Sintagmazione del testo: una scelta per disambiguare la terminologia e ridurre le variabili di un'analisi del contenuto di un corpus

Pasquale Pavone

Scuola Superiore S. Anna – Pisa – Italia

Riassunto

Nell'ambito dello studio di testi tecnico-specialistici questo lavoro suggerisce di considerare come unità d'analisi (unità lessicali) i sintagmi nominali. Si propone quindi di non limitare la tokenizzazione delle sole collocazioni nominali ma considerare tutti i sintagmi nominali, con una soglia minima di occorrenza. L'obiettivo si raggiunge mediante il riconoscimento delle strutture sintattiche che definiscono le locuzioni nominali. Tali lessie complesse, essendo costituite da forme tecnico specialistiche, polirematiche e non polirematiche, rappresentano l'universo dei soggetti ed oggetti disambiguati all'interno di un testo, ovvero la terminologia del testo. Si dimostra che l'utilizzo in analisi fattoriale di un numero limitato di sintagmi ricostruisce la stessa struttura dell'intero vocabolario in analisi. La procedura viene applicata al corpus formato dalle recensioni della Guida gastronomica del Gambero Rosso (edizioni 2004 e 2008).

Abstract

In a logic of study of technical-specialized texts, this work suggests to consider the noun phrase as the unit of analysis (lexical units). It is therefore proposed not to limit the tokenization of nominal collocation only, but to consider all the noun phrases, given a minimum threshold of occurrence. The objective is achieved through the recognition of syntactic structures that define the noun phrases. These complex lexias, being made of technical-specialized form, polysemous and not-polysemous forms, represent the universe of unambiguous subjects and objects within the text, that is the terminology of the text. It is shown how the use of factor analysis in a limited number of noun phrases is able to rebuild the same structure of the whole vocabulary in analysis. The procedure here presented is applied to the corpus made by the reviews of the gastronomic guide of Gambero Rosso Publ. (editions 2004 and 2008).

Keywords: noun phrase, automatic classification, collocation, meta-data, specialist text

1. Introduzione

Il presente lavoro mostra come la sintagmazione delle lessie complesse di tipo nominale del vocabolario permetta di disambiguare i termini di un testo tecnico-specialistico, ottenendo la terminologia del testo. Si dimostra inoltre come un gruppo limitato di sintagmi nominali sia in grado di ricostruire la struttura dell'intero vocabolario sottoposto ad analisi statistica multidimensionale.

Una particolare attenzione negli studi linguistici è dedicata al riconoscimento e tokenizzazione delle polirematiche e delle collocazioni, definite queste ultime come sequenze di due o più parole caratterizzate da un forte legame di associazione reciproca (Sinclair, 1991). All'interno

di quest'ultima vasta classe di lessie complesse si ritrovano, fra gli altri, i termini tecnici e i nomi propri composti. A riguardo, in linguaggi tecnico specialistici risulta essere significativa la presenza di lessemi complessi. Tali forme complesse sebbene non siano dotate del sovrappiù semantico rispetto ai loro componenti, come per le polirematiche, si specificano tuttavia in *accezioni non comuni, ma tecnico-specialistiche* (De Mauro 1999-2003, p. XXXII).

I sintagmi nominali rappresentano l'insieme delle polirematiche, delle forme tecnico-specialistiche e delle lessie non polirematiche. Quest'ultimo gruppo è formato dalle combinazioni di parole leggittimate dai principi generali che regolano la struttura del sintagma nominale con un significato immediatamente ricavabile dalla composizione del significato delle parole che lo formano (Lenci et al., 2005).

Il riconoscimento automatico di tali espressioni favorisce la disambiguazione delle forme semplici che le compongono, anche se non si tratta di polirematiche o collocazioni. Il senso delle parole può essere infatti determinato dalle forme che le circondano. Pertanto vincolando i significanti al loro contesto, mediante il riconoscimento e la lessicalizzazione dei sintagmi nominali, si chiarisce l'uso effettivo dei singoli sostantivi nel testo.

Il riconoscimento ed estrazione di tali lessie complesse avviene mediante la formalizzazione delle loro strutture sintattiche, grazie all'utilizzo di Regular Espressions [1] basate sulle meta-informazioni grammaticali del vocabolario in analisi.

All'interno di testi tecnico-specialistici si ritrovano un numero considerevole di forme semplici altamente ambigue in quanto associate a diversi cotesti, generatrici pertanto di diversi significati, o diversi ambiti tematici. Lo scopo del lavoro è quello di dimostrare che in corpus formati da testi tecnico-specialistici la sintagmazione delle più comuni lessie complesse di tipo nominale permette di disambiguare gli oggetti e soggetti dei testi, che costituiscono l'elemento centrale del messaggio veicolato da un enunciato. Inoltre, si dimostra come una ristretta selezione di lessie complesse sia in grado, in analisi fattoriale, di ricostruire la stessa struttura categoriale generata dall'intero corpus in analisi.

