

Language & IT

Dr. Abdullah Al Fraidan

Lecture10-Corpus Linguistics

علم لغة كوربوس-Corpus Linguistics

SOME INSIGHTS OBTAINABLE FROM CORPORA

- بعض الأفكار يمكن الحصول عليها من
- ... But not maybe all obtainable by us from corpora we have free access to...
- ولكن لا يمكن الحصول عليها ربما كل من قبلنا من corpora لدينا حرية الوصول إلى ...
- Most of these have fairly obvious use for both descriptive linguists and teachers... and maybe learners too (and others in the range of users)
- معظم هذه لها استخدام واضح إلى حد ما لكل من واصفي اللغويين والمعلمين ... وربما المتعلمين جدا (وغيرها في نطاق من المستخدمين)
- General English
- الإنجليزية العامة
- Frequencies of individual words across varieties: certain and sure
- ترددات من الكلمات الفردية عبر أصناف: محدد ومتأكد
- Characteristics of varieties and individual authors: frequencies overall; TTRs
- خصائص الأنواع والفردية للمؤلفين: الترددات الشاملة؛ TTRs
- Details of meaning of vocabulary items and collocation: qualitative details of synonyms sad-unhappy; mutual information for money and flatly
- التفاصيل في معنى المفردات والعناصر والتجميع: التفاصيل النوعية في المرادفات - حزين- تعيس؛ المعلومات المتبادلة - المال وبصراحة

- Homonym and sense frequencies: lookout
- Lexical grammar: verbs used with that clauses
- **Grammar: uses of with**
- **النحوي: يستخدم مع**
- Use of words with a heavy pragmatic dimension: flipping, right
- استخدام الكلمات ذات البعد الثقيل العملي : flipping, right
- Lexical phrases: You know what I mean...
- العبارات المعجمية: ... You know what I mean
- Translation
- الترجمة
- Frequency of translation equivalences
- تردد في معادلات ترجمة
- Learner Language
- لغة المتعلم
- Error and performance analysis information from teacher-made mini-corpora of their learners' language
- الخطأ و تحليل أداء المعلومات من المعلمين من صنع مصغر corpora في لغة المتعلمين '
- Ditto from large corpora of learner language
- ديتو من corpora كبيرة في لغة المتعلم
- Frequency of types of lexical error
- تردد في أنواع الخطأ المعجمية
- Research on error correct ability by dictionaries
- البحث في خطأ القدرة الصحيح من قبل القواميس
- **GENERAL TYPES OF CORPUS TASK/ INVESTIGATION**
- **الأنواع العامة كوربوس المهمة / التحقيق**
- It is possible classify most corpus projects, or generate new ones, as combinations of choices from these main dimensions (for any given language, assumed to be English here):
- من الممكن تصنيف المشاريع الأكثر corpus ، أو إنشاء جديدة، و(يفترض لأي لغة معينة، لتكون إنكليزية هنا) مجموعات من الخيارات من هذه الأبعاد الرئيسية:

1) What kind of language is of interest?

- ما هو نوع اللغة هي الفائدة؟
- From normal native speaker adults today.
- من الطبيعي اللغة الأصلية للبالغين هي اليوم.
- Then it could be spoken or written, standard or non-standard, UK or US or..., from everyday language or the specialist register of newspapers or poetry or academic prose or...etc.
- ثم يمكن منطوقة أو مكتوبة، ومعياري أو غير قياسي، المملكة المتحدة أو الولايات المتحدة أو ...، من اللغة اليومية أو سجل متخصص من الصحف أو الشعر أو النثر أو الأكاديمية ... الخ
- From the past. Literary or not...
- من الماضي. أدبي أو لا ...
- from foreign language learners
- من متعلمي اللغة الأجنبية
- from normal native speaker children
- من الطبيعي اللغة الأصلية للأطفال
- from speakers with language disabilities (e.g. aphasics)
- من المتحدثين ذوي إعاقة اللغة (على سبيل المثال الحبسة)

2) What level(s) of language are you interested in?

