعلومات من الدراسات الفردية.

(٢) إذا تعاملنا مع النتائج (دراسة أجراها مؤلفون مختلفون باستخدام المدونات نفسها وسؤال البحث ذاته) فقد نريد مقارنة نتائج الدراسة الأصلية والتكرار وتقييم المنهجيات، وبناء على ذلك يمكننا اختيار الدراسة الأكثر موثوقية أو التي توفر معلومات أكثر اكتمالاً للتحليل البعدي.

العنوان الرئيسي

سؤال البحث: هل هناك تأثير للجنس في استخدام الضمائر؟

الأحرف الأولى من المرمز: VB

الدراسة: (2008) Newman, M. L., Groom, C. J., Handelman, L. D., & Pennebaker, J. W.
الاختلافات بين الجنسين في استخدام اللغة: تحليل 14000 عينة النص. عمليات الخطاب، 45 (3)،
211-236.

الجودة: عالية ☒ مقبولة ☐ منخفضة ☐

ملاحظات:

المدونة

الاسم: غير متوفر - مدونة انتهازية جمعت من عدة دراسات نفسية

الرموز: 4570000

الحالات: الذكور = 5971؛ الإناث = 8353

التمثيلية: اللغة من مهمات مختلفة في التجارب النفسية. 93٪ مكتوبة، 7٪ منطوقة، 3/2 من

المشاركين في سن طلاب الجامعة، مدة الجمع 1980-2002

الطريقة

تصميم البحث: نص/ متحدث فردي ☒ خاصية لغوية ☐ مدونة كاملة ☐

الاختبار الإحصائي: اختبار-ت لـ: ويلتش Welch

المتغيرات اللغوية ذات الصلة: جميع الضمائر الشخصية

المتغيرات اللغوية الأخرى: مجموعة من المتغيرات المعجمية المحددة تلقائياً

المتغيرات المستقلة ذات الصلة: الجنس

المتغيرات المستقلة الأخرى: غير متوفر

النتائج: التأثير(ات) الملحوظ(ة)

حجم التأثير ES الوارد في التقرير: نعم ☒ لا ☐
إذا كانت الإجابة بنعم، فأيهما؟ كوهن-د
التأثير 1: الجنس -> كل الضمائر الشخصية
الحجم: د = 0.36 الاتجاه: أنثى +
التأثير 2: الجنس -> ضمير المتكلم
الحجم: د = 0.17 الاتجاه: أنثى +
المعلومات الإحصائية المهمة المبلغ عنها (الاختبارات، درجات الحرية، المتوسطات، الانحرافات
المعيارية، حجم التأثير): جميع الضمائر الشخصية: ذكر: المتوسط = 12.69، الانحراف المعياري =
4.63؛ أنثى: المتوسط = 14.24، الانحراف المعياري 4.06
...
ملاحظات

ينبغي استنتاج معلومات الدراسة من التقرير أو استقراؤها، وقد يشمل ذلك عنصرًا من عناصر الذاتية، فمن الممارسات الجيدة تكرار تشفير ٢٠٪ من البيانات وحساب إحصائية درجة الاتفاق بين الترميز (المبحث ٥، ٣)، وهذا مهم أيضًا في حالة وجود مَرَمَزَات متعددة.

لنلقي الآن نظرة على مثال لخمس دراسات تشير إلى تأثير الجنس في استخدام الضمائر I و me و my و you و your إلخ. يوضح الجدول (٦، ٨) تفاصيل هذه الدراسات المستخرجة من جداول الترميز مع التركيز في المعلومات الإحصائية المبلغ عنها. ويُقدم بعض المعلومات الإحصائية المباشرة أو التي تحتاج إلى مزيد من الاستنتاجات (المبحث ٤، ٨). لاحظ أنه قد أُستبعدت الدراسة الرابعة والسادسة من الجزء الكمي للتحليل البعدي، وقد استبعدت الدراسة الرابعة بسبب عدم اكتمال المعلومات الإحصائية. من الممارسات الجيدة في مثل هذه الحالات الاتصال بمؤلفي التحليل لطلب الحصول على معلومات إضافية. ينبغي أيضًا تشجيع مؤلفي الدراسات الأولية على استخدام معايير إعداد التقارير (انظر: مباحث التقرير الإحصائي في هذا الكتاب) التي توفر معلومات كافية لتقييم التحليل البعدي للنتائج وتوليده. استُبعدت الدراسة السادسة لأنها تستخدم تصميمًا بحثيًا غير متوافق (تصميم المدونة بأكملها).

جدول (٦، ٨). دراسات رُوِّجَت للتحليل البعدي.

الدراسة	المعلومات المقدّمة	المعلومات المستنتجة
١ Newman et al. (2008)	المدونة: ن _{ذكر} = 5971؛ ن _{أنثى} = 8353 النتيجة (الضماير): كوهن-د = 0.36	
٢ Argamon et al. قصص- (2003)	المدونة: ن _{ذكر} = 179؛ ن _{أنثى} = 179 النتيجة (الضماير): ذكر: المتوسط = 282؛ حجم التأثير = 12 أنثى: المتوسط = 390؛ حجم التأثير = 19	كوهن-د = 0.508
٣ Argamon et al. روايات - (2003)	المدونة: ن _{ذكر} = 123؛ ن _{أنثى} = 123 النتيجة (الضماير): ذكر: المتوسط = 860؛ حجم التأثير = 18 أنثى: المتوسط = 977؛ حجم التأثير = 18	كوهن-د = 0.586
٤ Argamon et al. (2007)	المدونة: ن _{ذكر} = 25065؛ ن _{أنثى} = 21682 النتيجة (الضماير): ذكر: المتوسط = 9.84؛ أنثى: المتوسط = 11.97؛	لا يمكن حساب كوهن-د لأن التنوع (الانحراف المعياري) لم يُبلغ عنه
٥ Colley & Todd (2002)	المدونة: ن _{ذكر} = 24؛ ن _{أنثى} = 30 النتيجة (you, yours): ذكر: المتوسط = 0.78؛ أنثى: المتوسط = 1.33؛ ف (1, 56) = 7.69	كوهن-د = 0.759
٦ Rayson et al. (1997)	المدونة: ن _{ذكر} = 536؛ ن _{أنثى} = 561 النتيجة (الضماير): ذكر = 13.37%؛ أنثى = 14.55% = χ^2 1016.27	تصميم بحث غير متوافق

الخطوة الثالثة: التوليف (التركيب) الإحصائي

تجمع الخطوة الأخيرة من التحليل البعدي المعلومات من شتى الدراسات، وتصدر تقريراً عن حجم التأثير (الدمج) الناتج وعن مجال الثقة. يأخذ التوليف الإحصائي في أيسر أشكاله مدخلات قياس حجم التأثير (على الأخص: كوهن-د) وعدد الحالات (النصوص/ المتحدثون) في كل مدونة فرعية (الجدول ٧، ٨). تُقاس أحجام تأثير الدراسات الفردية في الأغلب لإعطاء مزيد من الأهمية للدراسات المأخوذة مع المدونات الأكبر لأنها أقرب إلى قيمة المجموع اللغوي الحقيقية، وتقاس اختياريّاً أيضاً للدراسات ذات الجودة الأعلى باستخدام مؤشر الجودة. يُستخدم عادة التباين العكسي في القياس الذي يأخذ في الحسبان عدد الحالات، إذ كلما زاد عدد الحالات (النصوص/ المتحدثون) زادت الأهمية.

جدول (٨، ٧). بيانات الإدخال للتحليل البعدي البسيط.

الدراسة	كوهن-د	الحالات في المدونة الفرعية 1 (ن1)	الحالات في المدونة الفرعية 2 (ن2)	التباين العكسي
Newman et al. (2008)	0.36	5,971	8353	3428.0
Argamon et al. (2003) - non-fiction	0.508	179	179	86.7
Argamon et al. (2003) - fiction	0.586	123	123	59.0
Colley & Todd (2002)	0.759	24	30	12.4

يُحسب التباين العكسي لكوهن-د بوصفه مقياساً نموذجياً في التحليل البعدي عند مقارنة نوعين من النصوص أو مجموعات المتحدثين باستخدام تصميم البحث للنص أو المتحدث الفردي على النحو الآتي^(٣):

$$\text{التباين العكسي} = \frac{1}{\frac{2}{2\dot{n}+1\dot{n}} + \frac{1}{(2\dot{n}+1\dot{n}) \times 2}}}$$

(٨، ١)

على سبيل المثال، يُحسب التباين العكسي للدراسة الأولى (Newman et al. 2008) على النحو الآتي:

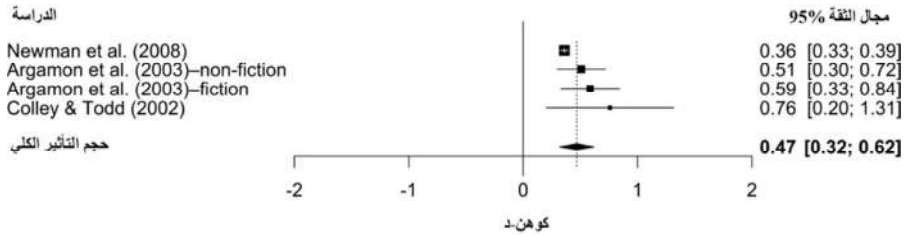
$$\text{التباين العكسي} = \frac{1}{\frac{2}{0.36} + \frac{8,353+5,971}{(8,353+5,971) \times 2}}}$$

(٨، ٢)

تحتوي العملية التي تجمع أحجام التأثير على خيارين: نموذج التأثيرات الثابتة ونموذج التأثيرات العشوائية. يفترض نموذج التأثيرات الثابتة أن جميع الدراسات في التحليل البعدي هي نتائج متطابقة استخدمت التصميم والطرق والمتغيرات نفسها وما إلى ذلك وتختلف في المدونات المستخدمة فقط. عندما يكون التنوع الوحيد الذي ينبغي مراعاته هو التغير من عينة إلى أخرى والخطأ في أخذ العينات فإن نموذج التأثيرات الثابتة يجمع بسهولة بين التأثيرات الموزونة لإنتاج ملخص التأثير. أما التأثير العشوائي

(٣) تتوفر معادلات لحساب التباين العكسي لتدابير حجم التأثير الأخرى في دراسة شاديش وهادوك (Shadish & Haddock 2009: 264)، ولا داعي للقلق بشأن تفاصيل المعادلات لأن أداة التحليل البعدي في أدوات لانكستر الشبكية تحسبها تلقائياً.

يفترض مصادر متعددة للتنوع عندما تَبَحَث الدراساتُ في بنيات مختلفة قليلاً باستخدام خيارات منهجية مختلفة إلى حد ما، وهو ما يحدث في الأغلب. لذلك يُوصى بالتأثير العشوائي بوصفه خياراً افتراضياً. عند استخدام نموذج التأثيرات العشوائية للبيانات الواردة في الجدول (٨، ٧) نحصل على ملخص تأثير قيمته (0.47) بمجال ثقة ٩٥٪ [0.32, 0.62].^(٤) يقدم الشكل (٨، ٥) ملخصاً مرئياً للتحليل البعدي، يستخدم المخطط الغابي forest plot الذي يُقدّم ملخصاً للتحليل البعدي بعرض أحجام تأثير الدراسات الفردية (المربعات المملوءة) ومجالات ثقتها (الشُّعيرات) وكذلك حجم التأثير الكلي (شكل معين)، ويمثل حجم المربع مقدار أهمية الدراسة الفردية في التحليل.



شكل (٨، ٥). مخطط غابي للتحليل البعدي لأربع دراسات.

بالإضافة إلى التحليل البعدي السهل يمكن استخدام نماذج أكثر تعقيداً تراعي ما يسمى المتغيرات المعدلة moderator variables التي تُعدّ متغيرات إضافية تُحدّد في الخطوة الثانية من عملية الترميز بوصفها ذات تأثير في أحجام تأثير الدراسات، ويمكن دمج هذه المتغيرات في نماذج التحليل البعدي. لمزيد من المعلومات عن هذا النوع من التحليل انظر في دراسة شميدت وهنتر (Schmidt & Hunter 2015: 38Iff). من المهم أن نلاحظ أن النماذج الأكثر تعقيداً من التحليل البعدي تكون ممكنة فقط إذا كانت هناك دراسات كافية مع المتغيرات المعدلة التي أُبلغ عنها لغرض التوليف الإحصائي. حريّ أن ندرك أهمية تراكم العلوم، ويوفر التحليل البعدي أدوات للتوليف الإحصائي الفعال لنتائج الأبحاث، ونجاح هذا الأسلوب التوليفي يعتمد اعتماداً مباشراً على جودة دراسات المراجعة للتوليف، وإدراج مجموعة واسعة من التقارير البحثية.

(٤) قيمة نموذج التأثيرات الثابتة هي (0.37) ومجال الثقة ٩٥٪ [0.34, 0.40].

التقرير الإحصائي: التحليل البعدي

١ - ما يجب الإفصاح عنه

التحليل البعدي هو إجراء معقد يتطلب الإفصاح المفصل عن القرارات المنهجية في كل خطوة، الأولى: تحديد الدراسات ذات الصلة، والثانية: الترميز، والثالثة: التوليف الإحصائي. من الممارسات الجيدة أيضًا توفير جدول الترميز وجدول البيانات أو قاعدة البيانات مع الدراسات المرمزة التي يستند إليها التحليل البعدي في ملحق الدراسة أو في موقع شبكي (للاطلاع على مناقشة تفصيلية عن الإفصاح عن التحليل البعدي، انظر في دراسة روزنتال 1995 Rosenthal).

٢ - مثال الإفصاح في التقرير

أجري التحليل البعدي مع الدراسات التي أفصحت عن تأثير الجنس بوصفه متغيرًا مستقلًا على استخدام الضمائر (المتغيرات اللغوية) في اللغة الإنجليزية. بُحث في المصادر المنشورة وغير المنشورة (أطروحات الدكتوراه)، وحُدِّدت الدراسات وفقًا لجدول الترميز مع ٢٠٪ من المُدخلات المحشاة من مرمزين (ملاءمة المُقدَّر البيني: ٩٥٪). في البداية، حُدِّدت ست دراسات، وأمكن استخدام أربعة منها فقط في التحليل البعدي (انظر: الملحق للحصول على البيانات المشفرة). قُدِّرَت أحجام تأثير الدراسات الفردية بالتباين العكسي، واستُخدم نموذج التأثير العشوائي. حُدِّد هذا الإجراء ملخص تأثير قيمته $d = 0.47$ بمجال ثقة ٩٥٪: [٠.٣٢, ٠.٦٢]. يمكن الاطلاع على ملخص مرئي للتحليل البعدي من المخطط الغابي في الشكل (٨, ٥).