2. Modello per la realizzazione di un dizionario terminologico

Il lavoro si sviluppa in tre fasi successive: tagging grammaticale delle forme del vocabolario; ricerca delle entità di interesse e loro lessicalizzazione; analisi statistico multidimensionale per la ricostruzione della struttura del testo. La prima fase del lavoro si può definire di pretrattamento e consiste nella corretta attribuzione alle forme del vocabolario delle metainformazioni di tipo grammaticale. Gli algoritmi per poter eseguire tale fase di pretrattamento non sono oggetto della nostra analisi, in questo lavoro ci siamo limitati ad usare per tali scopi il software TreeTagger [2].

La seconda fase del lavoro consiste nell'identificazione delle strutture sintattiche che permettono il recupero dei sintagmi nominali nel testo.

[1] Notazione algebrica che permette di definire in maniera formale e rigorosa dei modelli di stringhe (Lenci et al., 2005).

[2] TreeTagger è uno strumento che permette di annotare le parole contenute nei testi con la categoria grammaticale ed il lemma appropriati. È stato sviluppato nell'ambito del TC project (http://www.ims.uni-stuttgart.de/projekte/tc) all'Institute for Computational Linguistics dell' Università di Stoccarda.

Per individuare quali fossero le strutture grammaticali in grado di estrarre i sintagmi nominali si è partiti dall'osservazione delle polirematiche nominali presenti nel Lessico dei Poliformi, risorsa linguistica presente nel software Taltac2 [3]. Tale risorsa è stata ottenuta mediante la verifica in termini di occorrenza delle forme polirematiche nel dizionario di forme composte (Elia, 1995; 1996) sul lessico dell'Italiano standard (Bolasco and Morrone, 1998).

Si nota così come il 60% dei poliformi nominali, presenti nella risorsa di riferimento, è formato da strutture del tipo <N+A> o <A+N>, e il 32% è invece formato dalla struttura <N+PREP+N>, nello specifico dalla sequenza <N+[LEMMA(di)]+N>. Il valore aggiunto di quest'ultima struttura è dato dalla preposizione *di*, comprese le forme articolate, che avendo la caratteristica di indicare una proprietà (Rouget, 2000), introduce il secondo sostantivo aggiungendo un ulteriore significato al sintagma. Questa struttura lega solidamente i due sostantivi riconoscendo in essi la testa e il modificatore, o entrambi teste di un sintagma nominale (Sinclair, 1991).

Tali strutture sono state applicate quindi come queries testuali per l'estrazione di lessie complesse. Nella fase di sperimentazione e validazione di tali espressioni regolari (RE) si è evidenziato come la struttura:

<N+[LEMMA(di)]+N> **(1)**

può essere reiterata nella forma

<N+[LEMMA(di)]+N+[LEMMA(di)]+N> **(2)**

oppure svilupparsi nelle forme

<N+[LEMMA(di)] + <A+N>> **(3)**

<N+[LEMMA(di)] + <N+A>> **(4)**

Inoltre, nel momento in cui le strutture più lunghe **(2)**, **(3)** e **(4)** generano espressioni riconosciute e lessicalizzate, la successiva lessicalizzazione delle forme più brevi ottenute dalla **(1)** determina l'eliminazione automatica degli eventuali falsi positivi ottenuti dalla stessa.

Facciamo di seguito alcuni esempi di quanto affermato precedentemente utilizzando alcune lessie proprie del linguaggio specialistico di tipo gastronomico.

Se ad esempio nel testo viene riconosciuta mediante la **(2)** la lessia più estesa <*zuppa di frutti di mare*> automaticamente non verrà riconosciuta nel vocabolario la lessia vuota di senso <*zuppa di frutti*> ottenuta mediante la **(1)**, in quanto essa è sempre compresa nella struttura **(2)**. Al contrario rimane come unità d'analisi nel vocabolario la polirematica <*frutti di mare*> che ha un utilizzo in termini di occorrenze che va oltre la specificazione della zuppa (primo sostantivo).

Allo stesso modo accade che la lessicalizzazione dei sintagmi <*insalata di frutti di mare*>, ottenuta dalla **(2)**, e <*insalata di <frutti esotici>*>, ottenuta dalla **(4)**, determina l'eliminazione della lessia vuota <*insalata di frutti*>, lasciando inalterata la lessicalizzazione delle polirematiche <*frutti di mare*> e <*frutti esotici*>.

