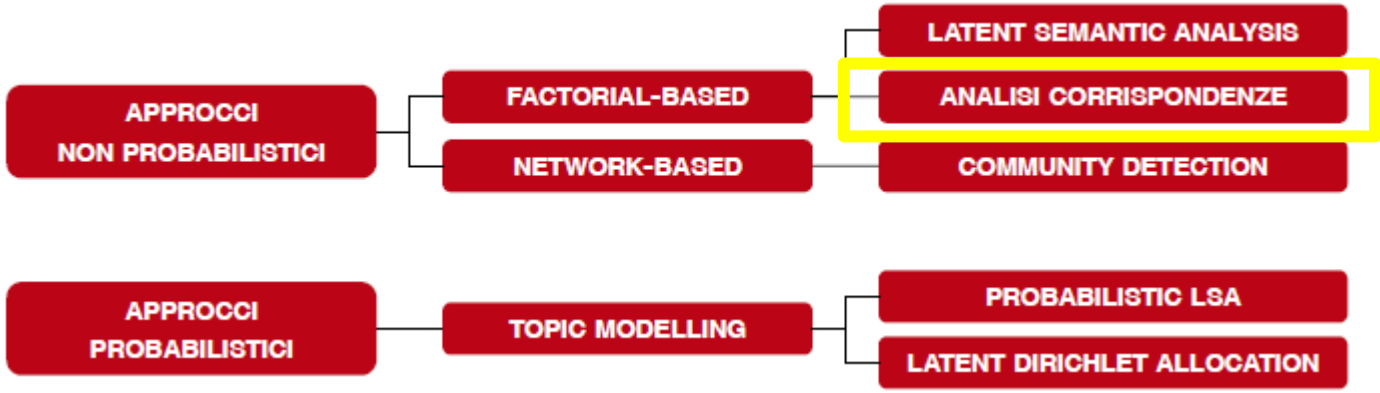
L' ANALISI DELLE CORRISPONDENZE LESSICALI (ACL)

L'ANALISI DELLE CORRISPONDENZE LESSICALI (ACL)

All'interno di un testo, attraverso i diversi termini utilizzati dall'autore, sono espressi una serie di concetti.

> ***Concepts are the glue that holds our mental world together*** (Murphy, 2002)

Uno degli obiettivi principali dell'analisi automatica dei testi è quello di estrarre informazioni interessanti a partire da una collezione di testi, evidenziando quali sono i principali concetti/tematiche



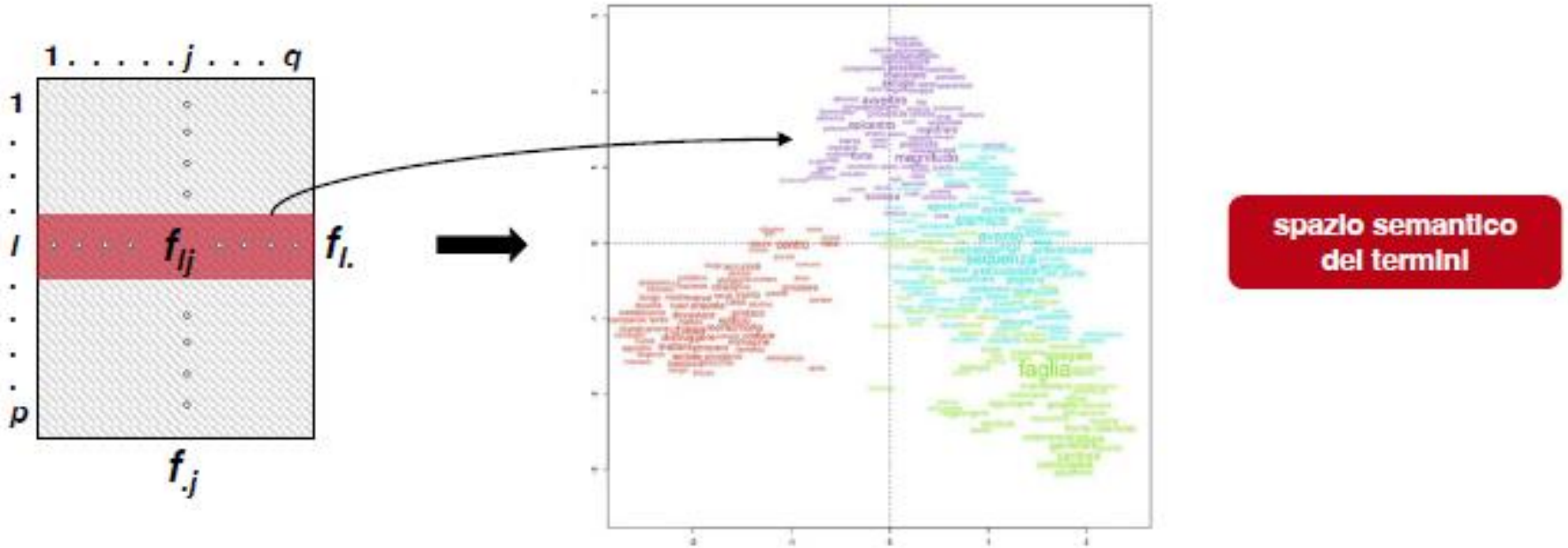
L'ANALISI DELLE CORRISPONDENZE LESSICALI (ACL)