(٨, ٤) أحجام التأثير: دليل للاستخدام الصحيح

فكّر...

قبل قراءة هذا المبحث فكر في الموقف الآتي:

كنت تستعد لنصف ماراثون وتريد تقييم تأثير التدريب. قبل التدريب تجري مسافة ٢٠٠ متر في الدقيقة الواحدة. بعد التدريب تجري ٢٢٠ ياردة في الوقت الزمني نفسه. هل أثمر التدريب؟

للإجابة عن السؤال من مهمة (فكّر) نحتاج أولاً إلى تحويل القيم إلى وحدة مشتركة. إن كنت معتاداً على الوحدات المترية فأنت بحاجة إلى تحويل الياردات إلى أمتار: (220 ياردة $\cong$ 201.2 متر). أما

إن كنت معتادًا على الياردات فإنك تحولها إلى الأمتار: (200 متر $\cong$ 218.7 ياردة). السؤال المطروح هو: هل يُعدّ التحسن بمقدار 1.2 متر (1.3 ياردة) معتبرًا أم لا؟ من الواضح أن في ذلك تحسنًا، لكن من النظرة الأولى يبدو تافهًا إلى حد ما، فهل يسرنا إنفاق كثير من الوقت والجهد في التحسن بأكثر من متر واحد أو ياردة واحدة؟ إذا أخذنا في الحسبان الصورة الأكبر في نصف الماراثون (21098 متر أو 23073 ياردة) فإن تأثير التدريب يعني أنك قد حسّنت تدريب الجري بمقدار 38 ثانية.

يوضح هذا المثال التفكير الأساسي في التأثيرات وأحجام التأثيرات التي يمكننا تطبيقها أيضًا على البيانات اللغوية. يقيس حجم التأثير، كما علمنا، المتغيرات اللغوية الملحوظة والاختلافات والتغيرات في تكراراتها. من المهم أن نفهم جيدًا معنى القياسات بالضبط وكيفية تفسيرها. في مثال نصف الماراثون يمكننا أن نبين بسهولة ما نعنيه بمتر واحد أو ياردة واحدة برسم خط يمثل وحدة القياس. نحن بحاجة إلى أن ندرك أن القرار المتعلق بالأهمية العملية للتأثير ليس له علاقة بالإحصاءات، فقرارنا الشخصي في مثالنا هنا عما إذا كان التحسن بفارق ٣٨ ثانية في نصف الماراثون يستحق وقت التدريب والطاقة أم لا. لاتخاذ قرار مستنير ينبغي وضع الثواني الـ ٣٨ في سياقها ومقارنتها بأوقات المتسابقين الآخرين.

يركز هذا المبحث على استخدام قياسات حجم التأثير في لغويات المدونة الحاسوبية. ورد مفهوم حجم التأثير في الفصل الأول ونوقشت قياسات حجم التأثير المختلفة في أجزاء هذا الكتاب. يوضح الجدول (٨، ٨) نظرة عامة على قياسات حجم التأثير المستخدمة في هذا الكتاب بما في ذلك وصف موجز لكل مقياس، وهي تُراوح بين إحصائيات بسيطة مثل المتوسط أو تغير النسبة المئوية إلى مقاييس أكثر تعقيدًا مثل تدابير الارتباط المستخدمة لتحديد التصاحب اللغوي، وكوهن-د، وقياس حجم التأثير بتربيع إيتا η^2 ، والاحتمالية اللوغارتمية، وما إلى ذلك.

جدول (٨، ٨). قياسات حجم التأثير الواردة في هذا الكتاب.

قياس حجم التأثير	وصف موجز	هل يمكن مقارنته بمتغيرات لغوية مختلفة؟	المبحث بالكتاب
المتوسط	إحصائية وصفية أساسية تُظهر تكرار المتغير اللغوي في المدونة	لا	١، ٢

تابع جدول (٨, ٨).

المبحث بالكتاب	هل يمكن مقارنته بمتغيرات لغوية مختلفة؟	وصف موجز	قياس حجم التأثير
٧, ٣	نعم	إشارة إلى زيادة / نقصان تكرار المتغير اللغوي بين نقطتين زمنيتين	تغير النسبة المئوية (% change)
٣, ٢	نعم. لاحظ أن قيمة-ت تعتمد على حجم المدونة، ومن ثم فهي مقياس حجم تأثير إشكالي	مقاييس الاقتتان لتحديد العلاقة التصاحبية بين الكلمات	قيمة المعلومات المتبادلة، قيمة-ت، اللوج دايس، إلخ
٣, ٤	نعم	إحصائيات لتحديد الكلمات الرئيسية	بارامتر رياضي بسيط، اختلاف النسبة المئوية (%DIFF)، إلخ
١, ٣ ٣, ٢ ٦, ٣ ٧, ٣	نعم	مقياس موحد للتعبير عن الفرق بين تكرار المتغير اللغوي في مدونتين فرعيتين في وحدات الانحراف المعياري	كوهن-د (Cohen's d)
٦, ٣	نعم	نسخة قوية من مقياس كوهن-د محسوبة مع المتوسطات المرتبة والانحرافات المعيارية المقلصة	كوهن-د المتين والفعال (robust Cohen's d)
١, ٣ ٥, ٢ ٦, ٣	نعم	أشهر أسائه معامل الارتباط، ولكنه يستخدم أيضًا كمقياس موحد لحجم التأثير في المواقف المختلفة	ر (r)
٥, ٢	نعم	نسخة من ر تستخدم للبيانات الترتيبية	ر _s (r_s)
٦, ٣	نعم	نوع من الارتباط يستخدم كمقياس لحجم التأثير لاختبار مان-ويتني U	ر _{تب} : معامل الارتباط الثنائي المتسلسل للترتب (r_{tb} : rank biserial correlation)
٥, ٢ ٥, ٤	نعم	مقياس يعبر عن مقدار التنوع الذي يُحسب بمتغير مستقل (متنبئ)	ر ^٢ (r^2) معامل التحديد

تابع جدول (٨، ٨).

المبحث بالكتاب	هل يمكن مقارنته بمتغيرات لغوية مختلفة؟	وصف موجز	قياس حجم التأثير
١، ٣ ٦، ٣	نعم	مقياس حجم تأثير عام يُبلغ عنه تقليدياً باستخدام اختبار تحليل التباين	إيتا تربيع (η^2)
٤، ٣	نعم	نسبة احتمالين يحدث بهما شكلان مختلفان لمتغير لغوي في سياق معين	نسبة الاحتمال (probability ratio) تعرف أيضاً بنسبة الخطر (risk ratio) أو النسبي (relative risk)
٤، ٣ ٤، ٤	نعم	مقياس يستخدم للتعبير عن تأثير المتنبئ في ناتج نماذج الانحدار اللوجستي ونماذج التأثير الممتزج	نسبة الاحتمالية اللوغارتمية [log] odds ratio أو ln(OR)
٤، ٣	نعم	مقياس لحجم التأثير يوضح الارتباط بين المتغيرات الاسمية (البيانات الفئوية)	معامل كريمر ث

جدول (٩، ٨). تحويل حجم التأثير والاستخلاص الرياضي $4 + 2_{0.5}$

المدخل	المخرج	التحويل والاستخلاص	المثال
ر	كوهن-د	$\frac{r^2}{r^2 - 1} = د$	$ر = 0.3$ $د = \frac{0.3 \times 2}{\sqrt{2_{0.3} - 1}} = 0.629$
كوهن-د	ر	$\frac{د}{\sqrt{\frac{د}{4+د}}} = ر$	$د = 0.5$ $ر = \frac{0.5}{\sqrt{4 + 2_{0.5}}} = 0.25$
η^2 [مجموعة متشابهة، أحجام كوهن-د]	كوهن-د	$\frac{2 \times \sqrt{\eta^2}}{\sqrt{1 - \eta^2}} = د$	$0.2 = \frac{\sqrt{0.01} \times 2}{\sqrt{0.01 - 1}} = د$ $0.01 = \eta^2$
نسبة الاحتمالية اللوغارتمية [1ن كوهن-د OR]	كوهن-د	$\frac{1ن (OR) \times \sqrt{3}}{\pi} = د$	$0.9 = (OR) 1ن$ $0.5 = \frac{\sqrt{3} \times 0.9}{3.14} = د$
ت؛ 1ن؛ 2ن	كوهن-د	$\sqrt{\frac{2ن + 1ن}{2ن \times 1ن}} \times ت = د$	$ت = 3؛ 1ن = 50؛ 2ن = 25$ $د = \sqrt{\frac{25 + 50}{25 \times 50}} \times 3 = 0.73$
تحليل التباين الأحادي ف؛ 1ن؛ 2ن	كوهن-د	$\sqrt{\frac{(2ن + 1ن) ف}{(2ن \times 1ن)}} \pm = د$	$ف = 9؛ 1ن = 50؛ 2ن = 25$ $د = \sqrt{\frac{(25 + 50) \times 9}{(25 \times 50)}} \pm = 0.73$

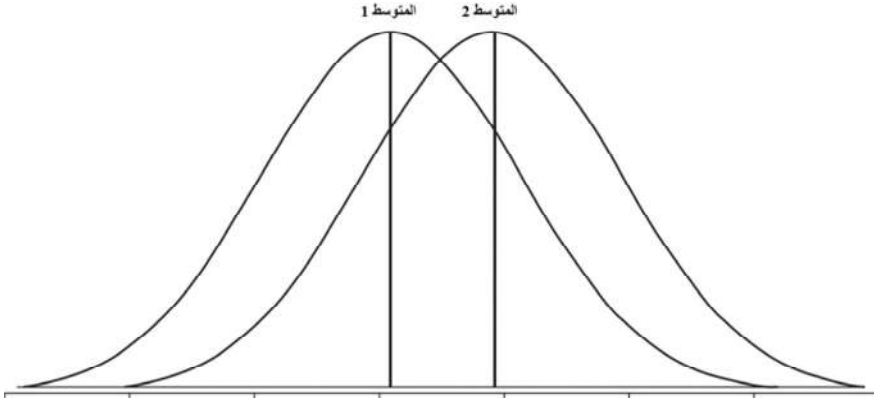
سنجيب هنا عن السؤالين الآتين. الأول: كيف ينبغي التعامل مع الإفصاح غير المتسق لأحجام التأثير؟ والثاني: كيف يمكن بالضبط تفسير قياسات حجم التأثير؟ بالنظر إلى السؤال الأول، فإنه عند الدراسات المختلفة التي قد تشير إلى مجموعة متنوعة من قياسات حجم التأثير، أو قد لا تبلغ عن هذه المقاييس على الإطلاق نحتاج في الأغلب إلى إجراء تحويلات أو استخلاص لحجم التأثير من الاختبارات الإحصائية كاختبار-ت وتحليل التباين، وأحجام العينات (ن؛ ن٢). يعد التحويل والاستخلاص مهمين عند إجراء التحليل البعدي، إذ يتعين علينا تجميع معلومات غير كاملة أو غير متسقة. يوضح الجدول (٨، ٩) التحويلات والاستخلاصات الأساسية التي يمكن إجراؤها في أبحاث المدونة (Borenstein 2009; Fritz et al. 2012).

لنتناول السؤال الثاني (كيفية تفسير قياسات حجم التأثير؟) نحتاج إلى فهم جيد لكيفية عمل مقياس حجم تأثير معين على أرض الواقع، ففي الكتب المتخصصة تُنسب قياسات حجم التأثير إلى تفسيرات قياسية، وتستند عادةً إلى توصيات كوهن (Cohen 1988) التي تُقسّم حجم التأثير الملحوظ إلى (صغير) و(متوسط) و(كبير). يعرض الجدول (٨، ١٠) ملخصًا للتفسير القياسي لأربعة من قياسات حجم التأثير المشتركة.

جدول (٨، ١٠). حجم التأثير: التفسير المعياري.

مقياس حجم التأثير	ر	كوهن-د	إيتا تربيع η^2	معامل كريمر ف [جدول 2 × 2]
صغير	0.1	0.3	0.01	0.1
متوسط	0.3	0.5	0.06	0.3
كبير	0.5	0.8	0.14	0.5

على الرغم من أن هذه التفسيرات القياسية مفيدة، فلا ينبغي تطبيقها آلياً كما أكد كوهن Cohen (١٩٨٨) مراراً وتكراراً. وعلى العكس من ذلك ينبغي أن تُراجع في كل تخصص مجموعة التأثيرات المبلّغ عنها وتُقيّم تأثيراتها العملية بأسلوب نقدي بالبحث عن كيفية تأثير وجود المتغير اللغوي المستهدف أو غيابه في القواعد النحوية أو نوع النص أو مجموعة المتحدثين أو الخطاب.



شكل (٦, ٨). مقارنة بين مدونتين فرعيتين.

ولأن أحجام التأثير جديدة نسبياً في لغويات المدونة الحاسوبية فإن بقية المناقشة ستوفر دليلاً لتفسير الأهمية العملية لقياسات حجم التأثير الموحدة. تعمل قياسات حجم التأثير الموحدة مثل كوهن-د أو (r) على مقياس موحد يختلف عن المقياس الذي تقاس به المتغيرات مثل التكرار لكل مليون. تتمثل إحدى طرق تفسير أحجام التأثير الموحدة مثل كوهن-د أو (r) أو تربيع إيتا η^2 في ربطها بمفهوم مقياس احتمالية الأفضلية الذي يمكن فهمه بالحدس على نحو أكبر. احتمالية الأفضلية **probability of superiority (PS)** هو احتمال أن يكون للمتكلم أو للنص الذي يُختار عشوائياً من المدونة الفرعية ذات قيمة المتوسط الأكبر للمتغير الهدف درجة أعلى من المتحدث أو النص الذي يُختار عشوائياً من المدونة الفرعية ذات قيمة المتوسط الأدنى. يمكن توضيح ذلك بيانياً في الشكل (٦, ٨)^(٥). يمثل المنحنيان على شكل جرس توزيع القيم في المدونتين الفرعيتين الأولى والثانية^(٦)، ويشار إلى المتوسطات بخطوط عمودية سميكة. يمكننا ملاحظة أن المنحنيات تتداخل إلى حد كبير (69% في الشكل ٦, ٨)، وهناك فرق واضح بين هذين المدونتين الفرعيتين بقيمة احتمالية أفضلية

(٥) للحصول على نسخة ديناميكية من هذا الشكل راجع تمثيل كريستوفر ماجنوسون Kristoffer Magnusson لكوهن-د على العنوان <http://rpsychologist.com/d3/cohend/>.

(٦) هذه هي الصورة المثالية، وتكون توزيعات المتغيرات اللغوية عادةً منحرفة انحرافاً إيجابياً (على سبيل المثال ذيل

يميني طويل كما في خط المؤشر التالي).