L'applicazione delle RE per il recupero delle forme del tipo <NA> e <AN> può generare invece un numero consistente di falsi positivi se la RE non risulta essere affinata. Il gran numero di falsi positivi è infatti caratterizzato dall'associazione fra un sostantivo e un aggettivo determinativo oppure dalla dissociazione fra genere e numero dei componenti (*selvatiche ricotta*, *peperoncino tipica*). Pertanto si è proceduto al perfezionamento delle RE vincolando le ricerche

[3] TaLTaC2, acronimo che sta per Trattamento automatico Lessicale e Testuale per l'analisi del Contenuto di un Corpus, sviluppato a partire da ricerche svolte presso l'Università di Roma La Sapienza (www.taltac.it).

alle entità che presentassero coerenza fra numero e genere degli elementi ricercati, oltre che considerando esclusivamente gli aggettivi qualificativi.

Una volta validate nella fase esplorativa dell'analisi le singole RE si definisce, grazie a Taltac2, un'unica *meta-query* testuale (modello come regola) (Bolasco and Pavone, 2010) che raccoglie le singole *f(x)*, costruendo in tal modo il modello nella sua struttura complessiva. Tale vocabolario di entità può presentare tuttavia al suo interno alcuni falsi positivi che vengono depurati: prima mediante l'esclusione delle forme con meno di 5 occorrenze; e successivamente mediante la lessicalizzazione per ordine di estensione, dalle più lunghe alle più brevi, delle entità riconosciute. A questo punto le entità lessicalizzate vengono assunte come meta-dizionario (modello come risorsa disponibile).

2.1. Identificazione delle dimensioni semantiche del modello

L'ultima fase del lavoro consiste nell'analisi multidimensionale del corpus finalizzata ad individuare le dimensioni semantiche lungo cui è strutturato il testo. In particolare si esegue l'analisi delle corrispondenze semplici applicata alla matrice [forme x testi] ottenuta dall'analisi lessico-testuale svolta.

La matrice in analisi ha pertanto in riga le forme grafiche del vocabolario e in colonna le variabili categoriali in base a cui si è deciso di ripartire il testo. L'analisi delle corrispondenze consente di visualizzare sul piano grafico alcune associazioni tra parole e variabili-modalità, tali da mostrare la lettura del testo attraverso fattori che suggeriscono dimensioni di senso latenti (Bolasco, 1999). Il piano fattoriale risultante può essere interpretato in qualità di dimensioni semantiche attraverso cui leggere il corpus. In questo modo più le forme sono distanti dall'origine maggiore è il loro contributo alla determinazione degli assi, mentre la vicinanza tra le forme rinvia ad una loro cooccorrenza nel testo originario.

L'analisi viene eseguita prima sul vocabolario a soglia 5 ottenuto dalla prima tokenizzazione delle forme senza il riconoscimento di alcuna lessia complessa, in modo da stabilire quale sia la struttura del corpus in analisi. Successivamente si procede a sottoporre ad analisi delle corrispondenze sia il vocabolario formato dai soli sintagmi nominali sia il vocabolario formato da una ristretta selezione di sintagmi in grado di ricostruire l'intera struttura del vocabolario originario. L'obiettivo comparativo di tali elaborazioni è quello di evidenziare da un lato le differenze semantiche nella distribuzione dei lessici e dall'altro sottolineare la similitudine tra le strutture categoriali generate dall'intero vocabolario o da un ridotto numero di sintagmi nominali.

3. Applicazione

Di seguito verrà illustrata un'applicazione del modello di riconoscimento dei sintagmi nominali, applicato al corpus formato da 4.159 recensioni della Guida dei Ristoranti del Gambero Rosso (2004 e 2008). Le recensioni consistono in brevi schede di descrizione dei ristoranti in cui vengono elencati i piatti e i vini proposti, le particolarità del luogo e dell'accoglienza, il tipo di servizio offerto. Il corpus sottoposto ad analisi risulta formato da 28.011 forme diverse per un totale di 636.851 occorrenze totali.

Nella fase di pretrattamento del testo si è effettuato un tagging grammaticale delle forme grafiche del corpus mediante l'utilizzo del software TreeTagger, in questo modo sono state attribuite le meta-informazioni grammaticali alle singole forme del vocabolario. Il corpus così annotato è stato sottoposto ad analisi lessico testuale utilizzando il software Taltac2.

La successiva fase del lavoro è consistita pertanto nell'estrazione dei sintagmi nominali, avvenuta mediante la validazione delle espressioni regolari che ricostruiscono le strutture sintattiche dei più comuni sintagmi nominali (Tab. 1).

