# Lingüística de corpus

Max Carey

Introducción a la lingüística aplicada

*"You shall know a word by the company it keeps"*
- J.R. Firth (1957:11)

23 de octubre de 2017

# Esbozo

# Esbozo

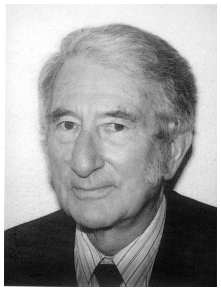
# J.R. Firth

J.R. Firth
Lingüísta inglés
1890-1960

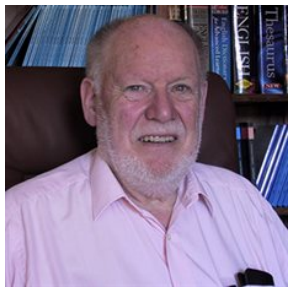
- Cabré and Lorente (2003): fundador del funcionalismo lingüístico
  - Incorporó las ideas del antropólogo Mainowski quien dijo: *"'el lenguaje no es un sistema en sí mismo (posición estructuralista extrema) sino que evoluciona por las demandas de la sociedad y su contexto"'* (11).
  - Se le atribuye ser el primero en estudiar colocaciones (Bartsch & Evert, 2014)

# Antecedentes históricas

J.R. Firth
Lingüísta inglés
1890-1960

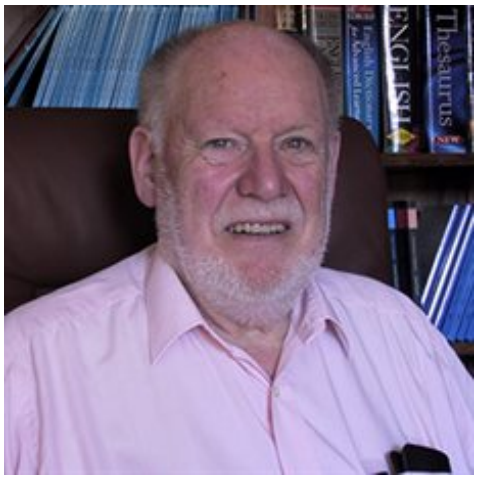


Michael Halliday
Fundador de la gramática sistémico funcional
1925-



John McHardy Sinclair
Fundador de la lingüística de corpus
1933-2007

# Sinclair




John McHardy Sinclair
Fundador de la lingüística de corpus
1933-2007

- Fundador del proyecto **COBUILD** (Collins and Birmingham University International Language Database)
  - Revolucionó la lexicógrafia (Fontenelle, 2011, p. 58)
- Famoso por dos principios de la lingüística de corpus
  - *the open choice principle*
  - *the idiom principle* Sinclair and Carter (2004)

# The Open Choice Principle

- [El libro de fonología variable de México]$_{\text{FN}}$ está sobre la cama.
  - que fue escrito por el lingüísta X
- [El libro de fonología introductoria]$_{\text{FN}}$ está sobre la cama.
  - que leímos para la clase de Fonética
- [El libro de fonología de las lenguas indígenas de México]$_{\text{FN}}$ está sobre la cama.
  - que se publicó en el Colegio de México

# Kahoot Question 1: The Idiom Principle

- Words tend to co-occur, so our lexicon is really composed of 'lexical bundles'.
  - Sinclair (39) calls these structures compound lexical items and he identifies the following component parts:
    - Collocation
    - Colligation
    - Semantic preference
    - Semantic Prosody
- Kahoot 1
- Let's run through Sinclair's analysis of the *true feelings* with the Corpus of Contemporary American English

# Esbozo

# True Feelings (Sinclair's Example p.35)

**Let's check out *True Feelings* with the Corpus of Contemporary American English**

- start with collocation: [true feelings]$_{\text{semantic core}}$
- List most frequent words that occur at n-1
  - [poss. adj.]$_{\text{colligation}}$ + [true feelings]$_{\text{s.c.}}$
- List common verb collocations in range n-4
  - [verb of expression]$_{\text{semantic preference}}$+ [poss. adj.]$_{\text{colli.}}$ + [true feelings]$_{\text{sc}}$
- Use KWIC to get semantic prosody
  - Semantic Prosody: Negative/Reluctance

**Final Lexical Bundle:**

[[[v. exp.]$_{\text{sem. pref.}}$+ [poss. adj.]$_{\text{colli.}}$ + [true feelings]$_{\text{sc}}$]$_{\text{sem pros: neg/rel.}}$]]