نسبتها ٧١٪ وهو ما يعني أنه في هذه النسبة من الحالات سيكون للمتغير اللغوي المعني للمتكلم أو في النص المختار عشوائياً من المدونة الفرعية الثانية قيمة أعلى من المختار عشوائياً من المدونة الأولى. ترتبط قيمة احتمالية الأفضلية (71) بقيمة كوهن-د (0.8) في الجدول (١١، ٨)، وهو تأثير كبير في التفسير القياسي.

للاطلاع على التأثير العملي للتأثيرات اللغوية الخاصة بالمدونة المعبر عنها بـ: [د، ر، η^2 (ر²)] يوفر الجدول (١١، ٨) مقارنة بين هذه القيم وقيم احتمالية الأفضلية ذات الصلة بالنسب المئوية^(٧) للدلالة على مدى أفضلية إحدى المدونات الفرعية على الأخرى من جهة ما سبق ذكره آنفاً. إضافة إلى ذلك، يوضح العديد من النقاط المرجعية من خط المؤشر مع مثال لغوي من المدونة الوطنية البريطانية. يُشجّع القراء على توفير المزيد من النقاط المرجعية للجدول استناداً إلى أبحاثهم أو الأبحاث التي عُثِرَ عليها في دراسات سابقة من أجل خلق نطاق تحليلي أدق. يمكن أن يساعد مقياس التأثيرات هذا على تحديد سياق التأثيرات التي لوحظت في الدراسات الفردية وتفسيرها.

جدول (١١، ٨). قياسات حجم التأثير: التحقق من صلاحية المدونة الوطنية البريطانية.

المثال من المدونة الوطنية البريطانية	احتمالية الأفضلية (%)	ر ² أو η^2	ر (الارتباط)	كوهن-د	
أداة التعريف (the) في مدونتين فرعيتين مختارتين عشوائياً من المدونة الوطنية البريطانية	50	0	0	0	
lovely في كلام الإناث في مقابل الذكور	53	0.002	0.05	0.1	
	56	0.01	0.1	0.2	
	58	0.022	0.15	0.3	
	61	0.038	0.2	0.4	
	64	0.059	0.24	0.5	
	66	0.083	0.29	0.6	
	69	0.11	0.33	0.7	

(٧) النظرة العامة إلى القيم [د و (ر) η^2 (ر²)] واحتمالية الأفضلية مأخوذة من دراسة فريتز وآخرين (8: Fritz et al. 2012).

تابع جدول (١١، ٨).

المثال من المدونة الوطنية البريطانية	احتمالية الأفضلية (%)	r^2 أو η^2	ر (الارتباط)	كوهن-د	
	71	0.14	0.37	0.8	
	74	0.17	0.41	0.9	
	76	0.2	0.45	1.0	
	78	0.23	0.48	1.1	
	80	0.27	0.51	1.2	
	82	0.3	0.55	1.3	
	84	0.33	0.57	1.4	
الضائير الشخصية في الكلام في مقابل الكتابة	86	0.36	0.6	1.5	
	87	0.39	0.63	1.6	
	89	0.42	0.65	1.7	
	90	0.45	0.67	1.8	
	91	0.47	0.69	1.9	
	92	0.5	0.71	2.0	
	94	0.55	0.74	2.2	
	96	0.59	0.77	2.4	
	97	0.63	0.79	2.6	
	98	0.66	0.81	2.8	
	98	0.69	0.83	3	
	99	0.72	0.85	3.2	
	99	0.74	0.86	3.4	
المبني للمجهول في الكتابة الأكاديمية في مقابل الكلام العامي	99	0.76	0.87	3.6	
	100	0.78	0.89	3.8	
	100	0.8	0.89	4	

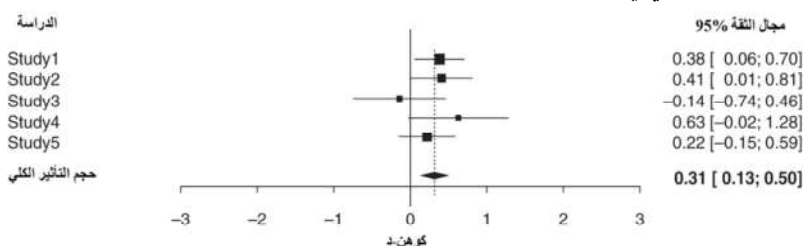
(٨، ٥) تمارين

- ١- ما أهم شيء تعلمته من هذا الكتاب؟ اكتب ذلك أدناه.
- ٢- قدم تحويلات لقياسات حجم التأثير في الجدول (١٢، ٨).

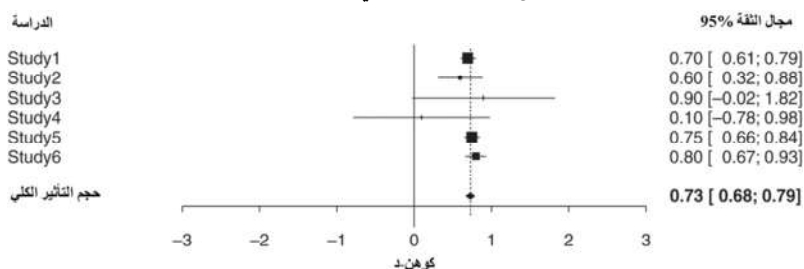
جدول (١٢، ٨). تحويلات حجم التأثير.

المدخل	الناتج (التحويل)
$r = 0.9$	كوهن-د =
كوهن-د = 1.3	$r =$
$\eta^2 = 0.05$	كوهن-د =
نسبة الاحتمال اللوغارثمية (OR) = 0.2	كوهن-د =
ت = 2؛ 1 ن = 100؛ 2 ن = 100	كوهن-د =
ف = 10؛ 1 ن = 100؛ 2 ن = 100	كوهن-د =

٣- فسر المخطط الغابي في الشكلين (٧، ٨) و (٨، ٨).



شكل (٧، ٨). مخطط غابي للمثال الأول.



شكل (٨، ٨). مخطط غابي للمثال الثاني.

تذكر الآتي

- تساعدنا الإحصائيات على التعبير عن المعلومات الكمية بدقة.
- يقدم التحليل البعدي ملخصاً إحصائياً لدراسات متعددة بالجمع بين أحجام التأثير.
- يمكن تمثيل نتائج التحليل البعدي باستخدام المخطط الغابي.

- للتعامل مع الإفصاح غير المتسق عن أحجام التأثير يمكننا تحويل قياس حجم التأثير إلى آخر أو استخلاصه.
- يمكن فهم مقاييس حجم التأثير الموحدة من جهة احتمالية الأفضلية.
- يمكن تفسير مقاييس حجم التأثير بمساعدة نقاط مرجعية، وهي تُظهر أمثلة على التأثيرات اللغوية التي يسهل تخيلها والقيم المقابلة لقياسات حجم التأثير الشائعة.

قراءات إضافية

- Borenstein, M., Hedges, L. V., Higgins, J. & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester: John Wiley & Sons.
- Brezina, V. & Meyerhoff, M. (2014). Significant or random? A critical review of sociolinguistic generalisations based on large corpora. *International Journal of Corpus Linguistics*, 19(1), 1-28.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.
- Cooper, H., Hedges, L. & Valentine, J. (eds.) *The handbook of research synthesis and meta-analysis*. New York: Russell Sage Foundation.
- Cumming, G. (2012). *Understanding the new statistics*. New York: Routledge.

أدوات لانكستر الإحصائية

- ١- يمكن استنساخ جميع التحليلات المبينة في هذا الفصل باستخدام أدوات لانكستر الإحصائية. تشمل الأدوات المتاحة لهذا الفصل:
 - حاسبة حجم التأثير.
 - حاسبة التحليل البعدي.
- ٢- يقدم الموقع أيضًا موادًا إضافية للطلاب والمعلمين.

ملحوظات ختامية

علم الإحصاء هو أداة تحليلية فعّالة. استعرض هذا الكتاب عددًا من الطرق المختلفة التي يمكن بها استخدام التقنيات الإحصائية لاستكشاف المدونات. توفر الأدلة القوية الموجودة في هذه المجموعات اللغوية الإلكترونية إمكانات لا حصر لها للبحث اللغوي والاجتماعي على حد سواء، وهو ما يوفر رؤية فريدة عن أنماط استخدام اللغة. إذا طُبّق علم الإحصاء على نحو مناسب فإنه سيُسَهِّل عملية التحليل بالعمل كعدسة تكبير يمكن بها ملاحظة الواقع اللغوي وتفاصيل الأمثلة لاستخدام اللغة وكذلك الصورة الأكبر للقواعد النحوية والمفردات والخطاب. يستلزم أن نتذكر أن العدسة الكبرى هذه ينبغي أن تكون دائمًا شفافة، فما نريد أن نلاحظه ليس الأدوات نفسها المملوءة بتقنيات إحصائية متطورة، بل البيانات اللغوية. ينبغي أن يركز تحليلنا تركيزًا أساسيًا في البيانات، وأن نُؤخذ البيانات على محمل الجد. إذا تعارضت معتقداتنا ونظرياتنا مع البيانات فلا ينبغي لنا ببساطة رفض البيانات بوصفها دليلًا غير مريح أو إخفاؤها وراء المصطلحات الإحصائية المعقدة، بل ينبغي أن نتعامل معها بالسعي إلى فهم وتفسير حقيقي للنتائج والاكتشافات. بهذه الطريقة فقط يمكن أن تكون دراستنا ذات معنى علمي حقا.

إتقان علم الإحصاء يمنحك مزيدًا من التمكن. قد يجعلك الإحصاء مثبطًا لأن الأدوات والتحليلات الإحصائية قد أصبحت أكثر تشابكًا وتعقيدًا، فالتحليل الإحصائي يتضمن العديد من المسارات. نحتاج إلى اختيار مدونة مناسبة وتقنية تحليلية فعّالة وتفسير مناسب للنتائج. تبدو هذه القرارات عادةً صعبة، وبخاصة للباحثين المبتدئين. يمكن أن يؤدي الطلب المتزايد في لغويات المدونة الحاسوبية بسبب التطور الإحصائي ونقص الموارد المناسبة لمستخدمي علم الإحصاء المبتدئين

والمتمسطين إلى الإحباط بسهولة. نأمل أن يكون هذا الكتاب موردًا يعالج هذه المشكلة بتقديم دليل للقراء لاتخاذ اختيارات مستنيرة عن الإحصاءات في تحليل اللغة.

الرسالة الرئيسية للكتاب ذات شقين. الأول: لا تتعلق الإحصائيات بمعادلات تحطيم الأرقام أو التذكر، فالحواسيب أفضل بكثير في هذه المهام من البشر، بل تتعلق بفهم المبادئ الأساسية المحورية للتحليل الكمي. الثاني: أود أن أشجع القراء على عدم السباح للتقنيات الإحصائية التجارية في السوق الإحصائية. في كل صيف، يفد عدد كبير من الطلاب من أنحاء مختلفة من العالم إلى لانكستر للتعرف على المدونات والإحصاءات في أحد الأسابيع من الفصول الصيفية. كثيرًا ما يسألني هؤلاء الطلاب عن أفضل اختبار إحصائي لاستخدامه مع المدونات، وما أفضل مقياس للتصاحب اللغوي وما إلى ذلك. عادةً أجيبهم قائلًا: يكون الأسلوب الإحصائي الأقوى هو المنطق السليم.

قائمة المراجع

- Argamon, S., Koppel, M., Fine, J. & Shmoni, A. R. (2003). Gender, genre, and writing style in formal written texts. *Text*, 23(3), 321–46.
- Argamon, S., Koppel, M., Pennebaker, J.W. & Schler, J. (2007). Mining the blogosphere: age, gender and the varieties of self-expression. *First Monday*, 12(9). http://firstmonday.org/issues/issue12_9/argamon/index.html
- Arppe, A. (2008). *Univariate, bivariate, and multivariate methods in corpus-based lexicography: a study of synonymy*. Helsinki: University of Helsinki. Available at <https://helda.helsinki.fi/bitstream/handle/10138/19274/univaria.pdf?sequence=2> (accessed 29/12/2015).
- Azen, R. & Walker, C. M. (2011). *Categorical data analysis for the behavioral and social sciences*. London: Routledge.
- Baker, H., Brezina, V. & McEnery, T. (2017). Ireland in British parliamentary debates 1803–2005: plotting changes in discourse in a large volume of time-series corpus data. In T. Säily, A. Nurmi, M. Palander-Collin & A. Auer (eds.), *Exploring future paths for historical sociolinguistics (Advances in Historical Sociolinguistics)*, pp. 83–107. Amsterdam: John Benjamins.
- Baker, P. (2009). The BE06 corpus of British English and recent language change. *International Journal of Corpus Linguistics*, 14(3), 312–37.
- (2011). Times may change, but we will always have money: diachronic variation in recent British English. *Journal of English Linguistics*, 39(1), 65–88.
- (2014). *Using corpora to analyze gender*. London: Bloomsbury.
- (2017) *American and British English: divided by a common language*. Cambridge University Press.
- Baker, P., Gabrielatos, C. & McEnery, T. (2013). Sketching Muslims: a corpus driven analysis of representations around the word ‘Muslim’ in the British press 1998–2009. *Applied Linguistics*, 34(3), 255–78.
- Baker, P., Gabrielatos, C., Khosravini, M., Krzyżanowski, M., McEnery, T. & Wodak, R. (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse & Society*, 19(3), 273–306.
- Balakrishnan, N., Voinov, V. & Nikulin, M. S. (2013). *Chi-squared goodness of fit tests with applications*. Waltham, MA: Academic Press.
- Barlow, M. (2013). Individual differences and usage-based grammar. *International Journal of Corpus Linguistics*, 18(4), 443–78.
- Baroni, M., Bernardini, S., Ferraresi, A. & Zanchetta, E. (2009). The WaCky Wide Web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43 (3), 209–26.
- Baroni, M. & Ueyama, M. (2006). Building general-and special-purpose corpora by web crawling. In *Proceedings of the 13th NIJL International Symposium, Language corpora: their compilation and application*, pp. 31–40.