CATGR(N) LEMMA(di) CATGR(N) LEMMA(di) CATGR(N)
CATGR(N) LEMMA(di) CATGR(N<sm>) CATGR(A<sm>) OR CATGR(N) LEMMA(di) CATGR(N<sm>) CATGR(A<s/pm>)
CATGR(N) LEMMA(di) CATGR(N<pm>) CATGR(A<pm>) OR CATGR(N) LEMMA(di) CATGR(N<pm>) CATGR(A<s/pm>)
CATGR(N) LEMMA(di) CATGR(N<sf>) CATGR(A<sf>) OR CATGR(N) LEMMA(di) CATGR(N<sf>) CATGR(A<s/pf>)
CATGR(N) LEMMA(di) CATGR(N<pf>) CATGR(A<pf>) OR CATGR(N) LEMMA(di) CATGR(N<pf>) CATGR(A<s/pf>)
CATGR(N) LEMMA(di) CATGR(N)
CATGR(N) LEMMA(di) CATGR(NM)
CATGR(N<sm>) CATGR(A<sm>) OR CATGR(N<sm>) CATGR(A<s/pm>)
CATGR(N<pm>) CATGR(A<pm>) OR CATGR(N<pm>) CATGR(A<s/pm>)
CATGR(N<sf>) CATGR(A<sf>) OR CATGR(N<sf>) CATGR(A<s/pf>)
CATGR(N<pf>) CATGR(A<pf>) OR CATGR(N<pf>) CATGR(A<s/pf>)
CATGR(A<sm>) CATGR(N<sm>) OR CATGR(A<sm>) CATGR(N<s/pm>)
CATGR(A<pm>) CATGR(N<pm>) OR CATGR(A<pm>) CATGR(N<s/pm>)
CATGR(A<sf>) CATGR(N<sf>) OR CATGR(A<sf>) CATGR(N<s/pf>)
CATGR(A<pf>) CATGR(N<pf>) OR CATGR(A<pf>) CATGR(N<s/pf>)
CATGR(N) LEMMA(a) CATGR(N)

Tabella 1: Meta-query testuale

Rispetto alle RE introdotte nel precedente paragrafo, sono state aggiunte altre strutture sintattiche nate dall'esplorazione del corpus in analisi.

In questa fase esplorativa è risultata infatti particolarmente significativa, date le caratteristiche del corpus, la struttura sintagmatica <CATGR(N) LEMMA(a) CATGR(N)>. Mediante questa RE si recuperano determinate espressioni tipiche della gastronomia, dove generalmente la preposizione *a* introduce la tipologia di preparazione della pietanza (*patate al forno, baccalà alla livornese*).

La RE <CATGR(N)+Lemma(di)+CATGR(NM)> (dove per NM si intendono Toponimi) ha permesso di estrarre i cibi relazionati con luoghi geografici. Tale struttura semantica si presenta come elemento di differeziazione qualitativa di un piatto o di un elemento culinario rispetto ad un'altro (*pasta di Gragnano*, *olive di Gaeta*) ma al tempo stesso ha permesso di estrarre anche alcune polirematiche tipicamente gastronomiche come ad esempio *pan di Spagna* e *fichi d'India*.

Per quanto riguarda le strutture <NA> e <AN>, gli aggettivi considerati sono esclusivamente di tipo qualificativo e vincolati per genere e numero al sostantivo rappresentante la testa del sintagma, in modo da ridurre l'estrazione di falsi positivi.

Una volta validate tutte le singole RE si è creata la meta-query che ha permesso dunque il riconoscimento di un totale di 44.302 lessie complesse aventi le predette strutture sintattiche.

Dall'osservazione dei sintagmi ottenuti dal testo si evidenzia particolarmente la variabilità linguistica nella gastronomia. Risulta interessante infatti osservare le numerose varianti tematiche di singole forme di per sé ambigue. Ad esempio la forma grafica *salsa* presenta 264 varianti sintagmatiche (*di pomodoro*, *di fragola*, *di cioccolato*, *d'acciughe*, *di birra*, *di cachi*), e il riconoscimento del sintagma nominale di appartenenza permette evidentemente la corretta attribuzione tematica (primo o secondo piatto, a base di carne o di pesce, rispetto al dolce). Allo stesso modo vengono tematicamente disambiguate le forme: *crema* con 257 varianti (*di patate*, *catalana*, *di limone*, *di broccoli*, *di ortiche*); *tortino* con 146 varianti (*di cioccolato*, *di alici*, *di castagne*, *d'agnello*, *di ananas*, *di cicoria*) e *mousse* con 104 varianti (*di cioccolato*, *di baccalà*, *di fichi*, *di gorgonzola*, *di fegato*).

Mediante il riconoscimento di tali espressioni sintagmatiche si è pertanto riusciti ad attribuire una corretta categoria semantica a determinati sostantivi, come nei casi sopra riportati, i quali se osservati singolarmente riescono ad esprimere solamente l'aspetto esteriore delle vivande.

La disambiguazione semantica non ha però riguardato solamente classiche parole di per sé incapaci di rappresentare esattamente il significato del cibo descritto, ma ha coinvolto anche termini che nel linguaggio comune hanno un senso univoco e comunemente inteso. Ad esempio, sono state rilevate 128 varianti di *ragù* (*di carne*, *di pesce*, *di broccoli*, *di lumache*) e 88 varianti di *carpaccio* [4] (*di tonno*, *di manzo*, *di carciofi*, *di noci*, *di melone*). Tali termini posseggono autonomamente un significato, o meglio un'appartenenza tematica, che viene smentita nel loro uso e che si mette in luce solamente se si evidenzia l'eventuale modificatore.

