A corpus is A collection of language material, made in some principled way either on tape , written in hard copy, or in electronic form.

المجمع : وكتب في نسخة ورقيه أو في شكل ألكتروني ,هو مجموعة من المواد اللغوية التي وردت في بعض الطرق المبدئيه أما على شريط

Such collections are used in many different ways by different people

وتستخدم هذه المجموعات بطرق مختلفة من قبل أشخاص مختلفين

We are concerned mainly with use

ونحن اساساً معنيون بستخدامها ":

1) by linguists to help describe language, and test theories

من قبل اللغوين للمساعدة في وصف اللغه واختبار النظريات

2) by teachers and learners to aid language learning i.e. a form of CALL

من قبل المعلمين والمتعلمين للمساعدة في تعلم لغة أي شكل من أشكال كال

In principled way means

الطرق المبدئية تعني

Haphazardly العشوائيه

To perform any electronic corpus-based task directly you need

لأداء اي مهمة قائمة على الحضور الالكتروني المباشر تحتاج الى

A corpus -1

الأحضار

2-. A search engine

محركات البحث

A corpus itself is just text (a form of data), which may have been originally

والاحضار في حد ذاته هو مجرد نص شكل من اشكال البيانات والتي قد تكون في الاصل

Written

كتابية

Transcribed speech

كلام مبرمج أو مدون

Corpora are not all stored in the same format

لايتم تخزينالمجاميعفي نفس الشكل

may have coded information (tags) added in and out of the text,

قد تكون المعلومات مشفرة و العلامات مضافة داخل وخارج النص

to show e.g. who was speaking, the register of the text, or the part of speech

لأظهار من الذي كان يتحدث وسجل النص أو جزء من الخطاب

—

Corpora, when they are in the plainest of DOS or ASCII text,

المجاميع عندما يكونون في of DOS or ASCII text

may have coded information.

تكون لديك المعلومات مشفرة

Coded information are called Tags

ويطلق على المعلومات المشفرة العلامات

———

Tags are added in and out of the text, to show

العلامات تضاف للدخول والخروج من النص لـ أظهار

Who was speaking

الذي كان يتحدث

The register of the text

سجل النصوص

The part of speech of each word

الجانب الخطابي من كل كلمة

—

A search engine is program which generally runs through the text

محرك البحث هو البرنامج الذي يمتد عادة من خلال النص

(or a precompiled index to the text)

أو الفهرس المترجم مسبقاً للنص

———

The plural form of the word CORPUS

صيغة الجمع لكلمةCORPUS

is Corpora

هي المجاميع

———

Descriptive grammarians

القواعد الوصفية

use it to to improve their descriptions to fit the facts of actual use of constructions

تستخدم لتحسين الاوصاف لتتناسب مع وقائع الاستخدام الفعلي للمنشئات

Stylisticians الاحصائية

use it to to see what differences there are in how frequently different authors use certain words

يستخدم لمعرفة الاختلافات في عدد المرات التي استخدم فيها المؤلفين مفردات مختلفة

Sociolinguists

اللغويات الاجتماعية

frequent certain constructions are in conversation

تتكرر تراكيب معينة في المحادثات

Computational linguists

اللغويات الحاسوبية

grammatical parsing programs will work on naturally occurring language

قواعد تحليل البرامج تعمل على طبيعة اللغة

Language learning researchers

باحثو متعلمي اللغة

often learners with a particular L1 get something wrong

في كثير من الاحيان المتعلمون خاصة يحصلون على شي خاطىء

Writers of teaching syllabuses

كتب مناهج التدريس

often the passive really occurs in academic English

غالبا ماتكون سلبية وهذا شي حقيقي يحدث في اللغة الانجليزية الاكاديمية

Writers of teaching course materials

كتب تدريس المواد الدراسية

to incorporate authentic examples into

their material

أدراج الامثلة الحقيقة في المواد

المحاضرة التاسعة ..

