**Collocazioni nell'opera giovanile di Federico García Lorca**

### 1.1 Cosa faremo

Utilizzeremo UCS per analizzare determinati aspetti linguistici e letterari di due diversi corpora. Ordineremo i bigrammi dei due corpora a partire da diverse misure di associazione (AM), al fine di ottenere tabelle in grado di mettere in risalto i bigrammi più significativi. In questo modo, scopriremo anche le proprietà delle diverse AM, i vantaggi e gli svantaggi che esse offrono, gli "eventi" che esse mettono in risalto e quelli che, invece, nascondono.

### 1.2 La nostra ipotesi e i corpora

L'aspetto che prenderemo in considerazione riguarda la differenziazione dei generi nell'opera giovanile di Federico García Lorca.

L'ipotesi è che (banalmente) prosa e poesia usino e combinino in maniera diversa le parole. Più nel dettaglio ci aspettiamo la poesia presenti una maggiore originalità e creatività a livello lessicale.

Per verificare questa ipotesi lavoriamo su due corpora:

(1) corpus contenente tutta la poesia (pubblicata e non) catalogata dai critici come **poesia** giovanile e

(2) corpus contenente tutti i testi in **prosa** (pubblicati e non) scritti tra il 1910 e il 1922 (inclusa la corrispondenza).

Lavoreremo a partire dai files

**bigrammi.poesia.lc** e **bigrammi.prosa.lc**

contenenti tutti i bigrammi presenti nei due corpora (tutti in minuscola), che troverete in **cl_shared_data/lorca** (copiateli nella vostra home nella cartella in cui desiderate lavorare).

### 1.3 UCS

Se, nel percorso che segue, avete problemi con i comandi e la loro formulazione, potete fare riferimento al manuale ucsdoc, accedendovi nel seguente modo:

ucsdoc nome-del-comando (es. ucsdoc ucs-make-tables)

Per un'introduzione a UCS, potete leggere:

ucsdoc ucsintro

Per conoscere tutti i programmi, il loro uso e esempi di applicazione, utilizzate il tutorial getting-started.txt, che trovate nella cartella cl_shared_data/lorca.

## 2. Preparazione dei files

Illustro di seguito il procedimento attraverso il quale otterremo i file utili all'analisi dei due corpora. Descrivo nel dettaglio il procedimento da seguire per il corpus della poesia, mentre darò solo

qualche veloce indicazione per la preparazione del file della prosa, visto che il procedimento è analogo.

### 2.1 Poesia

#### 2.1.1. Compilazione tabelle

Con **ucs-make-tables** creiamo una tabella contenente le "frequency signatures" dei bigrammi:

```
ucs-make-tables <bigrammi.poesia.lc -v poesia.ds
```

Con **ucs-print** visualizziamo il file:

```
ucs-print -i poesia.ds
```

Se tutto è andato bene, il file dovrebbe avere questo aspetto:

```
id       l1  l2          f   f1    f2     N
-- ---------- ----------- -- --- ---- -----
 1   voluntad  morena       1    1    7  56193
 2     eterna  es          1   35   427  56193
 3         al  beber       1  306     7  56193
 4      mucha  pena        1    5    51  56193
 5    mara<F1>as  de        1    2  3508  56193
 6        por  luisa       1  445     1  56193
```

A questo punto aggiungiamo alla tabella i valori relativi a diverse AM. (Per una lista completa delle AM calcolabili con UCS vedi ucsdoc UCS::AM)

```
ucs-add -v am.MI am.log.likelihood TO poesia.ds INTO poesia.data
```

#### 2.1.2 Sorting delle tabelle

Ordiniamo i bigrammi secondo le diverse misure utilizzando **ucs-sort** e vediamo cosa succede:

- LA FREQUENZA:

```
ucs-sort -v poesia.data BY f- |ucs-print -i
```

Il top della lista contiene per la maggior parte bigrammi composti da due parole funzionali ("de la", "en el", etc.).