Il modello di sintagmazione produce a questo punto una risorsa mediante la lessicalizzazione delle espressioni riconosciute aventi almeno 5 occorrenze.

Sono state pertanto riconosciute come nuove entrate del vocabolario, unità d'analisi lessicale, 2.570 sintagmi nominali aventi un totale di 38.299 occorrenze.

Il vocabolario è così passato da 28.011 entrate (prima della lessicalizzazione) a 30.534 forme. L'aumento delle unità in analisi risulta essere il fattore che costituisce la disambiguazione dei termini e la varietà dei significanti. Si noti come il numero delle entrate del vocabolario lessicalizzato non coincide con la somma delle entrate prima della lessicalizzazione e il numero dei sintagmi, in quanto le occorrenze di alcune forme sono state assorbite per completo all'interno dei sintagmi lessicalizzati.

In Tab. 2, vengono riportati alcuni risultati della lessicalizzazione, nello specifico riferiti al gruppo sintattico del tipo <N di N> che è costituito da un totale di 936 lessie complesse.

carta dei vini	1.950	*Verdure di stagione*	71	***gallinella di mare***	14
frutti di mare	236	*filetto di maiale*	69	*carpaccio di branzino*	13
selezione di formaggi	216	***tortino di cioccolato***	68	*gnocchi di zucca*	13
fiori di zucca	180	[...]	[...]	***millefoglie di melanzane***	13
lista dei vini	180	***ragù di pesce***	23	*ragù di cinghiale*	13
frutti di bosco	152	[...]	[...]	*tortino di alici*	13
dolci della casa	115	*mousse di ricotta*	18	[...]	[...]
petto d'anatra	108	***tortino di melanzane***	18	***asparagi di mare***	9
piatti della tradizione	107	*fragoline di bosco*	17	*carpaccio di baccalà*	9
filetto di manzo	107	***gelato di crema***	17	***tonno di coniglio***	9
Mozzarella di bufala	106	*ravioli di patate*	17	***millefoglie di patate***	8
gnocchi di patate	105	***crema di zucchine***	17	***flan di verdure***	8
[...]	[...]	*caponata di melanzane*	17	***caviale di melanzane***	8
tagliata di manzo	74	[...]	[...]	***sella di cervo***	8

Tabella 2: Sintagmi del tipo <N di N>

In neretto sono stati evidenziati alcuni sintagmi che maggiormente sottolineano la capacità di disambiguazione ottenuta con il procedimento di sintagmazione. Tale disambiguazione, come precedentemente anticipato, può essere di tipo tematico, come nel caso delle varianti di tortino, oltre che rappresentare una vera e propria disambiguazione di significato dei termini, si veda ad

[4] Carpaccio, pietanza di carne cruda affettata molto sottile e condita con olio e scaglie di formaggio grana, Dizionario Garzanti, www.garzantilinguistica.it.

esempio *gallinella di mare* [5], *asparagi di mare* [6] e *tonno di coniglio* [7], dotate evidentemente di uno *specifico sovrappiù semantico*.

Un ulteriore aspetto della capacità di disambiguazione del procedimento svolto emerge osservando l'attitudine dei sostantivi di generare lessie complesse. Infatti considerando solamente i sostantivi con almeno15 varianti sintagmatiche, si osserva come siano solo 58 sostantivi a generare 1.429 lessie complesse.

L'ultima fase del lavoro è consistita nella generazione delle matrici [forme x testi] da sottoporre ad analisi statistica multidimensionale. Si sono create quattro diverse matrici, tutte con gli stessi raggruppamenti per classi di prezzo, di punteggio, di numero di coperti e regione, aventi per unità lessicali, con almeno 5 occorrenze, quattro diversi vocabolari, qui di seguito specificati:

1. La prima matrice sottoposta ad analisi è costituita dal vocabolario formato da 7.339 forme risultante dalla prima tokenizzazione, senza alcuna lessicalizzazione.
2. La seconda matrice sottoposta ad analisi è costituita dal vocabolario ottenuto dopo aver lessicalizzato i sintagmi riconosciuti dal modello, costituito da 9.641 forme.
3. La terza matrice è composta esclusivamente dai 2.570 sintagmi nominali lessicalizzati.
4. La quarta matrice è formata solamente da 205 sintagmi nominali, selezionati sulla base del contributo nella spiegazione dei due assi fattoriali, frutto dell'analisi eseguita sulla precedente matrice.

