Language & IT

Dr. Abdullah Al Fraidan

Lecture- 9-Corpus Linguistics

Corpus Linguistics
علم لغة كوربوس

- **CURRENT GENERAL CORPUS ISSUES**
- القضايا الحالية العامة لكوربوس
- Corpus versus introspection. Is there a separate 'Corpus Linguistics'?
- كوربوس مقابل التأمل. هل هناك 'علم لغة مستقل لكوربوس " ؟
- Let the data speak for itself? (Sinclair)
- لندع البيانات تتحدث عن نفسها؟ (سنكلير)

I-language versus E-language (Chomsky)

- I-اللغة مقابل اللغة E (تشومسكي)
- Missing context, intention, 'ethnographic' information. Third person not 1st person view….(Widdowson)
- السياق مفقود ، القصد، والمعلومات 'الاثنوجرافية'. الشخص الثالث لا مشاهدة شخص ١.... (ودوسون)
- Corpus can't show what doesn't occur, or all that can occur
- لا يمكن أن يظهر كوربوس ما لا يحدث، أو كل ما يمكن أن يحدث
- Introspection may be surprised by what does occur
- قد يفاجأ التأمل من قبل ما يحدث
- Areas of language that corpora don't illumine
- مجالات اللغة التي corpora لا تنير
- Size of corpus and individual word frequency. How big should it be?
- مقاس كوربوس وتردد الكلمة الفردية. كيفية الكبير ينبغي أن يكون؟

- Cost effectiveness - more running words doesn't give more different words proportionally
- فعالية التكاليف - أكثر عمل للكلمات لا تعطي أكثر اختلافا نسبيا للكلمات
- 10-20 hours to process 2000 words of speech (prosodic tagging)
- ١٠-٢٠ ساعة لمعالجة ٢٠٠٠ كلمة في الكلام (علامات prosodic)
- Just because a population is vast does not mean samples have to be vast to be representative, as some think… Depends on feature of interest and variability.
- لمجرد ان السكان واسع لا يعني العينات يجب أن تكون واسعة لأن تمثل، كما يظن البعض ... يعتمد على ميزة في الفائدة والتباين
- Word frequency problem
- مشكلة تكرار الكلمة
- Static or dynamic (monitor) corpora?
- ثابت أو حيوي (راصد) corpora
- Sampling and how to be representative e.g. of general English?
- أخذ العينات وكيفية أن تمثل مثلا الإنجليزية العامة؟
- Any collection of texts is not a useful (principled) corpus. Problems…
- أي مجموعة نصوص ليست مفيدة كوربوس (مبدئية). المشاكل

- Opportunistic - biased to written, accessible varieties?
- الانتهازية - منحازة إلى الخطية، وأصناف الوصول إليها؟
- Systematic- balanced and representative: a corpus of corpora
- منهجية متوازنة وممثلة: أ كوربوس في corpora
- Exclude non-standard?
- تستبعد غير القياسية؟
- What national varieties?
- ما الأنواع الوطنية؟
- How far back?
- كيف يعود؟

- What proportions of varieties?
- ما أنسب الأنواع؟
- Speaker/writer factors as well (demographics)? Problem more with written than spoken (L1 from name?). Addressee
- المتكلم / كذلك عوامل الكاتب (الديموغرافية)؟ المشكلة أكثر مع مكتوبة من المنطوقة (L1 من الاسم؟). المرسل اليه
- Then: Random selection?
- ثم: الأختيار العشوائي؟
- Stratified sampling? What varieties?
- العينات الطبقية؟ ما الأنواع؟
- Weighting by how much read or by 'influence'? Expert judgment
- الوزن قبل كم القراءة أو بواسطة 'النفوذ'؟ آراء الخبراء
- Even genres like ‘academic writing’ are not homogeneous: depend on sub discipline (Business and Econs *I*, Computing and Physics *we*), genre within sub discipline (review, report), even the lecturer being written for
- حتى الأنواع مثل "الكتابة الأكاديمية" ليست متجانسة: تعتمد على الانضباط الفرعي (الأعمال وEcons الأول، الحاسبات والفيزياء)، والنوع داخل الانضباط الفرعية (الاستعراض، التقرير)، وحتى يتم كتابة المحاضرة
- How to sample each text, and sample size again? Copyright issues
- كيفية أخذ عينات كل نص، وحجم العينة مرة أخرى؟ قضايا حق المؤلف