Visto che non siamo interessati all'uso delle parole funzionali, ci conviene eliminare dalla nostra lista i bigrammi contenenti due parole con frequenza alta (giacché si tratterà probabilmente di parole funzionali). Lo facciamo con ucs-select (ucsdoc ucs-select):

```
ucs-select –v '%' FROM poesia.data WHERE '%f1%<=300 || %f2%<=300'|ucs-sort BY f-|ucs-print -i
```

In un corpus così limitato 300 risulta essere già una frequenza alta. Per stabilire questo valore abbiamo utilizzato il file contenente il corpus tokenizzato ordinato per frequenza discendente, verificando la frequenza della prima parola "lessicale" presente nella lista (e il caso vuole che la parola "lessicale" più frequente nella poesia del giovane Lorca sia "corazón", che compare nel corpus 283 volte). Ho considerato questa come spartiacque verosimile tra "frequenze alte" e "frequenze basse".

In questo modo ho mantenuto nel mio file tutti i bigrammi che contengono almeno una parola con frequenza inferiore a 300. Da un primo sguardo alla tabella è evidente che nella parte alta della lista compaiono bigrammi formati dalle parole "lessicali" più frequenti precedute da una parola funzionale.

Un output di questo genere non ci fornisce informazioni particolarmente interessanti (a meno che non fossimo interessati a verificare empiricamente quali articoli siano appropriati/piu' utilizzati con quali sostantivi). Ci conviene, pertanto, eliminare dal nostro file tutti i bigrammi che contengano anche solo una parola a frequenza alta. Di nuovo, uso ucs-select:

```
ucs-select -v '%' FROM poesia.data WHERE '%f1%<=300 && %f2%<=300'|ucs-sort BY f-|ucs-print -i
```

I bigrammi così ottenuti sembrano molto più interessanti, quindi li salviamo in un file che useremo come punto di partenza per le prossime "operazioni":

```
ucs-select -v '%' FROM poesia.data WHERE '%f1%<=300 && %f2%<=300' INTO poesia.data.filtered
```

- MUTUAL INFORMATION:

```
ucs-sort -v poesia.data.filtered BY am.MI |ucs-print -i
```

Cosa noto?

I primi 320 bigrammi sono bigrammi insoliti in cui f=f1=f2=1, ovvero parole che compaiono una volta sola nel corpus e capitano insieme:
e.g. "linfas alumna" (due sostanitivi indipendenti)

La MI di "linfas alumna" risulta molto maggiore della MI di un bigramma come "san francisco" (f=16 f1=31 f2=49) in cui le due parole hanno, evidentemente, una dipendenza mutua molto più forte. Verifichiamo, così, la nota tendenza della MI a sopravvalutare i bigrammi con frequenze molto basse.

Per ovviare a questo problema, eliminiamo i bigrammi in cui entrambe le parole hanno f=1:

```
ucs-select -v '%' FROM poesia.data.filtered WHERE "%f%>1" INTO poesia.data.max
```

- LOG-LIKELIHOOD:

Ordinate poesia.data.max in base a questa misura utilizzando ucs-sort.

```
ucs-sort poesia.data.max BY am.log.likelihood | ucs-print -i
```

Aprite due nuove finestre dove ordinerete lo stesso file per frequenza (f) discendente e mutual information. Comparate l'output: cosa notate?

- VERSIONE FINALE:

Poiché i risultati ordinati per log-likelihood seguono un andamento molto simile a quelli ordinati per frequenza discendente, per la nostra analisi utilizzeremo le tabelle contenenti i bigrammi filtrati per 1<f<300 (frequenza maggiore di 1 e minore di 300) [poesia.data.max], che ordino per MI e salvo in un file .txt:

```
ucs-sort poesia.data.max BY am.MI INTO poesia.txt
```

E sarà questo il file di partenza per le nostre analisi.