I risultati delle analisi fattoriali sono sintetizzati mediante grafici che consentono di definire le configurazioni dei punti sui piani di proiezione formati da coppie di assi fattoriali. In Fig. 1 è rappresentata la struttura del corpus su cui si distribuisce la nuvola di vocaboli (riportata nella miniatura). Tale mappa fattoriale è il risultato dell'analisi delle corrispondenze effettuato sulla prima delle quattro matrici. Si può osservare quindi come il primo fattore riesce a spiegare per completo la variabile prezzi mentre il secondo fattore spiega l'opposizione semantica fra cucina a base di carne e cucina a base di pesce. Il punteggio dei ristoranti, segnalato sulla mappa dalla linea verde, risulta essere strettamente dipendente dal prezzo, così come il numero di coperti, segnalati in mappa dalla linea grigia. Tendenzialmente ad un numero minore di coperti sono associati un prezzo ed un punteggio maggiori.

Nella miniatura della nuvola delle forme sono stati evidenziati 2 sostantivi protagonisti della sintagmazione. In questo modo si vuole sottolineare come in tale matrice di forme semplici, il loro contributo nella spiegazione degli assi fattoriali risulta essere poco significativo, occupando una posizione pressoché centrale nella mappa fattoriale.

In Fig. 2 sono riportate le distribuzioni degli individui della seconda matrice in analisi, formata da 9.641 forme del vocabolario a soglia 5 comprendente i 2.570 sintagmi lessicalizzati. Nello specifico sono stati evidenziati solamente i sintagmi riferiti alle forme tortino (a) e baccalà (b).

Questi due grafici mostrano come i sintagmi comprendenti le forme *tortino* e *baccalà* acquisiscono un ruolo differente nella spiegazione degli assi fattoriali rispetto alle forme semplici della precedente matrice in analisi.

[5] **Gallinella di mare o Chelidonichthys lucernus**, è un pesce teleosteo della famiglia Triglidae.

[6] **Asparagi di mare o Salicornia**, genere di piante erbacee dal fusto carnoso, commestibile, frequenti nei terreni salini (fam. Chenopodiacee).

[7] Il **tonno di coniglio** è una ricetta tipica del Monferrato (Piemonte). Si chiama in questo modo poiché la carne di coniglio, stando nell'olio per qualche giorno a macerare, diventa tenera come fosse un tonno.

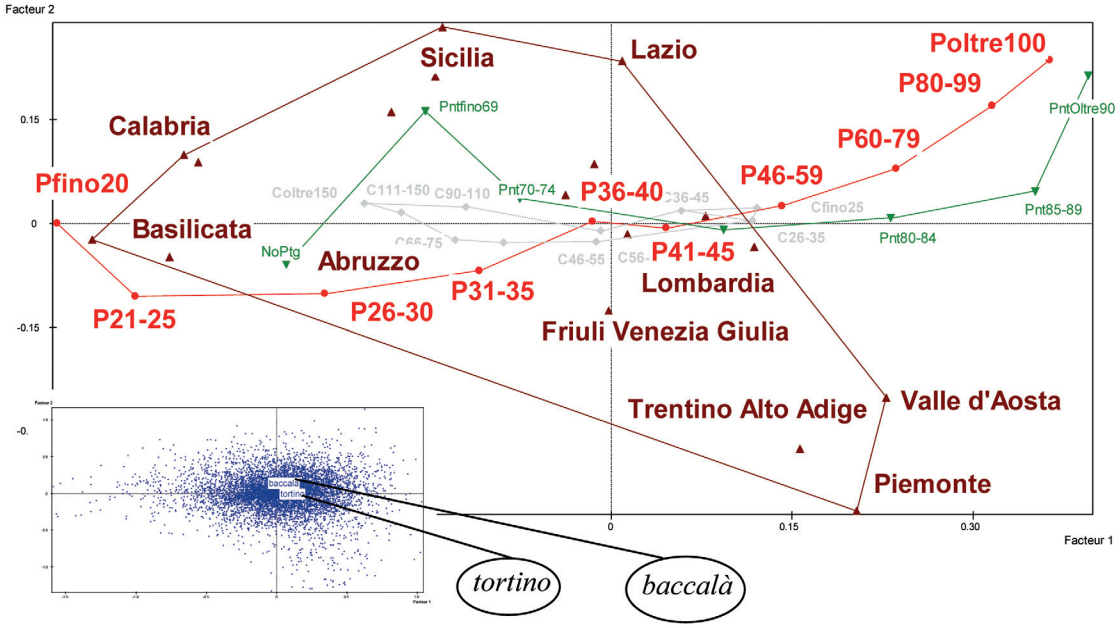
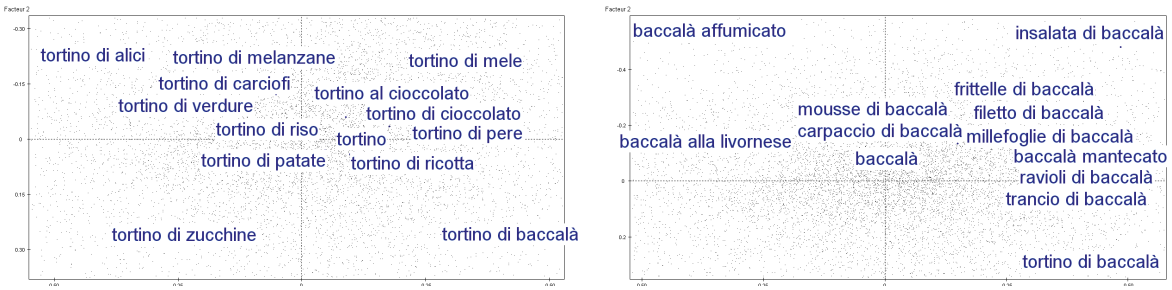