- Benzécri, J. P. (1992). Correspondence analysis handbook. New York: Marcel Dekker.
- Bestgen, Y. (2014). Inadequacy of the chi-squared test to examine vocabulary differences between corpora. *Literary and Linguistic Computing*, 29(2), 164–70.
- Biber, D. (1988). Variation across speech and writing. Cambridge University Press.
- Biber, D. & Conrad, S. (2009). Register, genre, and style. Cambridge University Press.
- Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. (1999). Longman grammar of spoken and written English. Harlow: Longman.
- Biber, D. & Jones, K. (2009). Quantitative methods in corpus linguistics. In A. Lüdeling & M. Kytö (eds.), *Corpus linguistics: an international handbook*, vol. 2, pp. 1287–1304. Berlin: Walter de Gruyter.
- Biber, D., Reppen, R., Schnur, E. & Ghanem, R. (2016). On the (non) utility of Juilland's D to measure lexical dispersion in large corpora. *International Journal of Corpus Linguistics*, 21(4), 439–64.
- Blythe, R. A. & Croft, W. (2012). S-curves and the mechanisms of propagation in language change. *Language*, 88(2), 269–304.
- Boneau, C. A. (1960). The effects of violations of assumptions underlying the t test. *Psychological Bulletin*, 57(1), 49.
- Borenstein, M. (2009). Effect sizes for continuous data. In H. Cooper, L. Hedges & J. Valentine (eds.), *The handbook of research synthesis and meta-analysis*, pp. 221–35. New York: Russell Sage Foundation.
- Brezina, V. (2013). BNC64 Search & Compare. Available at: <http://corpora.lancs.ac.uk/bnc64> (accessed 20/08/2016).
- Brezina, V. (2014). Effect sizes in corpus linguistics: keywords, collocations and diachronic comparison. Presented at the ICAME 2014 conference, University of Nottingham.
- Brezina, V. & Gablasova, D. (2015). Is there a core general vocabulary? Introducing the New General Service List. *Applied Linguistics*, 36(1), 1–22.
- Brezina, V., McEnery, T. & Baker, H. (in prep.) Usage fluctuation analysis: a new way of analysing shifts in historical discourse.
- Brezina, V., McEnery, T. & Wattam, S. (2015). Collocations in context. *International Journal of Corpus Linguistics*, 20(2), 139–73.
- Brezina, V. & Meyerhoff, M. (2014). Significant or random? A critical review of sociolinguistic generalisations based on large corpora. *International Journal of Corpus Linguistics*, 19(1), 1–28.
- Brezina, V. & Timperley, M. (2017). How large is the BNC? A proposal for standardized tokenization and word counting. CL2017, Birmingham. Available at: www.birmingham.ac.uk/Documents/college-artslaw/corpus/conference-archives/2017/general/paper303.pdf (accessed 08/03/18).
- Brezina, V., Timperley, M., Gablasova, D. & McEnery, T. (in prep.). #LancsBox: a new generation corpus tool for researchers, students and teachers.
- Cabin, R. J. & Mitchell, R. J. (2000). To Bonferroni or not to Bonferroni: when and how are the questions. *Bulletin of the Ecological Society of America*, 81(3), 246–8.
- Chernick, M. R. & LaBudde, R. A. (2014). An introduction to bootstrap methods with applications to R. Hoboken, NJ: John Wiley & Sons.
- Chomsky, N. (2000). New horizons in the study of language and mind. Cambridge University Press.
- Clausen, S. E. (1998). Applied correspondence analysis: an introduction. Thousand Oaks, CA: Sage.
- Cleveland, W. S. (1994). The elements of graphing data. Summit, NJ: Hobart Press.
- Cohen, J. (1988). Statistical power analysis for the behavioral sciences. Hillsdale, NJ: Erlbaum.
- Cohen, M. P. (2000). Note on the odds ratio and the probability ratio. *Journal of Educational and Behavioral Statistics*, 25(2), 249–52.
- Colley, A. & Todd, Z. (2002). Gender-linked differences in the style and content of e-mails to friends. *Journal of Language and Social Psychology*, 21(4), 380–92.

- Conrad, S. & Biber, D. (2001). Multidimensional methodology and the dimensions of register variation in English. In S. Conrad & D. Biber (eds.), *Variation in English: multidimensional studies*, pp. 18–19. Harlow: Pearson Education.
- Coupland, N. (2007). *Style: language variation and identity*. Cambridge University Press.
- Covington, M. A. & McFall, J. D. (2010). Cutting the Gordian knot: the moving-average type–token ratio (MATTR). *Journal of Quantitative Linguistics*, 17(2), 94–100.
- Crystal, D. (2003). *English as a global language*. Cambridge University Press.
- Cumming, G. (2012). *Understanding the new statistics*. New York: Routledge.
- Cumming, G., Fidler, F. & Vaux, D. L. (2007). Error bars in experimental biology. *Journal of Cell Biology*, 177(1), 7–11.
- Davies, H. T. O., Crombie, I. K. & Tavakoli, M. (1998). When can odds ratios mislead? *British Medical Journal*, 316(7136), 989–91.
- Davies, M. & Gardner, D. (2010). *A frequency dictionary of contemporary American English: word sketches, collocates and thematic lists*. London: Routledge.
- de Winter, J. C. (2013). Using the Student's t-test with extremely small sample sizes. *Practical Assessment, Research & Evaluation*, 18(10), 1–12.
- Diggle, P. J. & Chetwynd, A. G. (2011). *Statistics and scientific method: an introduction for students and researchers*. Oxford University Press.
- Divjak, D. & Gries, S. Th. (2006). Ways of trying in Russian: clustering behavioral profiles. *Corpus Linguistics and Linguistic Theory*, 2(1), 23–60.
- Dodge, Y. (2008). *The concise encyclopedia of statistics*. New York: Springer.
- Edgell, S. E. & Noon, S. M. (1984). Effect of violation of normality on the t test of the correlation coefficient. *Psychological Bulletin*, 95(3), 576. Efron, B. (1979). Computers and the theory of statistics: thinking the unthinkable. *SIAM Review*, 21(4), 460–80.
- Efron, B. & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. Boca Raton, FL: CRC Press.
- Erceg-Hurn, D. M. & Miroseovich, V. M. (2008). Modern robust statistical methods: an easy way to maximize the accuracy and power of your research. *American Psychologist*, 63(7), 591.
- Everitt, B. S., Landau, S., Leese, M. & Stahl, D. (2011). *Cluster analysis*. New York: John Wiley & Sons.
- Evert, S. (2008). Corpora and collocations. In A. Lüdeling & M. Kytö (eds.), *Corpus linguistics: an international handbook*, vol. 1, pp. 223–33. Berlin: Walter de Gruyter.
- Field, A., Miles, J. & Field, Z. (2012). *Discovering statistics using R*. London: Sage.
- Firth, J. (1957). *Papers in linguistics*. Oxford University Press.
- Francis, W. N. & Kučera, H. (1979). *Brown Corpus manual: manual of information to accompany a standard corpus of present-day edited American English for use with digital computers*. Brown University, Providence, RI. Available at <http://clu.uni.no/icame/brown/bcm.html>
- Friendly, M. (2002). A brief history of the mosaic display. *Journal of Computational and Graphical Statistics*, 11(1), 89–107.
- Friginal, E. & Hardy, J. (2014). Conducting multi-dimensional analysis using SPSS. In T. B. Sardinha & M. V. Pinto (eds.), *Multi-dimensional analysis, 25 years on: a tribute to Douglas Biber*, pp. 297–316. Amsterdam: John Benjamins.
- Fritz, C. O., Morris, P. E. & Richler, J. J. (2012). Effect size estimates: current use, calculations, and interpretation. *Journal of Experimental Psychology: General*, 141(1), 2–18.
- Gablasova, D., Brezina, V. & McEnery, A. M. (2017a). Exploring learner language through corpora: comparing and interpreting corpus frequency information. *Language Learning*, 67(S1), 130–54.
- (2017b). Collocations in corpus-based language learning research: identifying, comparing

- and interpreting the evidence. *Language Learning*, 67(S1), 155–79.
- Gablasova, D., Brezina, V., McEnery, T. & Boyd, E. (2017). Epistemic stance in spoken L2 English: the effect of task and speaker style. *Applied Linguistics*, 38(5), 613–37.
- Gabrielatos, C. & Marchi, A. (2012) Keyness: appropriate metrics and practical issues. Presented at CADS International Conference 2012, Corpus-assisted Discourse Studies: More than the sum of Discourse Analysis and computing? University of Bologna, Italy.
- Glass, G. V. (1965) A ranking variable analogue of biserial correlation: implications for short-cut item analysis. *Journal of Educational Measurement*, 2(1), 91–5.
- Greenacre, M. (2007). Correspondence analysis in practice. Boca Raton: Chapman & Hall/CRC.
- Gries, S. Th. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*, 13(4), 403–37.
- (2010). Dispersions and adjusted frequencies in corpora: further explorations. In S. Th. Gries, S. Wulff & M. Davies, *Corpus linguistic applications: current studies*, pp. 197–212. Amsterdam: Rodopi.
- (2013a). *Statistics for linguistics with R: a practical introduction*. Berlin: Walter de Gruyter.
- (2013b). 50-something years of work on collocations: what is or should be next. *International Journal of Corpus Linguistics*, 18(1), 137–66.
- Gries, S. Th. & Hilpert, M. (2008). The identification of stages in diachronic data: variability-based neighbour clustering. *Corpora*, 3(1), 59–81.
- (2010). Modeling diachronic change in the third person singular: a multifactorial, verband author-specific exploratory approach. *English Language and Linguistics*, 14(03), 293–320.
- Gries, S. Th., Newman, J., Shaoul, C. & Dilts, P. (2009). N-grams and the clustering of genres. Presented at workshop on Corpus, Colligation, Register Variation at the 31st Annual Meeting of the Deutsche Gesellschaft für Sprachwissenschaft, March.
- Grieve-Smith, A. (2007). The envelope of variation in multidimensional register and genre analyses. In E. Fitzpatrick (ed.), *Corpus linguistics beyond the word: corpus research from phrase to discourse*, pp. 21–42. Amsterdam: Rodopi.
- Gwet, K. (2002). Inter-rater reliability: dependency on trait prevalence and marginal homogeneity. *Statistical Methods for Inter-Rater Reliability Assessment Series*, 2, 1–9.
- Hand, D. J. (2010). Evaluating diagnostic tests: the area under the ROC curve and the balance of errors. *Statistics in Medicine*, 29(14), 1502–10.
- Hardie, A. (2014) Log ratio – an informal introduction. <http://cass.lancs.ac.uk/?p=1133>
- Harrington, J., Palethorpe, S. & Watson, C. I. (2000). Does the Queen speak the Queen's English? *Nature*, 408(6815), 927–8.
- Hayton, J. C., Allen, D. G. & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: a tutorial on parallel analysis. *Organizational Research Methods*, 7(2), 191–205.
- Healey, A. diPaolo (ed.) (2004). *The Complete Corpus of Old English in Electronic Form*. Dictionary of Old English Project. Centre for Medieval Studies, University of Toronto.
- Hill, T., Lewicki, P. & Lewicki, P. (2006). *Statistics: methods and applications: a comprehensive reference for science, industry, and data mining*. Tulsa, OK: StatSoft.
- Hilpert, M. (2011). Dynamic visualizations of language change: motion charts on the basis of bivariate and multivariate data from diachronic corpora. *International Journal of Corpus Linguistics*, 16(4), 435–61.
- Hilpert, M. & Gries, S. Th. (2009). Assessing frequency changes in multistage diachronic corpora: applications for historical corpus linguistics and the study of language acquisition. *Literary and Linguistic Computing*, 24(4), 385–401.

- Hosmer, D.W., Lemeshow, S. & Sturdivant, R. X. (2013). *Applied logistic regression*, 3rd edn. Hoboken, NJ: John Wiley & Sons.
- Hudson, T. (2015). Presenting quantitative data visually. In L. Plonsky (ed.), *Advancing quantitative methods in second language research*, pp. 78–105. London: Routledge.
- Ito, R. & Tagliamonte, S. (2003). Well weird, right dodgy, very strange, really cool: layering and recycling in English intensifiers. *Language in Society*, 32(02), 257–79.
- Jakubíček, M., Kilgarrieff, A., Kovář, V., Rychlý, P. & Suchomel, V. (2013). The TenTen corpus family. *Proceedings of the International Conference on Corpus Linguistics 2013*, pp. 125–7. Lancaster University.
- Jarvis, S. (2013). Capturing the diversity in lexical diversity. *Language Learning*, 63(s1), 87–106.
- Johnson, D. E. (2009). Getting off the GoldVarb standard: introducing Rbrul for mixed effects variable rule analysis. *Language and Linguistics Compass*, 3(1), 359–83.
- Juilland, A. G., Brodin, D. R. & Davidovitch, C. (1970). *Frequency dictionary of French words*. The Hague: Mouton.
- Juilland, A. G. & Chang-Rodríguez, E. (1964). *Frequency dictionary of Spanish words*. The Hague: Mouton.
- Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, 20(1), 141–51.
- Kepes, S., Banks, G. C. & Oh, I. S. (2014). Avoiding bias in publication bias research: the value of 'null' findings. *Journal of Business and Psychology*, 29(2), 183–203.
- Kerby, D. S. (2014). The simple difference formula: an approach to teaching nonparametric correlation. *Innovative Teaching*, 3, 1–9.
- Kilgarrieff, A. (1997). Putting frequencies in the dictionary. *International Journal of Lexicography*, 10(2), 135–55.
- (2005). Language is never, ever, ever, random. *Corpus Linguistics and Linguistic Theory*, 1(2), 263–76.
- (2009). Simple maths for keywords. In *Proceedings of the Corpus Linguistics Conference*, Liverpool, July.
- (2012). Getting to know your corpus. In *Proceedings of the 15th International Conference on Text, Speech and Dialogue*, pp. 3–15. Berlin: Springer.
- Kirk, R. E. (1996). Practical significance: a concept whose time has come. *Educational and Psychological Measurement*, 56(5), 746–59.
- (2005). Effect size measures. *Wiley StatsRef: Statistics Reference Online*. <http://dx.doi.org/10.1002/9781118445112.stat06242.pub2>
- Krippendorff, K. (2012 [1980]). *Content analysis: an introduction to its methodology*. London: Sage.
- Kruskal, W. H. & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260), 583–621.
- Kučera, H. & Francis, W. N. (1967). *Computational analysis of Present-Day American English*. Providence, RI: Brown University Press.
- Labov, W. (1966). *The social stratification of English in New York City*. Washington, DC: Center for Applied Linguistics.
- (1972). *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.
- (2010). *Principles of linguistic change*, vol. 3: Cognitive and cultural factors. Oxford: Wiley-Blackwell.
- Lakoff, G. & Johnson, M. (1980). *Metaphors we live by*. University of Chicago Press.
- Lakoff, R. T. (1975). *Language and woman's place*. New York: Harper & Row.
- Lavandera, B. R. (1978). Where does the sociolinguistic variable stop? *Language in Society*, 7(02), 171–82.
- Ledesma, R. D. & Valero-Mora, P. (2007). Determining the number of factors to retain in EFA: an easy-to-use computer program for carrying out parallel analysis. *Practical Assessment, Research & Evaluation*, 12(2), 1–11.