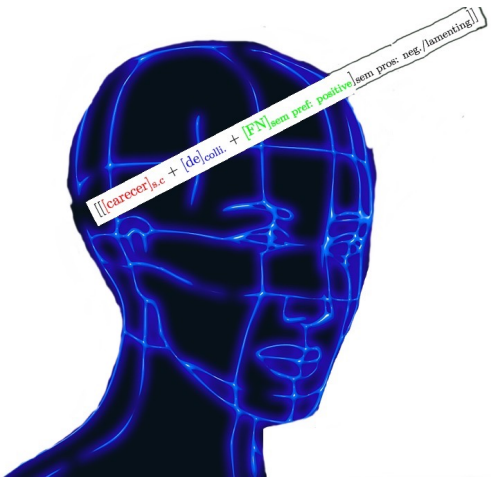
# Carecer - own analysis

**Vamos a analizar *carecer* con el Corpus de Español**

- Start with colligation/collocation: [carecer]$_{\text{semantic core}}$ + [de]$_{\text{colligation}}$
- List frequent noun collocations occurring in the range (n + 3)
  - [carecer]$_{\text{s.c}}$ + [de]$_{\text{colli.}}$ + [FN]$_{\text{semantic preference: something positive}}$
- What is the semantic prosody?
  - Kahoot Question 2

**Final Lexical Bundle:**

[[[carecer]$_{\text{s.c}}$ + [de]$_{\text{colli.}}$ + [FN]$_{\text{sem pref: positive}}$]$_{\text{sem pros: neg./lamenting}}$]]




- In academic blog post *Why Chomsky was wrong about corpus linguistics* Wallis (2016) explains theoretical clash with Sinclair.
  - Chomsky = I-language (linguistic competence) **not E-language**(linguistic performance)
  - Siclair: Build grammar from the **"bottom-up".**
- Today, few linguists rely on introspection and value-judgments; few applied linguistics lack theory in their analysis .

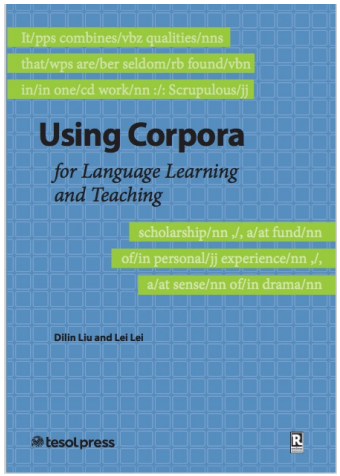
# Esbozo

# Types of Corpora

Table 1: Some Types of Corpora Listed By Adolphs (2011)

| **Type** | **Example** |
|---|---|
| Historical | Corpus diacrónico del español (CORDE) |
| Monitor | COCA, Corpus de Español |
| Learner | The International Corpus of Learner English |
| Parallel | Linguee.com |

# Language Teaching




Using Corpora for Language Learning and Teaching

(Liu & Lei, 2017)

- **Inductive** vs. **Deductive** Corpora Use in the Classroom
  - Inductive: Students "discover" the language patterns for themselves, engage in deep-learning
  - Deductive: Confirming or reinforcing what you already know (Come up with a small worksheet that students can fill out, completing some kind of information that they have)

# Language Teaching - Example 1

**Table 3.1 Common Collocations of *Make, Take, Do,* and *Have***

| Make | Take | Do | Have |
|---|---|---|---|
| *Make a case* | *Take a bath* | *Do business* | *Have an affair* |
| *Make a change* | *Take a break* | *Do chores* | *Have an accident* |
| *Make a choice* | *Take a bus/taxi* | *Do dishes* | *Have an argument* |
| *Make a commitment* | *Take a chance* | *Do drugs* | *Have a conversation* |
| *Make a contribution* | *Take a look* | *Do errands* | *Have difficulty* |
| *Make a decision* | *Take a nap* | *Do exercises* | *Have a dream* |
| *Make a difference* | *Take an offer* | *Do harm* | *Have experience* |
| *Make an effort* | *Take a rest* | *Do homework* | *Have a feeling* |
| *Make a living* | *Take a phone call* | *Do laundry* | *Have fun/a good time* |
| *Make a mistake* | *Take a shower* | *Do research* | *Have a look* |
| *Make a phone call* | *Take a test* | *Do things* | *Have a problem* |
| *Make progress* | *Take a walk* | *Do work* | *Have trouble* |

Table reproduced from (Liu & Lei, 2017, p.39)

- Is it easier to *make* or *take* something?
- Which verb is used with routines?
- What about have? Can you replace these verbs with another verb?

# Language Teaching - Example 2

- According to (Cowan, 2008) some phrasal verbs are separable: *set up, hand in, look up*
  - I set up [the machine]$_{\text{NP}}$
  - I set [the machine]$_{\text{NP}}$ up
- What can the following data tell us about separable phrasal verbs?