- Spoken? how natural are speeches, TV etc.?
- تكلم ؟ كيف طبيعية الخطب، والتلفزيون وما إلى ذلك؟
- Fully natural: observer’s paradox and how to be ethical? Permission. Labov’s tricks
- طبيعي تماما: مفارقة المراقب وكيف تكون أخلاقية؟ تراخيص. الحيل Labov
- Records of speakers (and addressees and…)
- السجلات في المتكلمين (والمرسل و...)
- Transcription issues: what to transcribe and who does it (expert or not)
- قضايا النسخ: ما لتدوين 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: عبر أنواع نصية
- richness of vocab: TTR
- غنى فوكب
- collocation AL strength: mi and t-score/z score
- قوة التجميع: mi و-t نتيجة z
- how to relate *go*, *goes* and *went*? Lemmatisation
- كيفية ربط*go, goes went* و - *went* المجموعات الصرفية
- concordance: the problem of large numbers. Qualitative into quantitative
- التوافق: مشكلة الأعداد الكبيرة. النوعي في الكمي
- 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 فرز والاختيار من بين قائمة KWIC؟

- Accessibility to general users – cost, computers etc.
- الوصول إلى عامة المستخدمين - التكاليف، أجهزة الكمبيوتر الخ
- The above issues all repeat for learner corpora. Further, issues (see ICLE solutions):
- القضايا المذكورة أعلاه كل تكرار للمتعلم المجاميع. علاوة على ذلك، القضايا (انظر حلول ICLE):
- What counts as a learner? Cf ICE
- ما يعتبر المتعلم؟
- Information about learner language that is not reflected in a learner corpus
- معلومات عن لغة المتعلم الذي لا ينعكس في المتعلم كوربوس
- What counts as 'authentic' for learners?
- ما يعتبر 'حقيقية' للمتعلمين؟
- Apart from L1, what variables would you want to have documented about the students and the tasks/setting for any collection of learner material in a corpus? (Cf Granger 2002 discussion) These all may make a difference
- وبصرف النظر عن L1، ما هي المتغيرات التي تريد أن توثق عن الطلاب والمهام / وضع لأي جمع مواد متعلم في كوربوس؟ (راجع مناقشة جرانجر ٢٠٠٢) كل هذه قد تحدث فرقا

- Problem therefore of comparability of such corpora collected by different people in different countries
- بالتالي المشكلة في المقارنة بين هذه المجاميع التي تم جمعها corpora من قبل مختلف الناس في مختلف البلدان
- Possibility of longitudinal corpora
- إمكانية طوليةcorpora
- Contrastive interlanguage analysis
- تحليل اللغة التقابلية
- NNS – NS To find errors and over/under use. But issues of:
- للعثور على الأخطاء وأكثر من / تحت الاستخدام. لكن القضايا
- Comparability of variety
- مقارنة متنوعة
- Linguistic imperialism (terms like *error*, *overuse*), but problem of learners' real wishes and lack of information on 'international proficient speaker English'
- الإمبريالية اللغوية (مصطلحات مثل الخطأ، الإفراط)، ولكن المشكلة في رغبات حقيقية ونقص المعلومات عن 'المتعلمين -الدولي يتقن الإنجليزية'
- NNS – NNS To distinguish transfer and non-transfer (e.g. developmental) errors.
- أخطاء - للتمييز والنقل وغير النقل (على سبيل المثال التنموي).
- Comparability again
- المقارنة مرة أخرى
- Parallel L1 corpus of the learners would be useful
- وبالتوازي كوربوس L1 للمتعلمين تكون مفيدة
- Computerised error analysis
- تحليل الأخطاء بالكمبيوتر
- **Method 1**: Think of an error and search for it
- الأسلوب ١: فكر في الخطأ وابحث عنه
- **Method 2**: Tag all errors in corpus and then search
- الأسلوب ٢: علامة كافة الأخطاء في كوربوس ثم البحث