- Leech, G. (1992). Corpora and theories of linguistic performance. In J. Svartvik (ed.), *Directions in corpus linguistics*, pp. 105–22. Berlin: Mouton de Gruyter.
- (2003). Modals on the move: the English modal auxiliaries 1961–1992. In R. Facchinetti, F. R. Palmer & M. Krug (eds.), *Modality in contemporary English*, 223–40. Berlin: Mouton de Gruyter.
- (2011). The modals ARE declining. *International Journal of Corpus Linguistics*, 16(4), 547–64.
- Leech, G., Garside, R. & Bryant, M. (1994). CLAWS4: the tagging of the British National Corpus. In *Proceedings of the 15th Conference on Computational Linguistics*, Kyoto, vol. 1, pp. 622–8.
- Leech, G., Rayson, P. & Wilson, A. (2001). *Word frequencies in written and spoken English: based on the British National Corpus*. London: Routledge.
- Leek, J. T. & Peng, R. D. (2015). Statistics: p values are just the tip of the iceberg. *Nature*, 520(7549), 612.
- Lijffijt, J., Nevalainen, T., Säily, T., Papapetrou, P., Puolamäki, K. & Mannila, H. (2016). Significance testing of word frequencies in corpora. *Literary and Linguistic Computing*, 31(2), 374–97.
- Lijffijt, J., Säily, T. & Nevalainen, T. (2012). CEECing the baseline: lexical stability and significant change in a historical corpus. In *Studies in Variation, Contacts and Change in English*, vol. 10. Helsinki: Research Unit for Variation, Contacts and Change in English (VARIENG).
- Love, R., Dembry, C., Hardie, A., Brezina, V. & McEnery, T. (2017). The Spoken BNC2014: designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics*, 22 (3).
- Lumley, T., Diehr, P., Emerson, S. & Chen, L. (2002). The importance of the normality assumption in large public health data sets. *Annual Review of Public Health*, 23(1), 151–69.
- Malvern, D. & Richards, B. (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing*, 19, 85–104.
- Mann, H. B. & Whitney, D. R. (1947) On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18(1), 50–60.
- Manning, C. D. (2011). Part-of-speech tagging from 97% to 100%: is it time for some linguistics? In A. F. Gelbukh (ed.), *International Conference on Intelligent Text Processing and Computational Linguistics*, pp. 171–89. Berlin: Springer.
- McEnery, T. (2006). *Swearing in English: bad language, purity and power from 1586 to the present*. Abingdon: Routledge.
- McEnery, T. & Baker, H. (2017). *Corpus linguistics and 17th-century prostitution: computational linguistics and history*. London: Bloomsbury.
- McEnery, T. & Hardie, A. (2011). *Corpus linguistics: method, theory and practice*. Cambridge University Press.
- Mehl, M. R., Vazire, S., Ramírez-Esparza, N., Slatcher, R. B. & Pennebaker, J.W. (2007). Are women really more talkative than men? *Science*, 317(5834), 82.
- Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P. . . & Pinker, S. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014), 176–82.
- Microsoft (2010). Microsoft Word [software].
- Millar, N. (2009). Modal verbs in TIME: frequency changes 1923–2006. *International Journal of Corpus Linguistics*, 14 (2), 191–220.
- Nevalainen, T. (1999). Making the best use of ‘bad’ data: evidence for sociolinguistic variation in Early Modern English. *Neuphilologische Mitteilungen*, 499–533.
- Nevalainen, T. & Raumolin-Brunberg, H. (2003). *Historical sociolinguistics: language change in Tudor and Stuart England*. London: Routledge.

- Newman, M. L., Groom, C. J., Handelman, L. D. & Pennebaker, J. W. (2008). Gender differences in language use: an analysis of 14,000 text samples. *Discourse Processes*, 45(3), 211–36.
- Nini, A. (2015) Multidimensional Analysis Tagger (v. 1.3) – manual. Available at: <https://sites.google.com/site/multidimensionaltagger/> (accessed 26/08/15).
- Osborne, J.W. (2012). *Best practices in data cleaning: a complete guide to everything you need to do before and after collecting your data*. Thousand Oaks, CA: Sage.
- Osborne, J.W. (2015). *Best practices in logistic regression*. Thousand Oaks, CA: Sage.
- Pearson, K. (1920). Notes on the history of correlation. *Biometrika*, 13(1), 25–45.
- Pechenick, E. A., Danforth, C. M. & Dodds, P. S. (2015). Characterizing the Google Books corpus: strong limits to inferences of socio-cultural and linguistic evolution. *PloS One*, 10(10), e0137041.
- Phillips, M. (1985). *Aspects of text structure: an investigation of the lexical organisation of text*. Amsterdam: North-Holland.
- Popper, K. (2005 [1935]). *The logic of scientific discovery*. London: Routledge.
- Rayson, P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics* 13(4), 519–49.
- Rayson, P., Berridge, D. & Francis, B. (2004). Extending the Cochran rule for the comparison of word frequencies between corpora. *Proceedings from 7th International Conference on Statistical Analysis of Textual Data (JADT 2004)*, pp. 926–36.
- Rayson, P., Leech, G. N. & Hodges, M. (1997). Social differentiation in the use of English vocabulary: some analyses of the conversational component of the British National Corpus. *International Journal of Corpus Linguistics*, 2(1), 133–52.
- Richardson, J. T. (2011). Eta squared and partial eta squared as measures of effect size in educational research. *Educational Research Review*, 6(2), 135–47.
- Rosenthal, R. (1995). Writing meta-analytic reviews. *Psychological Bulletin*, 118(2), 183.
- Rothstein, H. & Hopewell, S. (2009). Grey literature. In H. Cooper, L. Hedges & J. Valentine (eds.), *The handbook of research synthesis and meta-analysis*, pp. 103–26. New York: Russell Sage Foundation.
- Savický, P. & Hlaváčová, J. (2002). Measures of word commonness. *Journal of Quantitative Linguistics*, 9(3), 215–31.
- Schmider, E., Ziegler, M., Danay, E., Beyer, L. & M. Bühner. (2010). 'Is it really robust?' Reinvestigating the robustness of ANOVA against violations of the normal distribution assumption. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 6(4), 147–51.
- Schmidt, F. L. & Hunter, J. E. (2015). *Methods of meta-analysis: correcting error and bias in research findings*. Thousand Oaks, CA: Sage Publications.
- Scott, M. (1997). PC analysis of key words – and key key words. *System*, 25(2), 233–45.
- Scott, M. (2004). *WordSmith tools version 4*. Oxford University Press.
- Shadish, W. R. & Haddock, C. K. (2009). Combining estimates of effect size. In H. Cooper, L. Hedges & J. Valentine (eds.), *The handbook of research synthesis and meta-analysis*, pp. 257–78. New York: Russell Sage Foundation.
- Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology*, 46, 561–84.
- Shakespeare, W. (1992). *The Poems: Venus and Adonis, The Rape of Lucrece, The Phoenix and the Turtle, The Passionate Pilgrim*. Cambridge University Press.
- Sheskin, D. J. (2007). *Handbook of parametric and nonparametric statistical procedures*. Boca Raton, FL: Chapman & Hall/CRC.
- Siegel, S. (1956). *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.
- Sprent, P. (2011). Fisher Exact Test. In *International encyclopedia of statistical science*, pp. 524–5. Berlin: Springer.
- Stubbs, M. (2001). *Words and phrases: corpus studies of lexical semantics*. Oxford: Blackwell.
- Tagliamonte, S. A. (2006). *Analysing sociolinguistic variation*. Cambridge University Press.

- Theus, M. & Urbanek, S. (2008). *Interactive graphics for data analysis: principles and examples*. Boca Raton, FL: CRC Press.
- Toothaker, L. E. (1993). Multiple comparison procedures. *Sage University Paper Series on Quantitative Applications in the Social Sciences*, 07–089. Newbury Park, CA: Sage.
- Trafimow, D. & Marks, M. (2015) Editorial. *Basic and Applied Social Psychology*, 37(1), 1–2.
- Tufte, E. (1997). *Visual explanations*. Cheshire, CT: Graphics Press.
- _____. (2001). *Visual display of quantitative information*. Cheshire, CT: Graphics Press.
- _____. (2006). *Beautiful evidence*. Cheshire, CT: Graphics Press.
- Tweedie, F. & Baayen, R. H. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities*, 32, 323–52.
- Upton, G. J. (1992). Fisher's exact test. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 395–402.
- Valentine, J. (2009). Judging the quality of primary research. In H. Cooper, L. Hedges & J. Valentine (eds.), *The handbook of research synthesis and meta-analysis*, pp. 129–46. New York: Russell Sage Foundation.
- Verma, J. P. (2016). *Repeated measures design for empirical researchers*. Hoboken, NJ: John Wiley & Sons.
- Vine, B. (1999). Guide to the New Zealand component of the International Corpus of English (ICE-NZ). School of Linguistics and Applied Language Studies, Victoria University of Wellington.
- Vine, E. W. (2011). High frequency multifunctional words: accuracy of word-class tagging. *Te Reo*, 54, 71.
- Williams, G. (1998). Collocational networks: interlocking patterns of lexis in a corpus of plant biology research articles. *International Journal of Corpus Linguistics*, 3(1), 151–71.
- Wilson, D. B. (2009). Systematic coding. In H. Cooper, L. Hedges & J. Valentine (eds.), *The handbook of research synthesis and meta-analysis*, pp. 159–76. New York: Russell Sage Foundation.
- Xiao, R. (2009). Multidimensional analysis and the study of world Englishes. *World Englishes*, 28(4), 421–50.

ثبت المصطلحات

أولاً: (عربي-إنجليزي)

أ

Robust	(إحصاء) متين
Direction of the effect	اتجاه التأثير
Directionality	الاتجاهية
Log odds (logits)	أرجحيات اللوج
Probability of superiority (PS)	احتمالية الأفضلية
Statistical test	اختبار إحصائي
Log-likelihood test	اختبار الاحتمالية اللوغارتمية
Post-hoc tests	اختبارات بعدية
Null-hypothesis significance testing (NHST)	اختبار دلالة الفرضية الصفرية
Wald's z	اختبار زاي والد
Chi-squared test	اختبار كاي تربيع
Kruskal-Wallis test	اختبار كروسكال-والاس
Mann-Whitney U test	اختبار مان-ويتني U
Wilcoxon rank-sum test	اختبار مجموع رتب ويلكسون
Likelihood ratio test (G test)	اختبار نسبة الاحتمالية
Repeated measures tests	اختبارات القياسات المتكررة

Parametric tests	اختبارات بارامترية
t-test	اختبار-ت
Contractions	الاختصارات
Standard errors	أخطاء المعيار
Correlation	الارتباط
(multi)collinearity between predictors	ارتباط الخطية (المتعددة) بين المتنبئات
Rank biserial correlation	ارتباط المتسلسل الثنائي الرتبي
Positive correlation	ارتباط إيجابي
Pearson's correlation	ارتباط بيرسون
Spearman's correlation	ارتباط سيرمان
Negative correlation	ارتباط سلبي
Odds	الأرجحية
Individual styles	الأساليب الشخصية
Placement of registers	استبدال السجلات اللغوية
Extraction of factors	استخلاص العوامل
Independence of observations	استقلال الملاحظات
Diachronic polysemy	الاشتراك اللفظي التعاقبي
Sampling frame	إطار العينة
Lexico-grammatical frame	الإطار النحوي المعجمي
Error bars	أعمدة الخطأ
Normality assumption	الافتراض الاعتدالي
Assumptions	الافتراضات
Multinomial logistic regression	الانحدار الخطي المنطقي المتعدد الحدود
Simple logistic regression (VARBRUL)	الانحدار المنطقي البسيط

Logistic regression	انحدار منطقي
Binomial logistic regression	انحدار منطقي ثنائي الحدين
Standard deviation	الانحراف المعياري
Deviation of Proportions DP	انحراف النسب
Curvilinear	الانحنائية الخطية
Self-selection bias	انحياز الاختيار الذاتي
Practicality bias	الانحياز التطبيقي العملي
Legal considerations bias	انحياز لاعتبارات مقبولة
Practical importance	الأهمية التطبيقية

ب

Simple maths parameter	البارامتر الرياضي البسيط

ن

Fixed effects	التأثيرات الثابتة
Main effects	التأثيرات الرئيسية
Random effects	التأثيرات العشوائية
MU	التبادل
Variance (S2 or SD2)	التباين
Fluctuation of the meaning of linguistic forms	تبدل معاني الصيغ اللغوية
Data sparsity	تبعثر البيانات
Clustering	تجميع
Variability-based neighbor clustering (VNC)	تجميع الجوار القائم على المتغيرة
Meta-analysis	التحليل البعدي
One-way ANOVA	تحليل التباين الأحادي
Multi-way ANOVA (factorial ANOVA)	تحليل التباين المتعدد (التباين العاملي)

Homogeneity analysis	تحليل التجانس
Usage Fluctuation Analysis (UFA)	تحليل التقلب الاستعمالي
Correspondence analysis	تحليل التناظر
Factor analysis	تحليل العامل
Hierarchical agglomerative cluster analysis	تحليل العنقود المتكامل الهرمي
Multidimensional analysis	تحليل متعدد الأبعاد
Bias	تحيز
Publication bias	تحيز المجموع اللغوي
Topic bias	التحيز الموضوعي
Non-coverage bias	تحيز بدون تغطية
Text sample bias	تحيز عينة النص
Traditional text type bias	تحيز نوع النص التقليدي
Dispersion	التشتت
Collocations	التصاحبات اللفظية
Research design	تصميم البحث
corpus design	تصميم المدونة
Linguistic feature design	تصميم خاصية لغوية
Linguistic feature design	تصميم خصائص لغوية
Average profile	تعريف المتوسط
Covariance	التغاير
Change over time	التغير مع مرور الوقت
Interactions between predictors	التفاعلات بين المتنبئات
Homoscedasticity	التفاوت المتماثل
Functional interpretation of factors	التفسير الوظيفي للعوامل

Expected frequency of collocation	تكرار التصاحب المتوقع
Observed frequency of collocation	تكرار التصاحب الملحوظ
Average reduced frequency (ARF)	تكرار المتوسط المقلص
Absolute (or raw) frequency (AF)	التكرار المطلق
Relative (or normalized) frequency (RF)	التكرار النسبي (أو الممعيّر/المعتدل)
Mean frequency	تكرار الوسيط
Diachronic representativeness	التمثيلية التعاقبية
Bootstrapping	التمهيد
Replication	التنتاج
Frequency distribution	توزيع التكرار
Juilland's D	توزيع جولاند
Normal distribution	توزيع طبيعي

