

# Language & IT

Dr. Abdullah Al Fraidan

## Lecture- 8-Corpus Linguistics

### Corpus Linguistics

- A corpus is a collection of language material, made in some principled way (not haphazardly), either on tape or written in hard copy (e.g. books, student essays) or in electronic form.
- **corpus** كوربوس هو عبارة عن مجموعة من المواد اللغوية التي وردت في بعض الطرق المبدئية (وليس عشوائيا)، إما على شريط أو مكتوب في نسخة مطبوعة (مثلا الكتب ومقالات الطالب) أو بشكل إلكتروني.
- We are concerned only with the last type.
- **إننا نشعر بالقلق فقط لنوع آخر.**
- Such collections are used in many different ways by different people.
- **تستخدم هذه المجموعات بطرق عديدة ومختلفة من قبل مختلف الناس.**
- We are concerned mainly with use
- **إننا نشعر بالقلق أساسا مع الاستخدام**
- 1) by linguists to help describe language, and test theories
- **من قبل اللغويين للمساعدة في وصف اللغة، واختبار النظريات**
- 2) by teachers and learners to aid language learning (i.e. a form of CALL).
- **من قبل المعلمين والمتعلمين لمساعدة تعلم اللغة (أي شكل من أشكال CALL).**
- To perform any electronic corpus-based task directly you **need two things - a corpus and a search engine.**
- **لأداء أي مهمة مستندة إلى corpus كوربوس الإلكترونية مباشرة تحتاج لأمرين:**
- **corpus كوربوس ومحرك البحث**

- A corpus itself is just text (a form of data), which may have been originally written, or be transcribed speech.
- كوربوس نفسه هو فقط نص (شكل من أشكال البيانات)، وقد يكون قد كتب في الأصل، أو يكون خطاب مدون
- Corpora are not all stored in the same format (though often they are in the plainest of DOS or ASCII text), and they may have coded information (tags) added in and out of the text, to show e.g. who was speaking, the register of the text, or the part of speech of each word.
- لا يتم تخزين جميع Corpora بنفس التنسيق (بالرغم غالبا ما تكون في أبسط من نص DOS أو ASCII)، وأنها قد تكون مشفرة المعلومات (العلامات) وأضاف في دخول وخروج من النص، لإظهار مثلا الذي كان يتحدث، وسجل النص، أو جزء من كلمة لكل كلمة
- To use a corpus for any task you have to access it by using a search engine - a program which generally runs through the text (or a precompiled index to the text) and broadly does one of two things:
  - لاستخدام corpus كوربوس بأية مهمة لديك للوصول إليها باستخدام محرك البحث - عموما البرامج تمتد من خلال النص (أو فهرس مترجم مسبقا للنص) ويفعل على نطاق واسع لأحد الأمرين:

## • USERS OF CORPORA

- Dictionary makers - e.g. to find out how words are actually used, and how often, and improve dictionary entries
  - واضعي القاموس - مثلا لمعرفة كيفية استخدام الكلمات فعليا، وكيف أحيانا، تحسن مداخل القاموس
- Descriptive grammarians - e.g. to improve their descriptions to fit the facts of actual use of constructions
  - وصف النحويين - مثلا لتحسين أوصافها لتناسب وقائع الاستخدام الفعلي للإنشاءات
- Stylisticians - e.g. to see what differences there are in how frequently different authors use certain words
  - Stylisticians - مثلا لنرى ماذا هناك اختلافات في مدى تكرار استخدام مختلف الكتاب لبعض الكلمات

- Sociolinguists - e.g. to see how frequent certain constructions are in conversation  
• علماء اللغويين الاجتماعيين - مثلا نرى كيف تكرر بعض الإنشاءات في المحادثة
- Computational linguists - e.g. to see if their grammatical parsing programs will work on naturally occurring language  
• اللغويين الحسابية - مثلا لمعرفة ما إذا برامجها تحلل النحوية بالعمل على اللغة التي تحدث بشكل طبيعي
- Language learning researchers - e.g. to see how often learners with a particular L1 get something wrong  
• باحثين تعلم اللغة - مثلا لمعرفة عدد مرات المتعلمين مع L1 خاصة الحصول على شيء خاطئ
- Writers of teaching syllabuses - e.g. to see how often the passive really occurs in academic English  
• الكتاب من مناهج التدريس - مثلا لنرى كيف كثير من الأحيان السلبية تحدث حقا في الأكاديمية الإنجليزية
- Writers of teaching course materials - e.g. to incorporate authentic examples into their material  
• الكتاب في تدريس المواد الدراسية - مثلا لإدراج أمثلة أصيلة إلى وضعهم المادي
- Teachers making class tasks, or even learners directly themselves - e.g.  
• على سبيل المثال المعلمين لجعل درجة للمهام ، أو حتى المتعلمين أنفسهم مباشرة -
- to supply additional clues for context guessing word meaning  
• لتوفير أدلة إضافية لسياق تخمين معنى الكلمة
- for guidance on how to use word when writing  
• للحصول على إرشادات حول كيفية استخدام الكلمة عند الكتابة
- to help prompt self-correction  
• للمساعدة في موجه التصحيح الذاتي
- for word study  
• لدراسة الكلمة
- for 'language awareness' work on grammar  
• للعمل على "وعي اللغة بالقواعد

التاريخ في-History of Corpora

<http://privatewww.essex.ac.uk/~scholp/corpinintro.htm#hist>