الصراحة المحاضرة مادري وش تبي غريبة ومخربطة لكن استخدمت أسئلة ابو بكر جزاه الله خير وقدرت أطلع منها بعض المعلومات

Introspection

الاستبطان

means that you try to investigate different ideas while in corpus you Collect data and store them in one place

يعني محاولة تحقيق أفكار مختلفة بينما تقوم بـ احضار وجمع البيانات وتخزينها في مكان واحد

—

A corpus is a good representation of Daily life of people

الاحضار هو تمثيل جيد من الحياة اليومية للناس

——

One of the limitation of using carpus is that it …….

واحدة من حدود استخدام الاحضار هو أنه

Can't cover all what can occur

لايمكن أن يغطي كل مايحدث

—

If a population is vast, samples have to be vast to be representative

اذا كان عدد السكان هائل العينات يجب ان تكون هائلة لتكون ممثلة

This is a wrong belief

هذا أعتقاد خاطئ

___

To be **opportunistic** when you **design a corpus** means……

ان تكون انتهازياً عند تصميم الاحضار يعني أن

To benefit from available resources like media and internet

تستفيد من الموارد المتاحة مثل وسائل الاعلام والانترنت

_____

Sinclair سنكلير

who says Let the data speak for itself

هو الذي قال دع البيانات تتحدث عن نفسها

___

the sentences coming from your corpus called concordance

جملك الخاصة أحضارها يدعى بالتوافق

—

one of corpus linguistics use

واحدة من الاحضارات في استخدام اللغويات

is to do error analysis task

هو القيام بمهمة تحليل الاخطاء

—

.؟ How to relate go, goes and went

كيف الارتباط يذهب ذهب وتوجه ؟

preparing a corpus. It is called….

عند اعداد الاحضار ويسمى ذلك

Lemmatisation

المحاضرة 10

some insights obtainable from corpora

بعض الافكار يمكن الحصول عليها من المجاميع

but not maybe all obtainable

لكن ربما لايمكن للجميع الحصول عليها

Most of these have fairly obvious use for both descriptive linguists and teachers…
and maybe learners too and others in the range ofusers
معضمها له فائدة واضحة الى حد ما لكل من اللغوين وصفوف المعلمين وربما المتعلمين وبعضها في مجموعة المستخدمين

Frequencies of **individual words** across varieties:
ترددات الكلمات الفريدة خلال الاصناف
certain and sure
معينة ومؤكدة

Characteristics ofvarieties and individual authors:
خصائص الاصناف والمؤلفين
frequencies overall; TTRs
ترددات أجمالية

Details of meaningof vocabulary items and collocation:
تفاصيل معاني المفردات وبنود التجميع
qualitative detailsof synonyms
تفاصيل نوعيه من المفردات

It is **possible classify** most corpus projects, or generate new ones
من الممكن تصنيف معظم المشاريع وأحضارها أو أنتاج واحدة جديدة

from normal nativespeaker adults today
من الطبيعي اليوم للبالغين من المتحدثين في اللغة الام

it could be spoken orwritten, standard or non-standard,
أن يمكنه أن يتحدث بها أويكتبها قياسياً أو غير قياسي

from everyday language or the specialist register of newspapers or poetry or academic prose or…etc.
من اللغة اليومية أو السجلات المتخصصة في الصحف او الشعر او النثر الاكاديمي ,, الخ

it is possible sometimesto **merge** your own corpus with a readymade corpus
من الممكن في بعض الاحيان دمج الاحضار الخاص بك مع الاحضار الجاهز

—

cannot obtain fromcorpus…

لايمكن الحصول عليها من الاحضار

Mobile numbers to the American people

أرقام الهواتف النقاله للشعب الأمريكي

–

how people use thelanguage in their daily live

that's mean

كيفية استخدام اللغة في الحياة اليومية هذا يعني

pragmatics

البراغماتية

—

What kind of corpusinformation is needed..

أي نوع من المعلومات عند الاحضار نحتاجها

more **concordance**-typeinformation

مزيد من المعلومات المتوافقةمع هذا النوع

more frequency information **about words**

مزيد من المعلومات المتردده حول الكلمات

المحاضرة 11 ـ

British National Corpus BNC

is a 100 million word collection of samples of written

and spoken language from a wide range of sources,

هي عبارة عن مجموعة من 100 مليون كلمة من عينات مكتوبه ولغات منطوقه من مجموعة واسعه من المصادر

designed to represent a wide cross-section of British English from the later part of the 20th century,

both spoken and written.