Prof.ssa Rosanna Cataldo

L'analisi fattoriale ha lo scopo di descrivere la variabilità di un fenomeno complesso, a partire da un insieme ampio di caratteristiche, attraverso un numero ridotto di costrutti (**LATENTI**)

L'analisi delle corrispondenze è una tecnica per decomporre una matrice di dati in una serie di fattori, ciascuno dei quali rappresenta un aspetto latente dell'associazione presente nei dati osservati

Applicata all'analisi di dati testuali, **l'analisi delle corrispondenze lessicali** consente di ricercare e visualizzare strutture linguistiche latenti che esprimono concetti o temi prevalenti, e allo stesso tempo consente di ricercare e visualizzare similarità tra documenti in termini di vocabolario condiviso



OBIETTIVO DEL METODO

Obiettivo è **trovare il miglior sottospazio di rappresentazione** della nube dei punti **conservando quanta più informazione possibile**

Occorre **individuare i principali contenuti semantici del testo** e rappresentarli come assi cartesiani in spazi a più dimensioni, preferibilmente due o tre in modo da poterli visualizzare.

Punto di partenza è il Corpus composto da n documenti relativi ad uno specifico ambito rispetto al quale si ha un dato bisogno informativo.

Dopo aver effettuato tutte le procedure di pre-trattamento e analizzato il vocabolario delle forme che compongono i testi, si selezionano le *p* forme testuali a maggior contenuto informativo.

Si può costruire in questo modo una tabella con le **in colonna le *p* forme lessicali** ovvero le forme testuali del Corpus che sono state selezionate per l'analisi e **in riga i testi che costituiscono il Corpus** ogni testo su una riga

In ogni cella di questa matrice si avrà la **frequenza** con la quale una determinata parola è presente in un determinato testo.

DUE MODI DI LEGGERE I RISULTATI DELL'ACL

- **Interpretazione semantica**, basata sulla considerazione dei contributi, delle coordinate fattoriali e del valore test;

- **Interpretazione grafica**, basata sulla lettura del grafico fattoriale proiezione sugli assi delle parole e/o categorie e delle variabili/modalità per mezzo delle quali sono stati ripartiti i testi

L'INTERPRETAZIONE SEMANTICA DEI RISULTATI DELL'ACL

L'interpretazione semantica dei fattori estratti si basa sui seguenti parametri:

- il **contributo assoluto** di ciascuna variabile attiva, che indica la quota di inerzia totale del fattore spiegata dalla variabile stessa, in altre parole, questo parametro rappresenta quanta parte ha avuto tale variabile nella determinazione del fattore, in rapporto all'insieme delle variabili;