**2.2 La prosa**

Ripetiamo il procedimento illustrato sopra, applicandolo questa volta al file della prosa.

```
ucs-make-tables <bigrammi.prosa.lc -v prosa.ds

ucs-add -v am.MI am.log.likelihood TO prosa.ds INTO prosa.data
```

- FREQUENZA:

```
ucs-sort -v prosa.data BY f- |ucs-print -i
```

Il top della lista contiene per la maggior parte bigrammi composti da due parole funzionali ("de la", "en el", etc.).
Elimino le parole a frequenza alta. Mi servo del corpus tokenizzato e ordinato per frequenza discendente per scoprire che la parola "lessicale" a frequenza più alta è ancora "corazón" con fq=529, per cui stabilisco il limite tra fq alte e fq basse a 550.

```
ucs-select -v '%' FROM prosa.data WHERE '%f1%<=550 && %f2%<=550'
INTO prosa.data.filtered
```

- MUTUAL INFORMATION:

```
ucs-sort -v prosa.data.filtered BY am.MI |ucs-print -i
```

Anche in questo caso in cima alla lista troviamo una lunga serie di bigrammi per i quali f=f1=f2=1 che vengono notevolmente sopravvalutati rispetto a bigrammi decisamente più significativi: ad es. "grasientas tragando" (f=f1=f2=1) ha una MI più alta di "queridos padres" (f=30 f1=32 f2=63)

Elimino bigrammi la cui frequenza congiunta sia pari a 1 (f=1):

```
ucs-select -v '%' FROM prosa.data.filtered WHERE "%f%>1" INTO prosa.data.max
```

Ordino tutto per MI e salvo in file .txt:

```
ucs-sort -v prosa.data.max BY am.MI INTO prosa.txt
```

**3. Analisi**

A questo punto abbiamo finalmente i file che contengono informazioni utili per la nostra analisi.

Che cosa si nota comparando **poesia.txt** e **prosa.txt** rispetto a:

1) incidenza nomi propri e latinismi: sono più frequenti nella poesia o nella prosa?
2) coppie sostantivo-aggettivo/aggettivo-sostantivo: sono perifrasi fisse o originali? in quale corpus verifichiamo una maggiore incidenza di coppie che possiamo considerare collocazioni? In quale corpus gli aggettivi sono più spesso posposti al sostantivo?
3) molti bigrammi con MI alta contengono parole semanticamente correlate alla religione ("agnus dei" "vía crucis" "lux perpetua", "gratia plena", etc.) e alla musica ("adagio cantabile", "molto allegro", "allegro ma", "ma non", "chopinescas romanzas", etc.). Perché?

**3.1.Cosa ci dicono questi dati in merito all'uso delle parole nella prosa e nella poesia?**

In **PROSA:** alta incidenza di nomi propri (jean aubry, miguel pizarro, etc.) e di coppie nome+aggettivo codificate (guardia civil, conferencia telefónica, macho cabrío, letra mayúscula), coppie nome-aggettivo più creative si trovano più in basso nella lista (serpientes bailarinas, estatuas mórbidas, alfombras húmedas), alto numero latinismi

In **POESIA:** alta incidenza di perifrasi verbali e di coppie aggettivo-sostantivo o sostantivo-aggettivo “creative” già dall'inizio (barro mohoso, país marfileño).

Questo ci suggerisce che la prosa è più informativa della poesia e meno variegata a livello lessicale; nella POESIA la lingua è utilizzata in maniera più creativa e originale, grazie all’uso di combinazioni lessicali inaspettate e inusuali.

**3.2 Cosa ci dicono questi dati in merito alla selezione dei bigrammi che abbiamo operato?**

Soprattutto per quanto riguarda la prosa, i bigrammi con MI alta costituiscono di solito espressioni fisse la cui f, f1 e f2 sono molto basse e tendono a essere molto vicine in valore; ovvero, sono costituiti da parole molto rare che quando compaiono nel corpus, lo fanno in corrispondenza l'una dell'altra. Così, la parola *agnus* compare solo due volte nel corpus ed entrambe le volte precede la parola *dei*. Sebbene ciò indichi un'indubbia dipendenza mutua tra le due parole (e quindi ci dia informazioni relative alla lingua in generale), ci dice poco sulla lingua usata da Lorca.

Per ottenere risultati più interessanti in riferimento alla lingua di Lorca è meglio continuare a sperimentare con UCS, filtrando le tabelle in modi diversi. Ad esempio, cosa succede se escludo tutti i bigrammi in cui f=f1=f2 e f>8?

```
ucs-select -v '%' FROM poesia.txt WHERE '%f%!=%f1% && %f1%!=%f2%
&& %f%!=%f2% %f%>8'|ucs-print -i
```