Figura 1: Distribuzione delle variabili categoriali della matrice 1, formata da 7.339 forme



(a) (b)

Figura 2: Distribuzione del vocabolario lessicalizzato, 9.641 forme – selezione dei sintagmi contenenti le forme " tortino" (a) e "baccalà" (b)

Si possono a questo punto fare delle riflessioni qualitative sui risultati ottenuti. Ad esempio, si denota come il baccalà sia una pietanza specifica dei ristoranti medio alti, fatta eccezione del *baccalà alla livornese* e il *baccalà affumicato*, posizionati sul lato sinistro della mappa fattoriale e quindi maggiormente presenti nelle trattorie. D'altra parte, invece, si può osservare come il tortino sia una portata "salata" all'interno del lessico dei ristoranti medio bassi. Al contrario il tortino entra nel lessico dei ristoranti medio alti come dessert, fatta però eccezione del tortino di baccalà, sintagma ottenuto proprio dall'unione delle due forme e caratteristico dei ristoranti più cari e tendenzialmente presente in un ambito di cucina non di mare (vista la sua posizione sul piano fattoriale).

Procedendo nella legittimazione della scelta della sintagmazione i due seguenti grafici si giustificano per sottolineare come finanche una ristretta selezione di sintagmi nominali sia in grado di riassumere la stessa struttura dell'intero vocabolario in analisi.

In Fig. 3 è riportata quindi la struttura ottenuta dalla matrice formata dai soli 2.570 sintagmi nominali. Da quest'ultima analisi fattoriale sono stati quindi selezionati i sintagmi con un maggior contributo nella spiegazione degli assi fattoriali. A partire dai 205 sintagmi così selezionati si è costruita l'ultima matrice in analisi, i cui risultati sono rappresentati in Fig. 4.

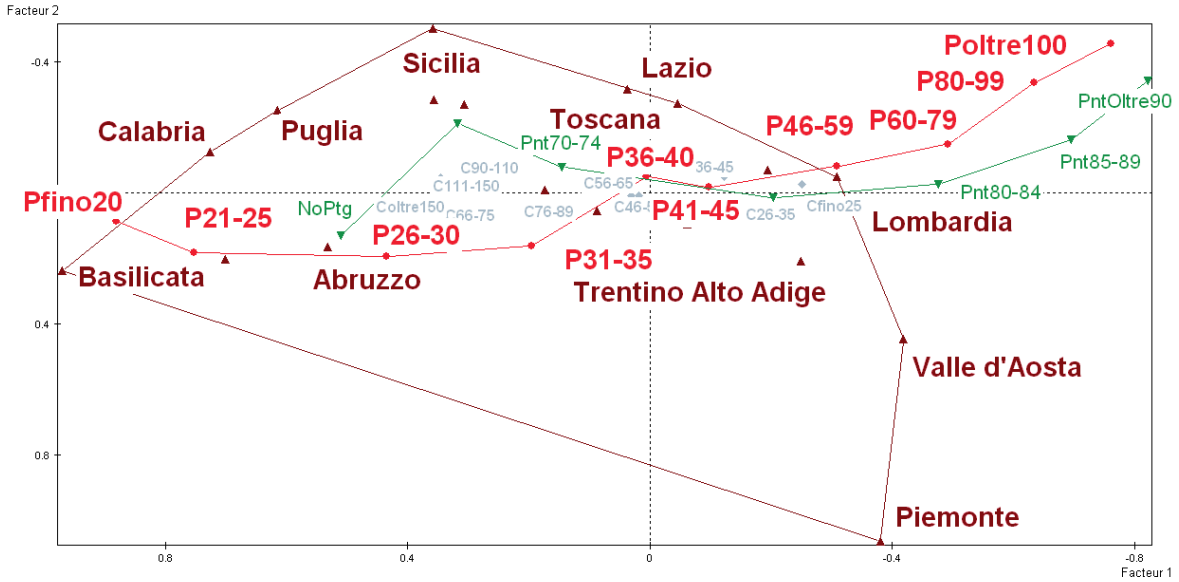


Figura 3: distribuzione delle variabili categoriali della matrice 3, formata dai 2570 sintagmi nominali