- il **coseno quadrato**, che indica il contributo del fattore alla spiegazione della variabilità di una determinata variabile;

- Il **valore test (V.T.)**, che è un test di significatività statistica che indica se la relazione delle modalità delle variabili con i fattori è statisticamente significativa per P-value=0,05, quando il V.T. ≥ 2

LA LETTURA DEL PIANO FATTORIALE

L' ACL permette di rappresentare graficamente le associazioni tra le righe e le colonne della tabella su un piano delimitato da due assi fattoriali.

Su tale piano ogni forma testuale e ogni documento sono rappresentati da un punto, e questo permette di considerare la prossimità tra due punti-riga o due punti-colonna e anche la prossimità incrociata tra un punto-riga e un punto-colonna.

Nel primo caso se due **documenti** (rappresentati dai due punti-riga) sono vicini se ne può concludere che essi condividono prevalentemente lo stesso linguaggio e quindi **hanno un vocabolario simile** (ovvero sono composti e caratterizzati dalle medesime forme testuali).

Nel secondo caso se due **forme testuali** (rappresentate da due punti-colonna) sono vicine se ne può concludere che esse **vengono** prevalentemente **utilizzate spesso in modo congiunto** e dunque anche tendenzialmente negli stessi documenti.

Infine nel terzo caso se un punto-riga e un punto-colonna sono prossimi se ne può concludere che la forma testuale corrispondente al punto-colonna «tende» a caratterizzare il testo rappresentato dal punto-riga.

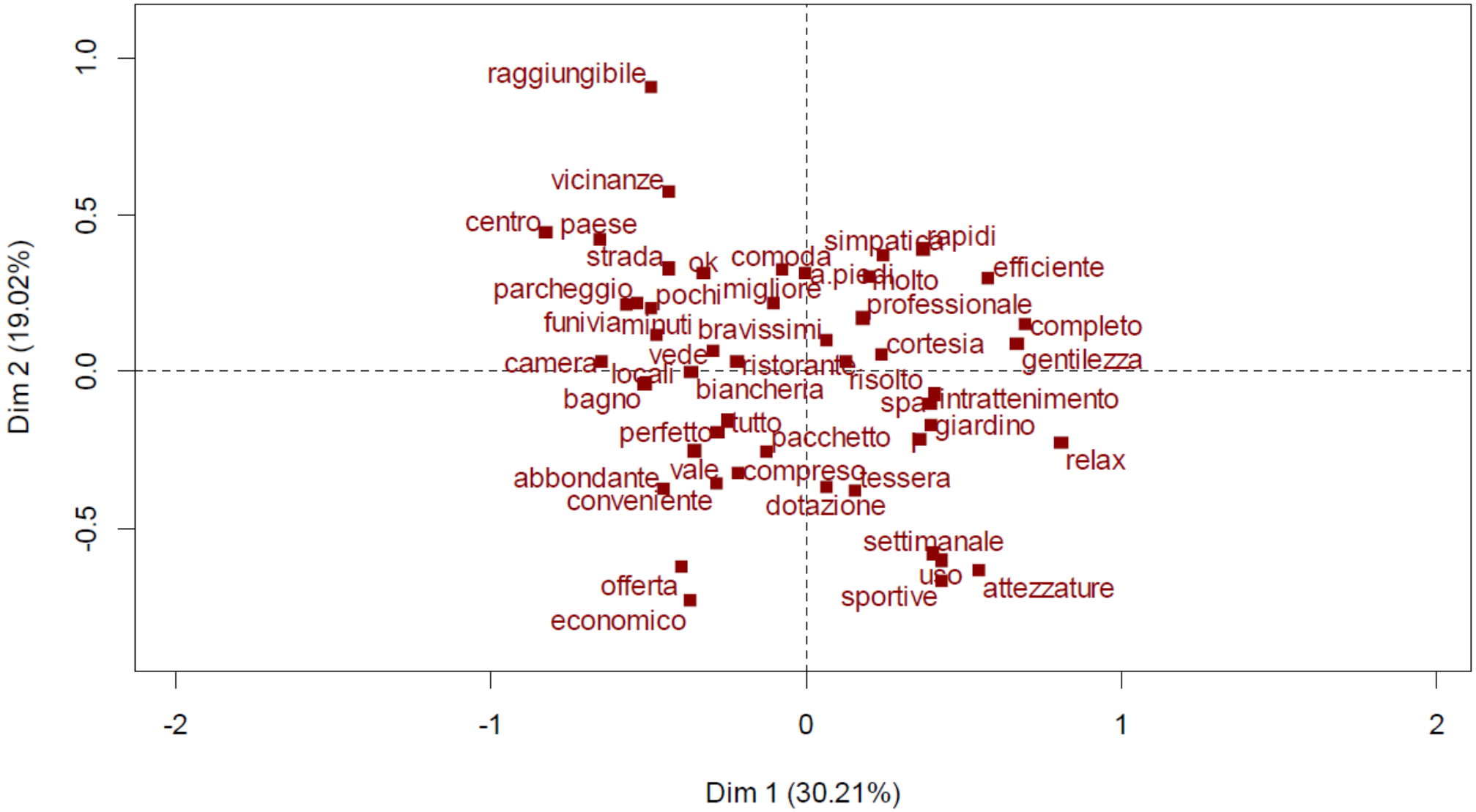
ESEMPIO: RECENSIONI DI UN ALBERGO DI MONTAGNA