ث

Stability over time	الثبات مع مرور الوقت

ج

Cross-tabulation tables	جداول النتائج المتقاطعة
Cross-tabulation	الجدولة المتقاطعة

ح

Effect size	حجم التأثير
Corpus size	حجم المدونة
Minimal sensitivity MS	الحساسية ذات النهاية الصغرى
Intercept (or constant)	الحصر (الكمية الثابتة)

خ

Random co-occurrence base-line	خط أساس توارد التصاحب العشوائي
Regression line	خط الانحدار
Line of the best fit	خط التوافق الأفضل
Line graph	خط بياني
Stacked bar chart	خطاطة أعمدة متكدسة
Bar chart	خطاطة الأعمدة
Linearity	الخطية

د

Statistically significant	دال إحصائيا
Dice	الدايس
Diachronic study	الدراسة التعاقبية
Longitudinal study	الدراسة الطولية
The degrees of freedom (df)	درجة الحرية
Delta P	دلتا ب
C-index (concordance index)	دليل الكشف (السياقي)

ر

CLINK (complete linkage)	الربط التام
Average linkage	الربط المتوسط
SLINK (single linkage)	الربط المنفرد
Circumscribe the variable context	رسم حدود المتغير

س

Registers	السجلات اللغوية

Variable context

سياق المتغير

ش

Collocation network

شبكة التصاحب

Collocation graph

الشكل البياني للتصاحب

ظ

Envelope of variation

ظرف التنوع

ع

Node

العقدة

Inferential statistics

علم الإحصاء الاستدلالي

Descriptive statistics

علم الإحصاء الوصفي

External factors

عوامل خارجية

Internal factors

عوامل داخلية

Welch's independent samples

عينات ويلتش المستقلة

Sample

عينة

Random sampling

عينة عشوائية

ف

Null hypothesis

الفرضية الصفرية

(quasi) complete separation

فصل (شبه) كامل

ق

Zipf's law

قانون زيف

Peaks and troughs (Generalized Additive Model GAM)

القمم والقيعان

Statistical measure

قياس إحصائي

Optimal scaling

القياس الأمثل

Measure of central tendency	قياس النزعة المركزية
Association measures (collocation measures)	قياسات التجاور (قياسات التصاحب)
Values	قيم
Baseline values	قيم المنطلق
p-value	قيمة الاحتمال
Eigenvalue	القيمة الكاملة (الجذر الكامن للتحليل)
T-score	قيمة-ت
Z-score	قيمة-ز

ك

κ	كبا
Lockwords	كلمات الإيقاف
Token (running word)	الكلمة الفعلية
Type	الكلمة النوعية
Hapax legomena (hapaxes)	كلمة فريدة
Cohen's d	كوهن-د
Cohen's κ	كوهن كبا

ل

Lexeme	اللكسيم
Sociolects	اللهجات الاجتماعية
Log Dice	اللوج دايس
Log-likelihood	اللوغارتمية الاحتمالية

م

Lemma	المادة المعجمية
Principle of accountability	مبدأ الموثوقية

Collocates	المتصاحبات
Outlier or rogue value	المتطرفة أو القيمة المتطرفة
Variable	متغير
Labovian sociolinguistic variable	متغير اللابوفيين اللغوي الاجتماعي
Ordinal variable	متغير ترتيبى
Judgement variable	متغير تقييمى
Outcome variable	متغير ناتج
Nominal variable	متغيرات اسمية
Predictor variables	متغيرات المتنبئ
Dummy variables	متغيرات صورية
Categorical variables	متغيرات فئوية
Cramer's V	متغيرات كرامر (معامل كرامر)
Linguistic variables	متغيرات لغوية
Ambient linguistic variables	متغيرات لغوية محيطية
Multiple variables	متغيرات متعددة
Explanatory variables	متغيرات مستقلة
Explanatory variables or predictors	متغيرات مستقلة أو متنبئات
Mean	المتوسط
Average	المتوسط
The grand mean	المتوسط الكلى
Trimmed mean	المتوسط المرتب
Confidence interval (CI)	مجال الثقة
Column/row totals	مجاميع الأعمدة والصفوف
Grand total	المجموع الكلى

Population	المجموع اللغوي
Dataset	مجموعة بيانات
Scatterplot	مخطط التشتت
Correspondence plot	مخطط التناظر
Candlestick plot	مخطط الشموع
Histogram	مخطط درجي
Boxplot	مخطط رسم المربع
Scree plot	مخطط ركامي
Mosaic plot	مخطط فسيفسائي
Stepwise entry	المدخل التدريجي
Block entry	المدخل الثابت
Diachronic corpora	المدونات التعاقبية
Population-based corpora	مدونات قائمة على المجموع اللغوي
Corpus	المدونة (الحاسوبية)
Corpus of interest (C) or focus corpus	المدونة الرئيسية
Reference corpus	المدونة المرجعية
Range	المدى
Span	المدى
Interquartile range	المدى الربعي
Range2 (R)	المدى ٢
Euclidean distance	مسافة إقليدية
Chi-squared distance	مسافة كاي تربيع
Manhattan distance	مسافة منهاتن
Reference levels	مستويات المرجع

Scatterplot matrix	مصفوفة مخطط التشتت
R2 (coefficient of determination)	معامل التحديد
Coefficient of variation (CV)	معامل التنوع
Relative ratio	المعدل النسبي
Mutual Information (MI)	المعلومات المتبادلة
Akaike information criterion (AIC)	معيار معلومات أكايكي
Linguistic and social meaningfulness	المغزى اللغوي والاجتماعي
Scale	مقياس
Inter-rater agreement	ملاءمة المقدّر البيني
Profiles	ملفات التعريف
Representative	ممثل
Baseline	المنطلق (خط الأساس)
Raw agreement	موائمة الصفوف

ن

Outcome of interest	ناتج الاكتراث
Collocation window	نافذة التصاحب اللفظي
Null results	نتائج صفريّة
Probability ratio (PR)	نسبة الاحتمال
Odds ratio	نسبة الأرجحية
Risk ratio	نسبة الخطر
Type-token ratio TTR	نسبة الكلمات النوعية إلى الكلمات الفعلية
Log ratio	نسبة اللوج
Standardized type/token ratio (STTR)	النسبة الموحدة للكلمات النوعية/ الفعلية
Moving average type/token ratio (MATTR)	نسبة متوسط الكلمة النوعية/ الفعلية المتحرك

Non-parametric counterparts	النظائر غير البارامترية
Data point or case	نقطة أو حالة البيانات
Mixed-effect modelling	نمذجة التأثير الممتزج
Parsimonious model	النموذج المقتر
Baseline (or null) model	نموذج المنطلق (أو الصفر)

9

Median (mdn)	الوسيط
Web as corpus	الويب كمدونة

ثانياً: (إنجليزي-عربي)

A

Absolute (or raw) frequency (AF)	التكرار المطلق
Akaike information criterion (AIC)	معييار معلومات أكايكي
Ambient linguistic variables	متغيرات لغوية محيطة
Association measures (collocation measures)	قياسات التجاور (قياسات التصاحب)
Assumptions	الافتراضات
Average	المتوسط
Average profile	تعريف المتوسط
Average reduced frequency (ARF)	تكرار المتوسط المقلص

B

Bar chart	خطاطة الأعمدة
Baseline	المنطلق (خط الأساس)
Baseline (or null) model	نموذج المنطلق (أو الصفر)
Baseline values	قيم المنطلق
Bias	تحيز
Binomial logistic regression	انحدار منطقي ثنائي الحدين
Block entry	المدخل الثابت
Bootstrapping	التمهيد
Boxplot	مخطط رسم المربع

C

Candlestick plot	مخطط الشموع
Categorical variables	متغيرات فئوية
Change over time	التغير مع مرور الوقت
Chi-squared distance	مسافة كاي تربيع

Chi-squared test	اختبار كاي تربيع
C-index (concordance index)	دليل الكشف (السياقي)
Circumscribe the variable context	رسم حدود المتغير
Clustering	تجميع
Coefficient of variation (CV)	معامل التنوع
Cohen's d	كوهن-د
Cohen's κ	كوهن كبا
Collocates	المتصاحبات
Collocation graph	الشكل البياني للتصاحب
Collocation network	شبكة التصاحب
Collocation window	نافذة التصاحب اللفظي
Collocations	التصاحبات اللفظية
Column/row totals	مجاميع الأعمدة والصفوف
Confidence interval (CI)	مجال الثقة
Contractions	الاختصارات
Corpus	المدونة (الحاسوبية)
corpus design	تصميم المدونة
Corpus of interest or focus corpus	المدونة الرئيسية
Corpus size	حجم المدونة
Correlation	الارتباط
Correspondence analysis	تحليل التناظر
Correspondence plot	مخطط التناظر
Covariance	التغاير
Cramer's V	متغيرات كرامر (معامل كرامر)

Cross-tabulation	الجدولة المتقاطعة
Cross-tabulation tables	جداول النتائج المتقاطعة
Curvilinear	الانحنائية الخطية

D

Data point or case	نقطة أو حالة البيانات
Data sparsity	تبعر البيانات
Dataset	مجموعة بيانات
Delta P	دلتا ب
Descriptive statistics	علم الإحصاء الوصفي
Deviation of Proportions DP	انحراف النسب
Diachronic corpora	المدونات التعااقبية
Diachronic polysemy	الاشتراك اللفظي التعااقبي
Diachronic representativeness	التمثيلية التعااقبية
Diachronic study	الدراسة التعااقبية
Dice	الدائس
Direction of the effect	اتجاه التأثير
Directionality	الاتجاهية
Dispersion	التشتت
Dummy variables	متغيرات صورية

E

Effect size	حجم التأثير
Eigenvalue	القيمة الكاملة (الجذر الكامن للتحليل)
Envelope of variation	ظرف التنوع
Error bars	أعمدة الخطأ
Euclidean distance	مسافة إقليدية

Expected frequency of collocation	تكرار التصاحب المتوقع
Explanatory variables	متغيرات مستقلة
Explanatory variables or predictors	متغيرات مستقلة أو متنبئات
External factors	عوامل خارجية
Extraction of factors	استخلاص العوامل

F

Factor analysis	تحليل العامل
Fixed effects	التأثيرات الثابتة
Fluctuation of the meaning of linguistic forms	تبدل معاني الصيغ اللغوية
Frequency distribution	توزيع التكرار
Functional interpretation of factors	التفسير الوظيفي للعوامل

G

Grand total	المجموع الكلي

H

Hapax legomena (hapaxes)	كلمة فريدة
Hierarchical agglomerative cluster analysis	تحليل العنقود المتكامل الهرمي
Histogram	مخطط درجي
Homogeneity analysis	تحليل التجانس
Homoscedasticity	التفاوت المتماثل

I

Independence of observations	استقلال الملاحظات
Individual styles	الأساليب الشخصية
Inferential statistics	علم الإحصاء الاستدلالي
Interactions between predictors	التفاعلات بين المتنبئات
Intercept (or constant)	الحصر (الكمية الثابتة)
Internal factors	عوامل داخلية

Interquartile range المدى الربعي

Inter-rater agreement ملاءمة المُقدّر البيني

J

Judgement variable متغير تقييمي

Juilland's D توزيع جولاند

K

Kruskal-Wallis test اختبار كروسكال-والاس

κ كَبَا

L

Labovian sociolinguistic variable متغير اللابوفيين اللغوي الاجتماعي

Legal considerations bias انحياز لاعتبارات مقبولة

Lemma المادة المعجمية

Lexeme اللكسيم

Lexico-grammatical frame الإطار النحوي المعجمي

Likelihood ratio test (G test) اختبار نسبة الاحتمالية

Line graph خط بياني

Line of the best fit خط التوافق الأفضل

Linearity الخطية

Linguistic and social meaningfulness المغزى اللغوي والاجتماعي

Linguistic feature design تصميم خاصية لغوية

Linguistic feature design تصميم خصائص لغوية

Linguistic variables متغيرات لغوية

Lockwords كلمات الإيقاف

Log Dice اللوج دايس

Log odds (logits) أرجحيات اللوج

Log ratio	نسبة اللوج
Logistic regression	انحدار منطقي
Log-likelihood	اللوغارتمية الاحتمالية
Log-likelihood test	اختبار الاحتمالية اللوغارتمية
Longitudinal study	الدراسة الطولية

M

Main effects	التأثيرات الرئيسية
Manhattan distance	مسافة منهاتن
Mann-Whitney U test	اختبار مان-ويتني U
Mean	المتوسط
Mean frequency	تكرار الوسيط
Measure of central tendency	قياس النزعة المركزية
Median (mdn)	الوسيط
Meta-analysis	التحليل البعدي
Minimal sensitivity MS	الحساسية ذات النهاية الصغرى
Mixed-effect modelling	نمذجة التأثير الممتزج
Mosaic plot	مخطط فسيفسائي
Moving average type/token ratio (MATTR)	نسبة متوسط الكلمة النوعية / الفعلية المتحرك
MU	التبادل
(multi) collinearity between predictors	ارتباط الخطية (المتعددة) بين المتنبئات
Multidimensional analysis	تحليل متعدد الأبعاد
Multinomial logistic regression	الانحدار الخطي المنطقي المتعدد الحدود
Multiple variables	متغيرات متعددة
Multi-way ANOVA (factorial ANOVA)	تحليل التباين المتعدد (التباين العاملي)
Mutual Information (MI)	المعلومات المتبادلة

N

Negative correlation	ارتباط سلبي
Node	العقدة
Nominal variable	متغيرات اسمية
Non-coverage bias	تحيز بدون تغطية
Non-parametric counterparts	النظائر غير البارامترية
Normal distribution	توزيع طبيعي
Normality assumption	الافتراض الاعتدالي
Null hypothesis	الفرضية الصفرية
Null results	نتائج صفرية
Null-hypothesis significance testing (NHST)	اختبار دلالة الفرضية الصفرية

O

Observed frequency of collocation	تكرار التصاحب الملحوظ
Odds	الأرجحية
Odds ratio	نسبة الأرجحية
One-way ANOVA	تحليل التباين الأحادي
Optimal scaling	القياس الأمثل
Ordinal variable	متغير ترتيبى
Outcome of interest	نتائج الاكتراث
Outcome variable	متغير ناتج
Outlier or rogue value	المتطرفة أو القيمة المتطرفة

P

Parametric tests	اختبارات بارامترية
Parsimonious model	النموذج المقتر
Peaks and troughs (Generalized Additive Model GAM)	القمم والقمور