- ما هو مستوى اللغة التي تحبها؟
- vocabulary/lexis
- المصطلحات / المفردات
- grammar/syntax
- قواعد اللغة / بناء الجملة
- sounds, intonation
- الأصوات -التجويد
- spelling, punctuation
- التهجية - الترقيم
- text/discourse/rhetorical structure
- النص / الخطاب / البنية الخطابية
- pragmatics
- البراغماتية

- Spoken? How natural are speeches, TV etc.?
- التكلّم ؟ كيف الطبيعية هي للخطب، والتلفزيون وما إلى ذلك؟
- Fully natural: observer's paradox and how to be ethical?
- Permission. Labov's tricks
- الطبيعية تماما: تناقض المراقبين وكيف تكون أخلاقية؟ تراخيص. Labov
- Records of speakers (and addressees and...)
- سجلات المتكلمين (والمرسل و...)
- Transcription issues: what to transcribe and who does it (expert or not)
- قضايا النسخ: ما التدوين ومن يفعل ذلك (خبير أو لا)
- Random sampling again; problem of accents and dialects
- العينات العشوائية مرة أخرى؛ مشكلة من اللكنات واللهجات
- Analysis - how to extract useful information automatically?
- التحليل - كيفية استخراج المعلومات المفيدة تلقائيا
- frequency and its derivatives:
- التردد ومشتقاته:
- range: over text types
- المدى: أنواع فوق النص
- richness of vocab: TTR
- ثراء فوكب: TTR
- collocational strength: mi and t-score/z score
- تجميع قوتها: mi و t- والنتيجة z
- How to relate *go*, *goes* and *went*? Lemmatisation
- كيفية ربط *go*, *goes* و *went* ؟ المجموعات الصرفية
- Concordance: the problem of large numbers. Qualitative into quantitative
- التوافق: مشكلة الأعداد الكبيرة. النوعية في الكمية

- how to distinguish *right* from *right*: pos and other annotation/tagging
- كيفية التمييز بين الصواب والحق: POS وغيرها الشرح / العلامات
- How to sort and select from a KWIC listing?
- كيفية تصنيف والاختيار من بين قائمة KWIC؟
- Accessibility to general users – cost, computers etc.
- الوصول إلى عامة المستخدمين - التكاليف، أجهزة الكمبيوتر الخ

3) What is the purpose of the investigation? More language research, or more pedagogical?

- ما هو الغرض من التحقيق؟ المزيد من بحوث اللغة، أو أكثر تربوية؟
- To describe an aspect of language or compare different styles, authors etc. I.e. more exploratory research.
- لوصف جانباً من جوانب اللغة أو مقارنة أنماط مختلفة، أي المؤلفين الخ. المزيد من البحوث الاستطلاعية.
- To check on a proposed 'rule' or past finding or a theory-based prediction in some area of language study. I.e. more hypothesis testing research.
- للتحقق على 'القاعدة' المقترحة أو حقائق الماضي أو التنبؤ المستند إلى نظرية لبعض مناطق دراسة اللغة. أي أكثر فرضية لأختبار البحوث.
- to test out a parser that some computational linguists have designed
- لاختبار محلل لبعض اللغويين الحسابية بالتصميم
- to help create language syllabuses or teaching materials
- للمساعدة على إنشاء مناهج اللغة أو المواد التعليمية
- to help evaluate syllabuses or teaching materials
- للمساعدة في مناهج تقييم أو المواد التعليمية
- to use or evaluate corpus work as a class task (i.e. a form of CALL)
- لاستخدام أو تقييم عمل كوربوس كفئة مهمة (أي شكل من أشكال CALL)

- to help write a dictionary or grammar book
- للمساعدة في كتابة قاموس أو كتاب نحوي
- To help evaluate a dictionary or grammar book etc.
- للمساعدة في تقييم القاموس أو الكتاب النحوي الخ

4) What kind of corpus information is needed?

- ما هو نوع معلومات كوربوس المطلوبة ؟
- More concordance-type information - examples of occurrences of things in context to analyse. I.e. qualitative
- أكثر أنواع المعلومات المتوافقه - أمثلة من أحداث الامور للتحليل في سياقها الصحيح. أي النوعية
- More frequency information about words or whatever. I.e. quantitative
- المزيد من المعلومات حول تردد الكلمات أو أي كلمة. أي الكمية
- Most of the combinations implied above are possible to some extent with existing corpora. However, they are not all available to us here.
- معظم المجموعات الضمنية أعلاه هي ممكن إلى حد ما مع corpora الموجودة. ومع ذلك، فهي ليست متاحة لنا جميعا هنا.