Si applica l' ACL a un Corpus di recensioni relative ad un albergo in montagna.

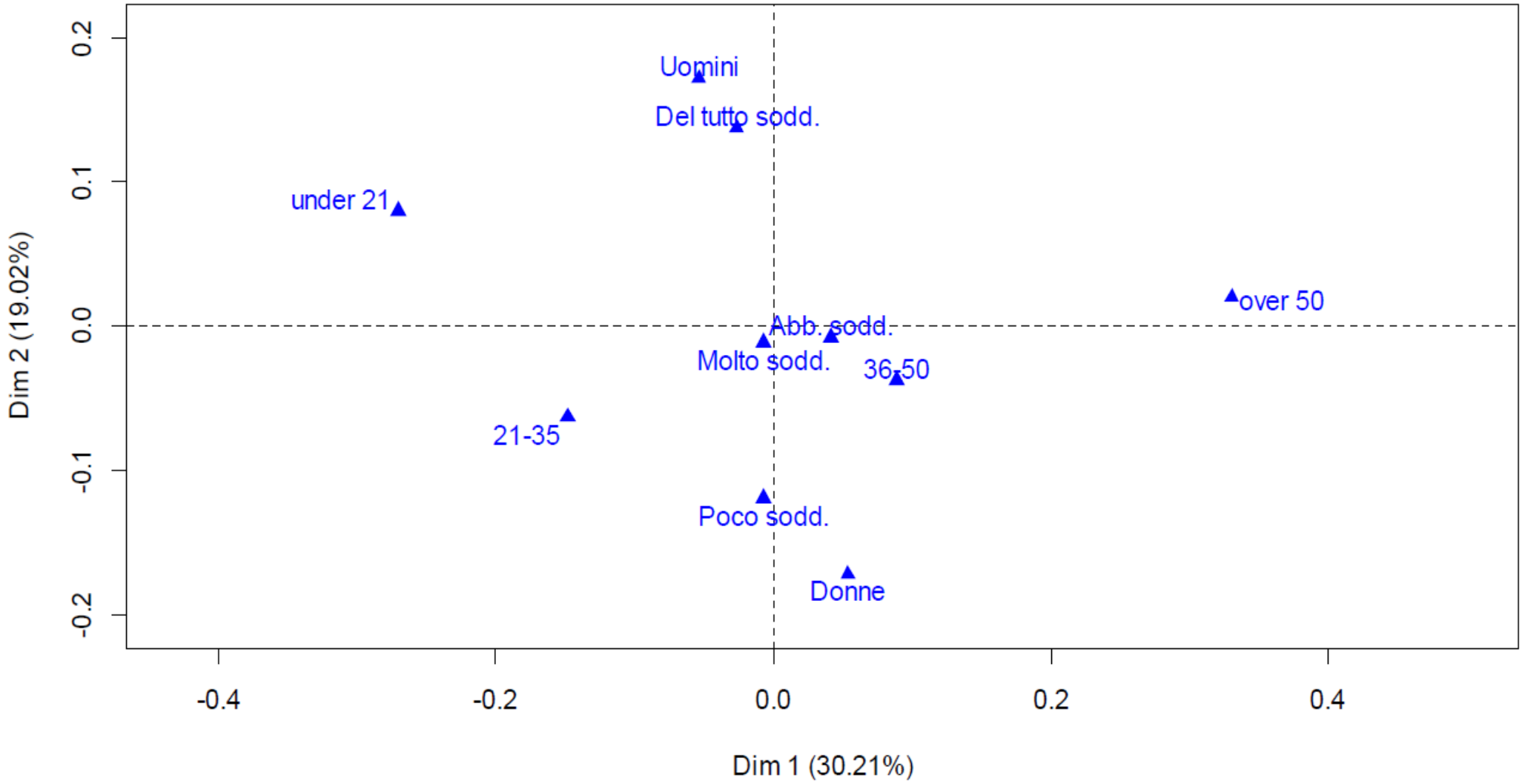
Per ogni recensione oltre al testo sono stati registrati alcuni dati strutturati tra i quali: il sesso dello scrivente, la fascia di età di appartenenza e il livello di soddisfazione indicato.

