Multilingual Corpora and Concordancing

Philip King
University of Aston

August 2011

Overview

- Multilingual corpora
- Data & data retrieval
  - Programs
- Uses
- Hands-on with Multiconcord

Multilingual corpora

- In more than one language
  - Often only 2, hence, bilingual, but can be more
  - Types of multilingual corpus:
    - Texts and their translations into one or more languages *(parallel)*
    - Comparable texts in ≥ 2 languages *(not parallel)*
      - And variations or combinations

Preparing your multilingual parallel corpus

- Problem for the software: how to find the corresponding bit of the target language.

- ***Alignment*** : preparing your texts (by marking sentence or paragraph beginnings in each pair of texts, and then checking that they match each other).

Multilingual corpora: three parallel texts

IN MY YOUNGER and more vulnerable years my father gave me some advice that I've been turning over in my mind ever since. 'Whenever you feel like criticising anyone,' he told me, 'just remember that all the people in this world haven't had the advantages that you've had.' He didn't say any more, but we've always …	Cuando era más joven y más vulnerable, mi padre me dio un consejo al que no he dejado de dar vueltas desde entonces. "Siempre que sientas deseos de criticar a alguien", me dijo, "recuerda que no todo el mundo ha disfrutado de las facilidades que tú has tenido." Eso fue lo único que dijo, pero como siempre nos lo hemos ...	Κάποτε, την εποχή που ήμουνα αρκετά πιο νέος και πολύ πιο ευαίσθητος, ο πατέρας μου μού έδωσε μια συμβουλή που δεν έπαψε από τότε να συνοδεύει τη σκέψη μου. "'Οταν ετοιμάζεσαι να κατακρίνεις κάποιον", μου είπε, "θυμήσου πρώτα ότι δεν είχαν όλοι οι άνθρωποι τις δικές σου ευκαιρίες στη ζωή". Αυτό ήταν όλο που μου …

Multilingual corpora: three parallel texts

Au coeur de l'Europe De par sa position géographique, à une trentaine de kilomètres à l'Est de Paris, sur le territoire de Marne-la-Vallée, l'une des cinq "villes nouvelles" de la région parisienne, Euro Disney Resort occupe une place privilégiée sur le continent européen (le mot resort sous-entend lieu de séjour, destination de vacances). La carte ci-dessous indique les grandes routes	**В СЕРДЦЕ ЕВРОПЫ** Благодаря своему географическому положению (он расположен в тридцати километрах от Парижа на территории Марн-ля-Валле, одного из пяти "новых городов" парижского района), парк аттракционов Евродисней занимает особое место на европейском континенте. На данной карте указаны дороги, ведущие к парку	**At the Crossroads of Europe** Thanks to its location some 30km - 20 miles east of Paris within the administrative limits of Marne-la Vallée, one of the five new towns in the Paris region, the Euro Disney Resort occupies a special place at the heart of the European continent. The map below shows the main routes for travelling from the major European cities.