Pearson's correlation	ارتباط بيرسون
Placement of registers	استبدال السجلات اللغوية
Population	المجموع اللغوي
Population-based corpora	مدونات قائمة على المجموع اللغوي
Positive correlation	ارتباط إيجابي
Post-hoc tests	اختبارات بعدية
Practical importance	الأهمية التطبيقية
Practicality bias	الانحياز التطبيقي العملي
Predictor variables	متغيرات المتنبئ
Principle of accountability	مبدأ الموثوقية
Probability of superiority (PS)	احتمالية الأفضلية
Probability ratio (PR)	نسبة الاحتمال
Profiles	ملفات التعريف
Publication bias	تحيز المجموع اللغوي
p-value	قيمة الاحتمال
(quasi) complete separation	فصل (شبه) كامل

R

R ² (coefficient of determination)	معامل التحديد
Random co-occurrence base-line	خط أساس توارد التصاحب العشوائي
Random effects	التأثيرات العشوائية
Random sampling	عينة عشوائية
Range	المدى
Range ² (R)	المدى ^٢
Rank biserial correlation	ارتباط المتسلسل الثنائي الرتبي
Raw agreement	موائمة الصفوف

Reference corpus	المدونة المرجعية
Reference levels	مستويات المرجع
Registers	السجلات اللغوية
Regression line	خط الانحدار
Relative (or normalized) frequency (RF)	التكرار النسبي (أو المعيّر/ المعتدل)
Relative ratio	المعدل النسبي
Repeated measures tests	اختبارات القياسات المتكررة
Replication	التنتاج
Representative	ممثّل
Research design	تصميم البحث
Risk ratio	نسبة الخطر
Robust	إحصاء متين
S	
Sample	عينة
Sampling frame	إطار العينة
Scale	مقياس
Scatterplot	مخطط التشتت
Scatterplot matrix	مصفوفة مخطط التشتت
Scree plot	مخطط ركامي
Self-selection bias	انحياز الاختيار الذاتي
Simple logistic regression (VARBRUL)	الانحدار المنطقي البسيط
Simple maths parameter	البارامتر الرياضي البسيط
SLINK (single linkage)	الربط المنفرد
Sociolects	اللهجات الاجتماعية
Span	المدى

Spearman's correlation	ارتباط سيرمان
Stability over time	الثبات مع مرور الوقت
Stacked bar chart	خطاطة أعمدة متكدسة
Standard deviation	الانحراف المعياري
Standard errors	أخطار المعيار
Standardized type/token ratio (STTR)	النسبة الموحدة للكلمات النوعية/ الفعلية
Statistical measure	قياس إحصائي
Statistical test	اختبار إحصائي
Statistically significant	دال إحصائيا
Stepwise entry	المدخل التدريجي

T

Text sample bias	تحيز عينة النص
The absence of evidence is not the evidence of the absence	انعدام الدليل ليس دليلا على الانعدام
The degrees of freedom (df)	درجة الحرية
The grand mean	المتوسط الكلي
Token (running word)	الكلمة الفعلية
Topic bias	التحيز الموضوعي
Traditional text type bias	تحيز نوع النص التقليدي
Trimmed mean	المتوسط المرتب
T-score	قيمة-ت
t-test	اختبار-ت
Type	الكلمة النوعية
Type-token ratio TTR	نسبة الكلمات النوعية إلى الكلمات الفعلية

U

Usage Fluctuation Analysis (UFA)	تحليل التقلب الاستعمالي

V

Values

قيم

Variability-based neighbor clustering (VNC)

تجميع الجوار القائم على المتغيرة

Variable

متغير

Variable context

سياق المتغير

Variance (S2 or SD2)

التباين

W

Wald's z

اختبار زاي والد

Web as corpus

الويب كمدونة

Welch's independent samples

عينات ويلتش المستقلة

Wilcoxon rank-sum test

اختبار مجموع رتب ويلكسون

Z

Zipf's law

قانون زيف

Z-score

قيمة-ز

كشاف الموضوعات

اختبار مجموع رتب ويلكسون، ٢٥٢	أ
اختبار نسبة الاحتمالية، ١٤٧، ١٦١، ١٦٣	اتجاه التأثير، ٢٤٦
اختبارات القياسات المتكررة، ٢٥٤	الاتجاهية، ٩٣، ٩٤، ٩٧
اختبار-ت، ١٧، ١٨، ٤٠، ٤١، ٢٣٥	احتمالية الأفضلية، ٣٤٨، ٣٤٩، ٣٥٠
٢٤٠، ٢٤١، ٢٤٣، ٢٤٤، ٢٤٥	٣٥١
٢٤٧، ٢٥٠، ٢٥٢، ٢٥٤	اختبار إحصائي، ١٦، ١٩، ٢١٣، ٢٤١
٢٥٥، ٢٧٢، ٢٧٣، ٢٧٦، ٢٧٨	٢٩٤، ٢٤٣
٣٣٨، ٢٩٤	اختبار الاحتمالية اللوغارتمية، ١٤٧، ١٦٢
الاختصارات، ٢١٠، ٢١٧، ٢٢١، ٢٢٢	٣٣٠، ١٦٤
٢٢٦، ٢٢٣	اختبار دلالة الفرضية الصفرية، ١٦
ارتباط الخطية (المتعددة) بين المتنبئات، ١٥٨	اختبار زاي والد، ١٦٣، ١٦٦، ١٦٨
ارتباط إيجابي، ١٨٤، ١٨٥، ١٩٥	١٧٤، ١٧٧، ١٧٨، ١٧٩
ارتباط بيرسون، ١٨٤، ١٨٥، ١٨٧، ١٩١	اختبار كاي تربيع، ١٣٣، ١٤٦، ١٤٩
٢٣٣، ٢٥٤، ٣٠٠	١٥٢، ١٥٧، ١٦٤، ١٧١، ١٧٦
ارتباط سبيرمان، ١٨٤، ١٩٠، ١٩١، ٢٣٣	١٧٩، ٢٦٠، ٢٦١، ٢٦٣، ٢٩٤، ٣٣٠
ارتباط سلبي، ١٨٤، ١٩٥، ٢٢٢	اختبار كروسكال-والاس، ٢٥٣
	اختبار مان-ويتني U، ٢٥٢، ٢٥٣

- الارتباط، ١١٩، ١٢٧، ١٣١، ١٥٨، ١٨١، ١٩٧، ٢١٧، ٢٤٢، ٢٤٥، ٢٤٦، ٢٤٧، ٢٥٢، ٢٦٨، ٢٧٧، ٣٠١، ٣٤٥، ٣٤٠، ٣٣٨، ٣٠٥
- انحراف النسب، ٦٤، ٦٧، ٦٨، ٦٩، ٨٣
- انحياز الاختيار الذاتي، ٢٤
- الأهمية التطبيقية، ٢٨، ٤٧
- أرجحيات اللوج، ١٦٥
- الأرجحية، ٩٥، ١٠٨، ١٠٩، ١١٠، ١١١، ١١٢، ١١٣، ١٢٦، ١٥١، ١٦١، ١٦٣، ١٦٥، ١٦٦، ١٦٧، ١٦٨، ١٧٤، ١٧٧، ١٧٨، ١٧٩
- استخلاص العوامل، ٢٠٧، ٢١٢
- استقلال الملاحظات، ١٤٦، ١٥٧، ٢٤٣، ٢٤٨
- التباديل، ٩٥
- التباين، ١٧، ٦٠، ١٣٨، ٢١٩، ٢٣٥، ٢٣٩، ٢٤٠، ٢٤٢، ٢٤٣، ٢٤٤، ٢٤٨، ٢٤٩، ٢٥٠، ٢٥٢، ٢٥٤، ٢٥٥، ٢٥٨، ٢٦٣، ٢٦٦، ٢٦٨، ٢٧٦، ٢٧٧، ٢٧٨، ٢٩٤، ٣٤٠، ٣٤١، ٣٤٦، ٣٤٧
- تبعثر البيانات، ٢٥
- الاشترك اللفظي التعاقبي، ٢٨٦
- إطار العينة، ٢١، ٢٢، ٢٣، ٢٤
- أعمدة الخطأ، ٤١، ٢٥٤، ٢٨٩
- الافتراض الاعتدالي، ١٧
- الافتراضات، ١٢، ١٦، ١٧، ١٩٠، ٢٤٨
- انحدار منطقي، ١٦٧
- الانحراف المعياري، ٦، ١٥، ٦٢، ٦٣، ٦٤، ٦٥، ٦٦، ٦٩، ٨٣، ١٨٥، ١٨٧

- تجميع الجوار القائم على المتغيرية، ٢٩٩،
٣٠٠، ٣٠١، ٣٠٢، ٣٠٤، ٣٠٥،
٣١٦، ٣١٥
- التحليل البعدي، ١٩، ٥٩، ٣٢١، ٣٣٣،
٣٣٤، ٣٣٥، ٣٣٦، ٣٣٧، ٣٣٩،
٣٤٠، ٣٤١، ٣٤٢، ٣٤٣، ٣٤٧،
٣٥٢، ٣٥١
- تحليل التباين الأحادي، ٢١٩، ٢٢٨، ٢٤٧،
٢٤٨، ٢٥٠، ٢٧٦، ٣٢٧، ٣٤٦
- تحليل العامل، ٢٢٥، ٢٢٦
- تحليل متعدد الأبعاد، ٢٢٥
- التحيز الموضوعي، ٢٣
- تحيز عينة النص، ٢٢
- تحيز نوع النص التقليدي، ٢٣
- تحيز، ٢٢، ٢٣، ٣٣١، ٣٣٦
- التصاحبات اللفظية، ٨٧، ٨٨، ١٢٥
- تصميم البحث، ١، ٢٨، ٣٦، ٤٦، ١٤٠،
١٥٥، ١٧٥، ٣٣٦، ٣٣٧، ٣٣٨، ٣٤١
- تصميم المدونة، ٢٩، ٤٦، ١٣٣، ١٣٦،
٣٣٩
- تعريف المتوسط، ٢٦١
- التغاير، ١٨٥، ١٨٧، ١٩٠
- التفاوت المتماثل، ٢٤٤، ٢٤٨، ٢٦٠
- التفسير الوظيفي للعوامل، ٢٠٧، ٢١٦
- تكرار التصاحب المتوقع، ٩١
- تكرار التصاحب الملحوظ، ٨٩، ٩٢
- تكرار المتوسط المقلص، ٦٩، ٧٠، ٧١، ٧٢،
٧٣، ٨٠، ٨٣، ٨٤، ٨٥
- التكرار المطلق، ٤، ٢٩، ٥٤، ٥٥، ٥٦، ٥٧،
٥٩، ٦١، ٦٢، ٦٧، ٦٨، ٦٩، ٧٠، ٧١،
٧٣، ٧٨، ٧٩، ٨٢، ٨٣، ٨٤، ٩٤
- ١٠٩
- التمثيلية التعاقبية، ٢٨٤، ٢٨٧، ٣٢٠
- التمهيد، ٢٥١، ٢٨١، ٢٩٢، ٢٩٤، ٢٩٥،
٢٩٦، ٢٩٧، ٢٩٨، ٣١٤، ٣٢٠
- النواتج، ٣٣٤، ٣٣٧
- توزيع التكرار، ١١، ٤٧
- توزيع جولاند، ٦٤، ٦٦، ٦٧، ٧٨، ٧٩،
٨٣
- التوزيع، ٦، ١١، ١٢، ١٤، ١٥، ١٨، ٣٤،
٥٧، ٦٠، ٦٦، ٦٧، ٧٩، ١٣٥، ١٧٠،
١٩٥، ٢٤٤، ٢٤٧، ٢٥٢، ٢٥٤

ج

جداول النتائج المتقاطعة، ١٤١، ١٤٣،

١٥٠، ١٥٢، ١٧٩، ٢٥٨، ٢٦٠

الجدولة المتقاطعة، ٩٢، ٩٣، ١٣٣، ١٤١،

١٤٢، ١٤٣، ١٤٤، ١٤٥، ١٤٧،

١٤٨، ١٥٠، ١٥٢، ١٧٠، ١٧٣،

١٧٦، ١٧٩، ١٨٠، ٢٥٩، ٢٦٣، ٢٦٥

خ

خط الانحدار، ٥، ٣٥

خط بياني، ٢٨٨

خطاطة الأعمدة، ٣١، ٣٢، ١٣٤، ٢٩٢،

٣٢٤

الخطية، ١٨، ١٥٧، ١٥٨، ٢٨٨، ٢٨٩،

٢٩١

ح

حجم التأثير، ٦، ١٩، ٢٠، ٢٧، ٢٨، ٤٢،

١١٩، ١٤٨، ١٤٩، ١٥٠، ١٥٢،

١٦٦، ١٨٧، ١٩٤، ٢٤٥، ٢٥٢،

٢٥٣، ٢٥٤، ٢٥٥، ٢٩٦، ٢٩٧،

٢٩٨، ٣٢١، ٣٢٨، ٣٢٩، ٣٣٧،

٣٣٨، ٣٤٠، ٣٤١، ٣٤٢، ٣٤٤،

٣٤٥، ٣٤٦، ٣٤٧، ٣٤٨، ٣٤٩،

٣٥٠، ٣٥١، ٣٥٢

حجم المدونة، ٢٥، ٢٦، ٥٧، ٧٠، ٨٢،

٢٩٣، ٣٤٥

الحصر، ١٥٦، ١٦١، ١٦٣، ١٦٤، ١٦٥،

١٦٨، ١٧٤، ١٧٧، ١٧٨، ١٧٩،

٢٦٨، ٢٧٧، ٢٧٨

د

الدائس، ٩٣، ٩٥

درجة الحرية، ١٤٨، ٢٤٢، ٢٤٣، ٢٤٥،

دلتاب، ٩٣، ٩٤، ٩٥، ٩٧، ١٢٦،

ر

الربط المنفرد، ٢٠٠، ٢٠٢، ٣٠١،

رسم حدود المتغير، ٢٣٩، ٢٦٦،

س

السجلات اللغوية، ٢، ٧٧، ١٣٦، ٢٠٦،

٢٠٧، ٢٠٨، ٢١٢، ٢١٨، ٢١٩،

٢٢٠، ٢٢٣، ٢٢٤، ٢٢٦، ٢٢٨،

٢٣٣، ٣٣١،

سياق المتغير، ١٣٩

ش

شبكة التصاحب، ١٠٠، ١٠٢، ١٢٣
الشكل البياني للتصاحب، ٩٩، ١٠٠

ع

عدد الحالات، ١١، ١٣، ٩٧، ١١٩، ١٧١،
٢٤٣، ٢٤٤، ٢٤٩، ٢٥٠، ٣٤٠
العقدة، ٨٩، ٩٠، ٩١، ٩٢، ٩٣، ٩٤، ٩٧،
٩٨، ٩٩، ١٠١، ١١٥، ١١٦، ١٢٥،
١٢٦، ١٢٧، ٣١٠