مصممة لتمثيل قطاع عريض من الانجليزية البريطانية من الجزء الاخير من القرن العشرين تحدثا وكتابة

BNC XML Edition, released in 2007.

الطبعة الاخيرة منها صدرت عام 2007

—

The **written part** of the BNC (90%) includes,

الجزء الخطي منها ويشمل

extracts from regional and national newspapers, specialist periodicals and journals for all ages and interests, academic books and popular fiction, published and unpublished letters

مقتطفات من الصحف الاقليمية الوطنية والدوريات المتخصصة والمجلات لجيمع الاعمار والاهتمامات والكتب الاكاديمية والخيال الشعبي المنشور وغير المنشور

The **spoken part** 10%

الجزء المنطوق

consists of orthographic transcriptions of unscripted informal conversations

يتكون من التدوين الهجائي للمحادثات غير الرسمية المرتجلة

recorded by volunteers selected from different age,region and social classes in a demographically balanced way

المسجلة من قبل متطوعين تم أختيارهم من مختلف االعمار والمنطقة اوالطبقات الاجتماعية بطريقة متوازيه ديموغرافيا

---

corpora and search engines primarily constitute **tools or research methods**,

المجاميع ومحركات البحث تشكل أساسا أدوات واساليب البحث

rather than **areas of enquiry**

بدلامن مجالات التحقيق

The bulk of the project

الجزء الاكبر من المشروع

has to come from the user's prior knowledge of linguistics, teaching

يجب أن يأتي من المعرفة المسبقة لمستخدم علم اللغه والتعليم

---

if you want benefit from corpus

اذا كنت تريد الاستفادة من الاحضار

you should have a **background** of

linguistic

يجب أن يكون لديك خلفية لغوية

---

if you want to choose a corpus task for yourself or your students ....

أذأ كنت تريد أختيار مهمة أحضاريه لنفسك أو طلابك

think in something you are already strong in

فكر في شي قوي بالفعل

المحاضرة 12 ـ

Natural Language Processing NLP

معالج اللغات الطبيعية

Computers use analyze, understand, generate natural language

يستخدم أجهزة الكمبيوتر لتحليل وفهم وتوليد اللغة الطبيعية

Ambiguity of language

الالتباسات اللغوية

Phonetic لفظي

write, right, rite

Lexical معجمي

can = noun, verb, modal

Structural هيكلي

I saw the man with the telescope

رأيت رجلا ومعه تلسكوب

Semantic دلالات الالفاظ

dish = physical plate, menu item

أطباق = لوحات مادية , قائمة العناصر

All of these make NLP difficult

كل هذه تجعل من ان ال بي صعبة

Simply writing down linguistic insights

ببساطة تدوين الأفكار اللغوية

isn't sufficient to have a working system

ليس كافياً للحصول على نظام العمل

Programs need to run in real-time

البرامج التي تحتاج للشتغيل في الوقت الحقيقي

be efficient

تكون فعالة

There are thousands of grammar rules

هناك الآلاف من القواعد اللغوية

which might be applied to a sentence

التي يمكن تطبيقها على الجملة

—

Use insights from

أستخدام الأفكار

computer science

من علوم الكمبيوتر

———

To find the best parse,

للعثور على أفضل تحليل

use chart parsing, a form of dynamic programming

استخدم التخطيط التحليلي وهو شكل من أشكال البرمجة الديناميكية

——

NLP is هو

a somewhat more applied

أكثر تطبيقية الى حد ما

——

NLP has agoals

أهداف معالج اللغات الطبيعية

Scientific Goal

هدف علمي

Engineering Goal

هدف هندسي

Scientific goal of LNP means….

الاهداف العلمية لمعالج اللغات الطبيعية

Identify the computational machinery needed

تحدد الالية الحسابية اللازمه

forms of linguistic behavior -------لـ أشكال السلوك اللغوي