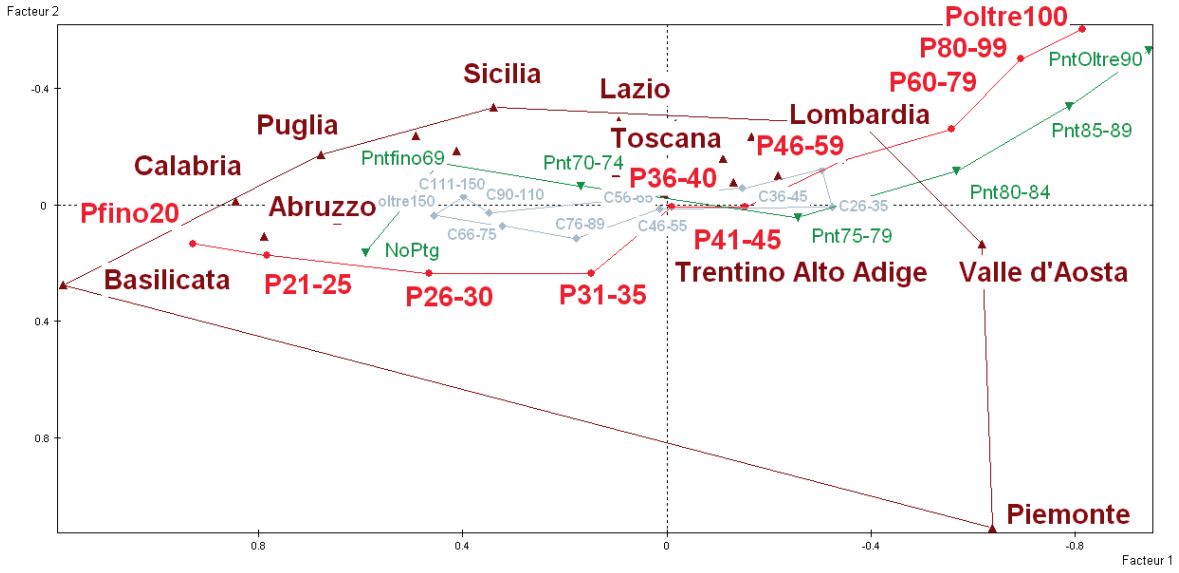


Figura 4: distribuzione delle variabili categoriali della matrice 4, formata da 205 sintagmi nominali

Si può osservare quindi che la distribuzione strutturale delle variabili categoriali rimane pressocchè identica, dalla matrice del vocabolario originario formato dalle 7.339 forme fino alla matrice formata dai 205 sintagmi nominali. Sottolineando, così, come sia la terminologia a sostenere la struttura di un vocabolario.

4. Conclusioni

Come si è visto la scelta della sintagmazione delle lessie nominali di un testo specialistico, come può esserlo il corpus di critica gastronomica, ha permesso di disambiguare i termini del testo sia rispetto ai temi che rispetto al significato stesso delle parole. Utilizzando le strutture più comuni di rappresentazione dei sintagmi nominali come modello per la ricerca delle entità,

è stato possibile estrarre le poliremathiche e le collocazioni presenti nel testo, oltre che delle forme sintagmatiche "semplici", ovvero non dotate di alcun sovrappiù semantico, che però non hanno pregiudicato l'analisi ma al contrario sono riuscite a far risaltare espressioni tipiche di differenti ambiti lessicali.

Mediante la sintagmazione si è passati dalle parole alla terminologia, con un notevole incremento qualitativo dell'analisi automatica del testo.

Riferimenti bibliografici

Bolasco S. (1999). *L'analisi multidimensionale dei dati*. Carocci: Roma.

Bolasco S. and Morrone A. (1998). *La construction d'un lexique fondamental de polyformes selon leur usage*. In *JADT 1998*, Université de Nice, pp. 155-166.

Bolasco S. and Pavone P. (2010). Automatic Dictionary and Rule-Based Systems for Extracting Information from Text. In Palumbo, F., Lauro, C.N. and Greenacre, M., editors, *Data Analysis and Classification. Proceedings of the 6th Conference of the Classification and Data Analysis Group of the Società Italiana di Statistica*, Berlin: Springer, pp. 189-198.

De Mauro T. (1999-2003). *Grande Dizionario Italiano dell'Uso*. Torino: Utet.

Elia A. (1995). Per una disambiguazione semi-automatica di sintagmi composti: i dizionari elettronici lessico-grammaticali. In Cipriani, R. and Bolasco, S., editors, *Ricerca qualitativa e computer*, Milano: Franco Angeli.

Elia A. (1996). *Per filo e per segno: la struttura degli avverbi composti.* In D'Agostino, E., editor, *Sintassi e semantica*, Napoli: ESI, pp. 167-263.

Lenci A., Montemagni S. and Pirrelli V. (2005). *Testo e computer*. Roma: Carocci.

Rouget C. (2000). *Distribution et sémantique des constructions Nom de Nom*. Paris: Honoré Champion.

Sinclair J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.