علم الإحصاء الاستدلالي، ١٦

علم الإحصاء الوصفي، ٥، ١٩، ٢٥
عوامل خارجية، ٢٦٩

عينات ويلتش المستقلة، ٢٤٤

عينة عشوائية، ١٢٠، ١٢٤، ١٢٩، ١٣٧،
٢٦٧، ٢٩٤

عينة، ٤، ٨، ٢٠، ٢١، ٢٢، ٢٣، ٢٤، ٢٧،
٤٠، ٤٤، ٤٥، ٥٦، ٦٠، ٦٩، ١٠٧،

١١٤، ١٢٠، ١٢٤، ١٢٩، ١٣٧،

١٤٦، ١٥٠، ١٥٧، ١٥٩، ١٦٦،

١٦٨، ١٨٢، ٢٣٩، ٢٤١، ٢٤٢،

٢٤٣، ٢٦٧، ٢٧١، ٢٧٢، ٢٧٤،

٢٨٤، ٢٩٤، ٣٣٨، ٣٤١

ف

الفرضية الصفريّة، ١٦، ١٧، ١٨، ٢٧، ٢٨،
٤٠، ٤٥، ١١٠، ١٢٠، ١٤٧، ١٤٨،
١٨٧، ١٨٩، ١٩١، ٢٤٥، ٢٥٠، ٢٨٣

ق

قانون زيف، ٥٧، ٥٩، ٧٨، ٨٢، ٨٣، ٨٤
القسم والقصور، ٢٨١، ٣٠٥، ٣٠٦، ٣٠٧،
٣١٠، ٣١١، ٣١٢، ٣٢٠

قياس إحصائي، ١٩

القياس الأمثل، ٢٥٨

قياس النزعة المركزية، ١٣

قيم، ١٠، ١٣، ١٤، ١٥، ١٧، ١٨، ٢٧،

٣٢، ٣٤، ٤٥، ٤٧، ٥٧، ٦٢، ٦٣، ٧١،

٧٦، ٧٩، ٩٩، ١٠٢، ١١١، ١٢٠،

١٢٣، ١٢٩، ١٤٣، ١٤٧، ١٤٨،

١٤٩، ١٥٨، ١٥٩، ١٦٤، ١٧٤،

١٨٤، ١٩٢، ١٩٤، ١٩٥، ١٩٧،

٢١٧، ٢١٨، ٢٢٨، ٢٤٦، ٢٥٣،

٢٥٤، ٢٥٨، ٢٦٧، ٢٩٥، ٢٩٦،

٣٠٧، ٣١١، ٣٢٣، ٣٢٩

قيمة الاحتمال، ٦، ١٦، ١٧، ٤٥، ١١٠،

١٢٠، ١٤٨، ١٥٢، ١٦٣، ١٦٦،

اللوح دایس، ٩٥، ٩٩، ١٢٦، ٣٤٥	١٦٨، ١٧٤، ١٧٧، ١٧٨، ١٧٩
م	١٨٧، ١٨٩، ١٩١، ٢٤٥، ٢٥٠
المادة المعجمية، ٥١، ٥٢، ٥٣، ٧٨	٢٥٤، ٢٥٥، ٢٦٠، ٢٧٦، ٢٩٥
مبدأ الوثوقية، ٢٦٥، ٢٧٠	٢٩٦، ٢٩٧، ٢٩٨
المتصاحبات، ٥١، ٨٨، ١٠١، ١٠٢، ١٢٦	قيمة-ت، ٩٥، ٣٣٧، ٣٤٥
٣١٤	قيمة-ز، ٩٥، ١٩٧، ٢٠٠، ٢٠٥
متغير ترتيبي، ١٢٩	ك
متغير تقيمي، ١٥٦، ٢٦٧	كلمات الإيقاف، ١٠٤، ١٠٧، ١٢٨، ٢٨٣
متغير ناتج، ٢٧٠	٣٣١
متغير، ٦، ١٠، ١١، ٢٠، ٣٤، ٣٩، ٤٤	الكلمة الفعلية، ٥٠، ٥١، ٧٠، ٧٤، ٧٥
٤٧، ١١٥، ١٢٩، ١٣٣، ١٣٩، ١٤٣	٨٤، ٩٠، ٢٢١، ٢٢٧، ٢٧٠
١٥٠، ١٥٢، ١٥٣، ١٥٤، ١٥٦	الكلمة النوعية، ٥١، ٥٢، ٥٣، ٥٦، ٧٠
١٥٨، ١٦٧، ١٦٨، ١٧٧، ١٧٩	٧٤، ٧٥، ٧٨، ٢٢١، ٢٢٧، ٢٧٠
١٩١، ١٩٧، ٢١٣، ٢١٦، ٢١٧	كوهن-د، ١٥، ٢٠، ٢٧، ٩٣، ٩٥، ٩٧
٢٥٢، ٢٦٧، ٢٦٩، ٢٧٠، ٢٨٢	١١٢، ١١٣، ٢٤٥، ٢٤٦، ٢٤٧
٣٠٦، ٣٠٩	٢٥٢، ٢٥٥، ٢٧٦، ٢٩٦، ٢٩٧
متغيرات اسمية، ١٠، ١٣٧	٢٩٨، ٣٢٩، ٣٣٨، ٣٤٠، ٣٤١
متغيرات المتنبئ، ١٥٣، ١٥٦، ١٥٨، ١٧٣	٣٤٥، ٣٤٦، ٣٤٧، ٣٤٨، ٣٤٩
١٧٩، ٢٦٦، ٢٦٧	٣٥٠، ٣٥١
متغيرات لغوية، ١٠، ٣٤، ١٩١، ١٩٥	ل
٢١٢، ٢٢٢، ٢٤١، ٢٩٢، ٢٩٦	اللكسيم، ٥٢، ٥٣، ٧٨
	اللهجات الاجتماعية، ٢٣٨

١٨٧ ، ١٨٨ ، ١٨٩ ، ١٩١ ، ١٩٤ ،	متغيرات متعددة، ١٣٤ ، ٢٠٤ ، ٢١٢ ،
٢٤٢ ، ٢٤٤ ، ٢٤٦ ، ٢٤٧ ، ٢٥٢ ،	٢٧٨ ، ٢٦٣
٢٨٤ ، ٢٩٥ ، ٣٢٤ ، ٣٤٠ ،	متغيرات مستقلة، ١٤١ ، ١٥٣ ، ٢٦٨ ،
مجموعة بيانات، ٢٠ ، ٦١ ، ١٣٦ ، ١٣٧ ،	المتوسط الكلي، ٢٤٨ ، ٢٤٩ ،
١٤٧ ، ١٥٩ ، ١٦٢ ، ٢١١ ، ٢٥١ ،	المتوسط، ٤ ، ٥ ، ١٣ ، ١٤ ، ١٥ ، ١٨ ، ٢٠ ،
٢٩٦ ، ٣٢٤ ، ٣٣٧ ،	٣٢ ، ٤٢ ، ٤٣ ، ٤٤ ، ٥٥ ، ٦٢ ، ٦٣ ، ٦٤ ،
مخطط التشتت، ٣٤ ، ٣٥ ، ٢٦١ ، ٢٦٤ ،	٦٥ ، ٦٩ ، ٧٢ ، ٧٦ ، ٨٠ ، ٨٤ ، ٩٥ ،
مخطط الشموع، ٢٨٩ ، ٢٩١ ، ٣١٣ ، ٣١٤ ،	١٣٥ ، ١٨٤ ، ١٩٠ ، ١٩٧ ، ٢٠٠ ،
مخطط درجي، ٣٤ ، ٣٥ ،	٢٠٣ ، ٢٤٢ ، ٢٤٨ ، ٢٤٩ ، ٢٥٥ ،
مخطط رسم المربع، ٣٢ ، ٤٠ ، ٤١ ، ٢٩٠ ،	٢٦١ ، ٢٧٤ ، ٢٨٩ ، ٢٩٤ ، ٣٠١ ،
٣٣٢	٣٠٥ ، ٣٣٨ ، ٣٤٠ ، ٣٤٤ ، ٣٤٨ ،
مخطط ركامي، ٢١٥ ، ٣٠٤ ، ٣٠٥ ،	مجال الثقة، ٦ ، ١٨ ، ١٩ ، ٤٥ ، ١٥٢ ، ١٦٣ ،
المدخل الثابت، ١٦٠	١٦٦ ، ١٦٧ ، ١٦٨ ، ١٧١ ، ١٧٤ ،
المدونات التعاقبية، ٢٨٣	١٧٧ ، ١٧٨ ، ١٧٩ ، ١٨٧ ، ١٨٩ ،
المدونة الحاسوبية، ١ ، ٢ ، ٣ ، ٤ ، ٦ ، ١٣ ،	١٩٥ ، ٢٢١ ، ٢٢٢ ، ٢٤٦ ، ٢٥٤ ،
١٥ ، ١٩ ، ٢١ ، ٢٧ ، ٢٩ ، ٤٧ ، ٥٢ ، ٥٣ ،	٢٥٥ ، ٢٩٧ ، ٣٠٦ ، ٣١١ ، ٣٣٤ ، ٣٤٠ ،
٥٥ ، ٦٢ ، ٨٨ ، ١٥٠ ، ١٦١ ، ١٩٩ ،	مجاميع الأعمدة والصفوف، ١٤٢
٢٠٤ ، ٢٦٩ ، ٢٧٠ ، ٢٨٧ ، ٣٠١ ،	المجموع الكلي، ٢٦ ، ١٤٢ ، ١٤٣ ، ١٤٤ ،
٣٢١ ، ٣٢٢ ، ٣٢٨ ، ٣٣٠ ، ٣٣١ ،	١٤٩ ، ١٧٦ ، ١٨٥ ،
٣٣٦ ، ٣٤٤ ، ٣٤٨ ،	المجموع اللغوي، ٨ ، ١٧ ، ١٨ ، ١٩ ، ٢٠ ،
المدونة الرئيسية، ١٠٥ ، ١٠٧ ، ١٠٨ ، ١٠٩ ،	٢١ ، ٢٢ ، ٢٥ ، ٢٦ ، ٢٧ ، ٣٣ ، ٤٤ ، ٤٥ ،
١١٠ ، ١١٢ ، ١٢٨ ، ١٢٩ ،	٤٧ ، ٦٣ ، ١٠٧ ، ١٥١ ، ١٦٦ ، ١٦٧ ،

المدونة المرجعية، ١٠٥، ١٠٨، ١٠٩، ١١٠،	مقياس، ٤، ١٠، ١٣، ٤٥، ٥٧، ٦٠، ٦١،
١١١، ١١٢، ١١٣، ١٢٨	٦٢، ٦٥، ٦٦، ٦٧، ٦٩، ٧٥، ٨٤، ٨٨،
المدى الربيعي، ١٥، ٣١	٩٤، ٩٧، ١٠٥، ١١١، ١١٧، ١٢٠،
المدى، ١٥، ٣١، ٦١، ٦٢، ٦٦، ٦٨، ٧٨،	١٢٥، ١٣٠، ١٣١، ١٥٠، ١٥١،
٧٩، ٨٣، ٨٤، ٨٩، ٩٢، ٩٨، ١١٩،	١٥٥، ١٦٣، ١٦٤، ١٦٦، ١٧٢،
المدى ١، ١٥، ٦١	١٧٧، ١٧٨، ١٧٩، ٢٠٠، ٢١٤،
المدى ٢، ١٥، ٦١، ٦٢	٢١٧، ٢١٨، ٢٤٥، ٢٤٦، ٢٥٢،
مسافة كاي تربيع، ٢٥٨، ٢٦١، ٢٦٢،	٢٥٨، ٢٨٩، ٣٠٥، ٣٣٧، ٣٤٤،
٢٦٣، ٢٦٥	٣٤٥، ٣٤٦، ٣٤٧، ٣٤٨، ٣٤٩،
مستويات المرجع، ١٥٨	ملاءمة المُقدَّر البيني، ١١٤، ١١٦، ٢٦٧،
معامل التحديد، ١٩١، ٣٤٥	٣١٠، ٣١٢، ٣٤٣،
معامل التنوع، ٦٤، ٦٥، ٦٦، ٦٧، ٨٣، ٣٠١،	ملفات التعريف، ٢٦١
معامل، ٦٤، ٦٥، ٦٦، ٦٩، ٨٣، ١٤٩،	المواءمة الخام، ١١٧، ١١٨،
١٥٢، ١٥٤، ١٥٦، ١٧٩، ١٨٧،	ن
١٨٩، ١٩١، ١٩٤، ١٩٥، ٢٥٤،	ناتج الاكتراث، ١٥٤، ١٥٨، ١٦٦، ١٦٧،
٢٧٠، ٣٠٠، ٣٠١، ٣١١، ٣١٢،	نافذة التصاحب اللفظي، ٨٩،
٣٤٥، ٣٤٦، ٣٤٧	نسبة الاحتمال، ١٥٠، ١٥١، ١٥٢، ١٦٤،
المعلومات المتبادلة، ٩٣، ٩٥، ٩٨، ٩٩،	١٧٧، ١٧٨، ١٧٩، ٢٦٨، ٢٧٨،
١٠٢، ١٢٣، ١٢٦، ٣١١، ٣١٢،	٣٣٠، ٣٤٦، ٣٥١،
٣١٥، ٣٤٥	نسبة الأرجحية، ٢٠، ٩٧، ١٥١، ١٦٥،
معيار معلومات أكايكي، ١٦٠، ١٦١،	١٦٦، ١٦٧، ١٦٨،
١٦٢، ١٦٣، ١٧٧، ١٧٨، ١٧٩، ٢٧٧،	نسبة الخطر، ١٥٠

نسبة الكلمات النوعية إلى الكلمات الفعلية،	نقطة البيانات، ٣١، ١٨٥
٨٤، ٨٢، ٧٦، ٧٥، ٧٤، ٧٣	نمذجة التأثير الممتزج، ١٥٧، ٢٣٥، ٢٦٥،
نسبة اللوج، ٩٥، ٩٧، ١١٢، ١١٣	٢٦٨، ٢٦٩، ٢٧١، ٢٧٧، ٢٧٨
نسبة متوسط الكلمة النوعية/الفعلية	النموذج المقتر، ١٦٠
المتحرك، ٧٥	نموذج المنطلق (أو الصفر)، ١٦٠
النقاط في الشكل البياني، ٢٦١	٩
النقاط، ٦، ١١٩، ١٢٧، ١٤٨، ١٩٦،	الوسيط، ٥، ١٤، ١٥، ٢٥٦، ٢٧٧
١٩٧، ٢٦١، ٢٨١، ٢٩٣، ٣٠٨،	الويب كمدونة، ٢٤
٣٠٩، ٣١٠، ٣١٢، ٣٢٣، ٣٤٩	