Dai testi delle recensioni sono state selezionate le forme testuali a maggior contenuto informativo, e l'ACL è stata eseguita solo su di esse.

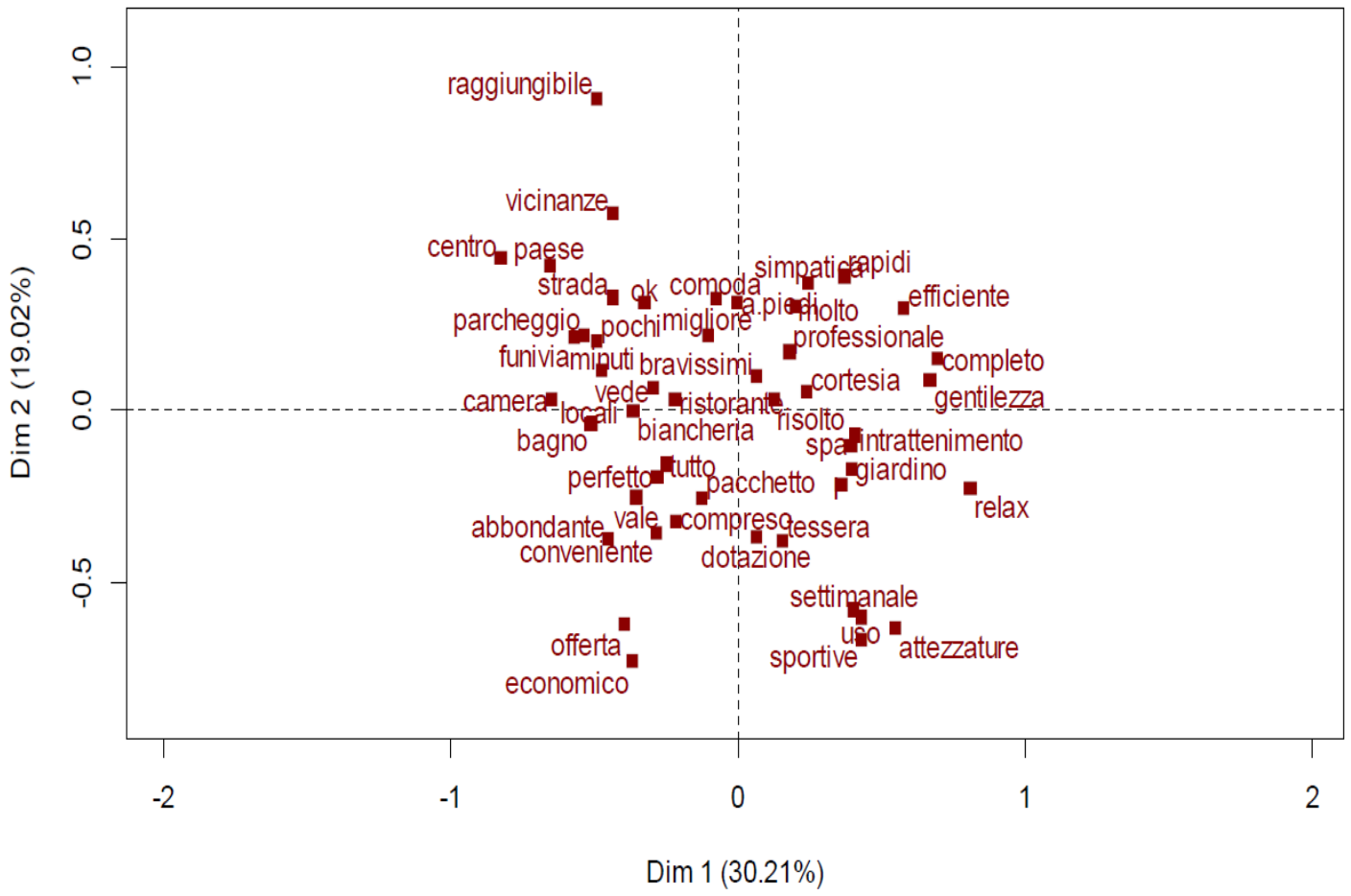
ESEMPIO: RECENSIONI DI UN ALBERGO DI MONTAGNA



ESEMPIO: RECENSIONI DI UN ALBERGO DI MONTAGNA



ESEMPIO: RECENSIONI DI UN ALBERGO DI MONTAGNA



- Dalla disposizione delle forme testuali e dei punti modalità nel piano si evince come le recensioni cui è associato il livello più alto di soddisfazione sono rilasciate principalmente da uomini e parlano principalmente della posizione dell'albergo raggiungibile, vicinanze, centro, a piedi, parcheggio, etc).

- Intorno all'origine degli assi si dispongono le recensioni dei molto ed abbastanza soddisfatti, in cui si parla prevalentemente di pulizia e di caratteristiche del servizio camere, bagno, biancheria, locali, ristorante, e poi cortesia, professionale, gentilezza, efficiente,, simpatica, etc).

- In basso ci sono recensioni dei poco soddisfatti, prevalentemente donne, in cui si parla in particolare di dotazioni dell'albergo e di aspetto economico attrezzature, dotazione, sportive, giardino, relax, tessera, settimanale, offerta, economico, compreso, conveniente, etc).

- L'età si muove fondamentalmente sul solo primo asse, da sinistra verso destra; alle età più giovani si associano i più alti livelli di soddisfazione

CONCLUSIONI

Non si può affermare che l'analisi delle corrispondenze descriva in via definitiva i significati contenuti in un insieme di testi

Ma **l'esplorazione delle associazioni tra le parole può far individuare alcune dimensioni di senso che possono contribuire alla lettura/descrizione del testo**

Prof.ssa Rosanna Cataldo

Si ringrazia il Prof. Michelangelo Misuraca per aver condiviso il suo materiale didattico dal quale è stato estratto parte delle slide presentate per l'Analisi delle Corrispondenze lessicali