Table 2: set up + NP

| Rank | Example | Hits |
|---|---|---|
| 1 | SET UP THE TENT | 24 |
| 2 | SET UP THE SYSTEM | 20 |
| 3 | SET UP THE EQUIPMENT | 19 |
| 4 | SET UP THE MEETING | 18 |
| 5 | SET UP THE CAMERA | 17 |
| 6 | SET UP THE TELESCOPE | 10 |
| 7 | SET UP THE CONDITIONS | 9 |
| 8 | SET UP THE COURSE | 9 |
| 9 | SET UP THE GAME | 9 |

Table 3: set (NP) up

| Rank | Example | Hits |
|---|---|---|
| 1 | SET IT UP | 618 |
| 2 | SET ME UP | 241 |
| 3 | SET HIM UP | 235 |
| 4 | SET THEM UP | 194 |
| 5 | SET YOU UP | 185 |
| 6 | SET HER UP | 114 |
| 7 | SET HIMSELF UP | 110 |
| 8 | SET US UP | 87 |
| 9 | SET YOURSELF UP | 84 |

# Example 2 - All Data

## Table 4: set up + NP

| Rank | Example | Hits |
|---|---|---|
| 1 | SET UP THE TENT | 24 |
| 2 | SET UP THE SYSTEM | 20 |
| 3 | SET UP THE EQUIPMENT | 19 |
| 4 | SET UP THE MEETING | 18 |
| 5 | SET UP THE CAMERA | 17 |
| 6 | SET UP THE TELESCOPE | 10 |
| 7 | SET UP THE CONDITIONS | 9 |
| 8 | SET UP THE COURSE | 9 |
| 9 | SET UP THE GAME | 9 |

## Table 5: set + NP + up

| Rank | Example | Hits |
|---|---|---|
| 1 | SET THE CAMERA UP | 7 |
| 2 | SET THE LEVEL UP | 3 |
| 3 | SET THE COURSE UP | 3 |
| 4 | SET THE BATHROOM UP | 2 |
| 5 | SET THE BOAT UP | 2 |
| 6 | SET THE CONTRAPTION UP | 2 |
| 7 | SET THE COMPUTER UP | 2 |
| 8 | SET THE COMPANY UP | 2 |
| 9 | SET THE DOCTOR UP | 2 |

## Table 6: set + prn. + up

| Rank | Example | Hits |
|---|---|---|
| 1 | SET IT UP | 618 |
| 2 | SET ME UP | 241 |
| 3 | SET HIM UP | 235 |
| 4 | SET THEM UP | 194 |
| 5 | SET YOU UP | 185 |
| 6 | SET HER UP | 114 |
| 7 | SET HIMSELF UP | 110 |
| 8 | SET US UP | 87 |
| 9 | SET YOURSELF UP | 84 |

## Table 7: set + up + prn.

| Rank | Example | Hits |
|---|---|---|
| 1 | SET UP SOMETHING | 21 |
| 2 | SET UP EVERYTHING | 14 |
| 3 | SET UP YOU | * |
| 4 | SET UP HIMSELF | * |
| 5 | SET UP ONE | * |
| 6 | SET UP IT | * |
| 7 | SET UP I | * |
| 8 | SET UP WE | * |
| 9 | SET UP ANYTHING | * |

# Esbozo

# Building a local corpus

- You want your corpus to be representative
- Corpus can be scientific or practical
- Differences between Online and Offline Corpus

Table 8: Pros and Cons of Using a Local Corpus

| **Pros** | **Cons** |
|---|---|
| Flexibility, complex queries | Requires technical skill to preprocess data: |
| Control over data | Website owns data |
| Free | Pay for Premium Features |

# Software from Laurence Anthony

AntFileConverter

AntWordProfiler

AntConc

- Let's create our own corpus from PDF Files
- And see a practical application of AntConc

# Esbozo

# Bibliography I

Adolphs, S. (2011). Corpus Linguistics. In J. Simpson (Ed.), *The Routledge handbook of applied linguistics* (1st ed ed.). Milton Park, Abingdon, [UK] ; New York: Routledge.

Bartsch, S., & Evert, S. (2014). Towards a Firthian notion of collocation. *Network Strategies, Access Structures and Automatic Extraction of Lexicographical Information. 2nd Work Report of the Academic Network Internet Lexicography, OPAL–Online publizierte Arbeiten zur Linguistik. Institut für Deutsche Sprache, Mannheim, to appear.*

Cabré, M. T., & Lorente, M. (2003). *Panorama de los paradigmas en lingüística* (Vols. Ciencias exactas, naturlaes y sociales). A. ESTANY. Madrid: Consejo Superior de Investigaciones Científicas, 2004.

# Bibliography II


Cowan, R. (2008). *The teacher's grammar of English: a course book and reference guide*. Cambridge ; New York: Cambridge University Press. (OCLC: ocn212410033)

Fontenelle, T. (2011). Lexicography. In J. Simpson (Ed.), *The Routledge handbook of applied linguistics* (1st ed ed., pp. 597–610). Milton Park, Abingdon, [UK] ; New York: Routledge.

Liu, D., & Lei, L. (2017). *Using Corpora for Language Learning and Teaching*. TESOL International Association.

Sinclair, J., & Carter, R. (2004). *Trust the Text Language, Corpus and Discourse*. (OCLC: 999116876)

Wallis, S. (2016, November). *Why Chomsky was Wrong About Corpus Linguistics*. Retrieved 2017-09-21TZ, from `https://corplingstats.wordpress.com/2016/11/02/why-chomsky-was-wrong/`