



علم الإحصاء في لغويات المدونة الحاسوبية

دليل عملي

تأليف

Vaclav Brezina

ترجمة

د. سلطان بن ناصر المجيول

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

ح دار جامعة الملك سعود للنشر، ١٤٤٣ هـ (٢٠٢١ م)

فهرسة مكتبة الملك فهد الوطنية أثناء النشر

بريزينا، فاكلاف.

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

٣٩٥ ص؛ ١٧ سم × ٢٤ سم

ردمك: ١ - ٩٤٧ - ٥٠٧ - ٦٠٣ - ٩٧٨

١- الاحصاء ٢ - الطرق الاحصائية أ. المجيلول، 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 ١٤٤٢ هـ المعقود

بتاريخ ١٩/٦/١٤٤٢ هـ الموافق ١/٢/٢٠٢١ م

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 سيُمثل 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 ووفّر أيضاً الأكواد الخاصة بتلك التحليلات، وذلك كله في موقع أدوات لانكستر الإحصائية.

تعتمد لغويات المدونة الحاسوبية على البيانات اللغوية، ومن المهم أن يُدرك القارئ أنّ المدونة corpus في هذا الكتاب تشير إلى مفهوم البيانات اللغوية linguistic data. وقد مرّ مصطلح corpus بصفته مدونةً حاسوبيةً بثلاث مراحل زمنية متتابعة، 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 ورموزٌ تُمثّل حقائقٌ وأرقامًا تفصيلية عن حالة (لغوية) معينة تُدرس وتُعالج بصفته وحدة كاملة. وسيتدرّب 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 والرميز 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: التعرّف 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 مكارني Tony McEnery ودانا جابلاسوفا Dana Gablasova على دعمهما غير المتقطع وعلى تشجيعهما في أثناء فترة تأليف هذا الكتاب، وأشكرهما على ملحوظاتها الدقيقة على كل فصل من فصول الكتاب. تلقى هذا الكتاب عدداً من الملحوظات المستنيرة من جابرييلي 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 حقلًا 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 تعقيداتهما. 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 وُقِف على الإحصاءات وقوفًا مقننا، ويُركز الكتاب في الأوجه التطبيقية لتحليل 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 علمًا منهجيًا متعدد الاستعمالات لتحليل اللغة، وتطبيقًا لحقول عدّة كاللغويات، والعلوم الاجتماعية، والإنسانيات 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. وبناء على ذلك، يتضمن هذا الكتاب أيضًا أسئلة للمناقشة (مباحث: فِكر Think about)، وتمارين يمكن للقراء التدرّب 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 زُوّد 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 مستخدم تلك الأدوات ممن ليسوا على مَلاك مؤسسات تدعم لهم الحزم 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). أو من بأن علم الإحصاء ينبغي ألا يكون عصيًا عن الاستيعاب، لكون الحاسب الآلي قادرًا على مضع 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.....
- ٦٩..... (٢, ٦) التنوع المعجمي: نسبة الكلمات النوعية إلى الكلمات الفعلية والنسبة الموحدة والنسبة المتوسطة المتحركة بينها.....
- ٧٣..... (٢, ٧) التطبيق وأمثلة أخرى: هل يتحدث البريطانيون عن الطقس طوال الوقت؟.....
- ٧٦..... (٢, ٨) تمارين.....
- ٨١..... قراءات إضافية.....
- ٨٥..... أدوات لانكستر الإحصائية.....

الفصل الثالث: علم الدلالة والخطاب: التصاحبات

اللفظية، والكلمات المفتاحية، وموثوقية الترميز اليدوي

- ٨٧..... (٣, ١) ما الذي يتناوله هذا الفصل؟.....
- ٨٨..... (٣, ٢) التصاحبات اللفظية ومقاييس التصاحب.....
- ٩٩..... (٣, ٣) مخططات التصاحب والشبكات: استكشاف المتجاورات المتقاطعة.....
- ١٠٣..... (٣, ٤) الكلمات المفتاحية وكلمات الإيقاف.....
- ١١٤..... (٣, ٥) مقاييس ملاءمة المُقدَّر البيئي.....
- 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 والفردى

- (٦, ١) ما الذى يتناوله هذا الفصل؟ ٢٣٥
- (٦, ٢) الأسلوب الفردى والتنوع الاجتماعى: 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 الفردى: تحليل التناظر ٢٥٧
- (٦, ٥) السياق اللغوى: 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 الذى يتناوله هذا الفصل؟ ٢٨١
- (٧, ٢) قياس الزمن وتمثله بوصفه متغيرًا ٢٨٢
- (٧, ٣) إيجاد الاختلافات وتفسيرها: تغير النسبة المئوية واختبار التمهيد ٢٩٢
- (٧, ٤) تجميع الفترات الزمنية: تحليل التجميع المجاور ٢٩٩
- (٧, ٥) نمذجة التغيرات فى الخطاب: القمم والقعور وتحليل التقلب الاستعمالى ٣٠٥
- (٧, ٦) التطبيق: الألوان فى 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 ما الذي يتناوله هذا الفصل؟ (٨, ١)
٣٢٢ عشرة مبادئ للتفكير الإحصائي (٨, ٢)
٣٣٣ التحليل البعدي: التوليف الإحصائي لنتائج الأبحاث (٨, ٣)
٣٤٣ أحجام التأثير: دليل للاستخدام الصحيح (٨, ٤)
٣٥٠ تمارين (٨, ٥)
٣٥٢ قراءات إضافية
٣٥٢ أدوات لانكستر الإحصائية
٣٥٣ ملحوظات ختامية
٣٥٥ قائمة المراجع
٣٦٣ ثبت المصطلحات: أولاً: (عربي-إنجليزي)
٣٧٥ ثانياً: (إنجليزي-عربي)
٣٨٧ كشف الموضوعات