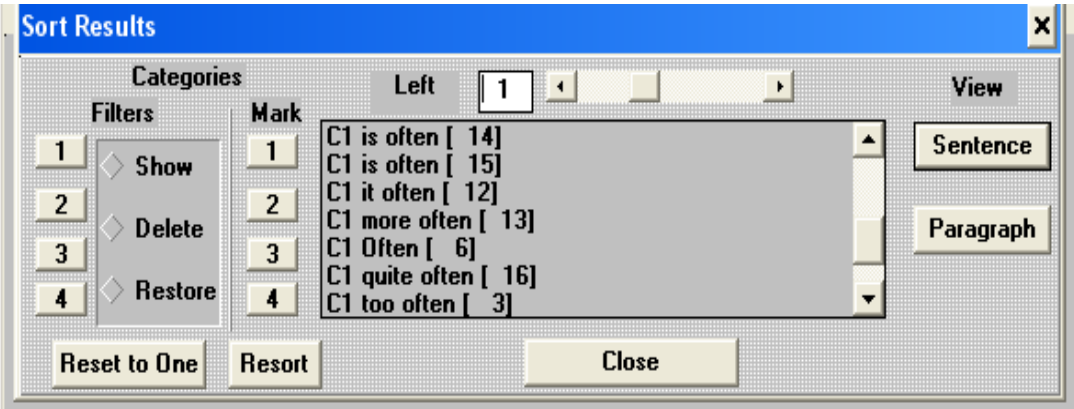
Sample of marked-up text

- <p><s>Giansily (UPE). <s>- _ (FR) _ Mijnheer de Voorzitter, zo is de stemming niet verlopen. <s>Bij de opening van de vergadering heeft de Voorzitter gezegd dat hij niet wist of dit verslag aan het eind van de ochtend in stemming zou worden gebracht of niet. <s>Aangezien de beslissing hiertoe pas op het moment van de stemming werd aangekondigd, konden wij ons verzoek niet van tevoren indienen.
- <p><s>Green (PSE). <s>- _ (EN) _ Mijnheer de Voorzitter, ik verwijs naar de Notulen en de stemming over het actualiteitendebat (bezwaren) van gisteren. <s>Mijn excuses voor de late kennisgeving, maar ik zou uw diensten willen verzoeken er nota van te nemen dat mijn fractie haar naam van de gezamenlijke ontwerp-resolutie over de mensenrechten in Colombia geschrapt wenst te zien.
- <p><s>Klaß (PPE). <s>- _ (DE) _ Mijnheer de Voorzitter, ik heb me gisteren kennelijk vergist bij de eindstemming over het verslag-Dury/Maij-Weggen. <s>Ik wilde vóór het verslag stemmen, maar ik heb een fout gemaakt en ik wilde dat nu in de notulen laten opnemen.
- <p><s>Gebhardt (PSE). <s>- _ (DE) _ Mijnheer de Voorzitter, ik heb bij de hoofdelijke stemming gisteren ook een fout gemaakt, die ik graag wilde corrigeren. <s>Bij punt 6, sub 1, van de notulen staat dat ik tegengestemd zou hebben, ik wilde echter vóór stemmen.
- <p><s>Wijsenbeek (ELDR). <s>- _ (EN) _ Mijnheer de Voorzitter, over de Notulen. <s>Gisteren is mijn verslag voor de tweede lezing vanwege een uiterst slechte planning door het Voorzitterschap niet behandeld...

Retrieving data

- Useful starting points
  - Frequency lists
  - Alphabetical lists
- Display
  - On screen manipulation (sorting, editing)

Sample results



ep960417.en P67 S3

I thank him on behalf of the Confederal Group of the European United Left, but also on behalf of my party in Italy, the _ Movimento dei Comunisti Unitari _ , which has had positive relations with his government **and often** been criticized for it.

ep960417.it P67

Lo ringrazio a nome del gruppo della sinistra europea, ma anche a nome del mio gruppo in Italia, quel movimento dei comunisti unitari che ha avuto con il suo governo un rapporto positivo ma spesso critico.

OUTPUT EDITED INTO A TABLE IN *WORD*

Unter _neuartig_ versteht man im Lebensmittelbereich die ursprüngliche Zusammensetzung, wenn sie insbesondere mit genetisch veränderten Organismen, die ich _GVO_ nenne, hergestellt wurde.	The word "novel" in connection with foodstuffs is taken to mean original foods and ingredients derived, in particular, from genetically modified organisms, or GMOs.
Es ist nämlich bedeutsam festzustellen, daß die häufigsten Zwischenfälle im wesentlichen die Überweisungen kleinerer Beträge, meist **unter** 10 000 ECU, betreffen.	It is significant, in fact, to note that the most frequent incidents essentially relate to the smallest transfers, mostly under ECU 10 000.
Ich denke folglich, daß der Wunsch, Wirtschaftsentwicklungen durch ständige Wechselkursanpassungen auszugleichen, schlecht für die betreffende Wirtschaft und **unter** allen Umständen gefährlich für den Fortbestand, das Überleben und die Vertiefung des gemeinsamen Marktes ist	Consequently, I believe that any attempt to control economic developments through exchange-rate measures is bad for the economy in question and, in any event, jeopardizes the continuation and deepening of the single market.
Im übrigen gibt es natürlich auch viele Menschen, die **unter** Krankheiten, **unter** Allergien leiden.	For the rest, there are of course many people who suffer from illnesses and allergies.

USES: WHAT CAN WE LEARN

- Who?
  - Language teachers
  - Language learners
  - Translation trainees
- What?
  - “gaps” in a language
  - Dictionary and beyond
    - Real correspondences
    - collocations
    - Complexity of equivalence