Automatic tagging

9.1 Introduction

In Chapter 8, I used automatic semantic tagging to verify the validity of different types of sampling procedures. On that occasion, automatic tagging of the whole datasets (the elicited, as well as the Web ones) was compared to automatic tagging of sampled subsets. In the current chapter, the same automatic semantic tagger – Wmatrix, the automatic semantic tagger developed at the University of Lancaster – will be applied in order to assess whether it could fruitfully replace manual coding in establishing cultural associations of the given node words (R.Q. 4). More concretely, this chapter compares the results obtained by manual tagging (see Chapter 6) to those obtained using Wmatrix. Since Wmatrix does not treat Italian and no semantic tagger based on a similar coding scheme exists for this language, the current chapter will analyse only the English elicited datasets.

As we have seen in the previous chapters, manual semantic tagging is not only time-consuming, but also highly demanding: it requires the work of at least two well trained coders, as well as an intense effort from each of them in terms of coherent and cohesive application of the given coding scheme. On the other hand, an automated coding procedure would reduce the number of coders to a single researcher, take only a few minutes, and guarantee effortless systematic application of the coding scheme.

In a preliminary experiment (Bianchi, 2010; see also Chapter 4), the *chocolate* and *wine* elicited datasets underwent automatic tagging using Wmatrix and the results of the automatic tagging were compared to manual coding at the level of conceptual domains (superordinate, broader categories) and of semantic fields, by applying the USAS-Codebook conversion scheme described further on in the current chapter. At the level of conceptual domains, the conversion scheme was applied to the top 30 items in the semantic frequency list and in the semantic keyword list of the elicited data as offered by Wmatrix, excluding grammatical items. As an intermediate step between manual tagging (sentence-based) and semantic tagging (word-based), it was decided to consider also the top 30 items of the raw frequency list and of the keyword list, as this allowed manual tagging to be applied on the basis of individual words. Therefore, the top 30 semantic items in the lists (excluding the node word) were manually mapped to one or more of the conceptual domains described in the Codebook. Those analyses were then compared to the results of manual coding of the whole elicited datasets, which showed that the semantic frequency list performed

generally better than the other lists. In fact, it retrieved the same or a higher number of domains and systematically showed strong correlation values at the Spearman test. At the level of semantic fields, comparison was performed using the most frequent 50 items in the semantic frequency list and in the semantic keyword list. When using the semantic frequency lists, the data consistently showed levels of correlation in the modest range, with results for *chocolate* being $r = 0.505$ (at $p < 0.01$), and for *wine* $r = 0.558$ (at $p < 0.01$); when using the semantic keyword list, results were less consistent, with strong correlation results for *chocolate* ($r = 0.703$ at $p < 0.01$) and modest correlation results for *wine* ($r = 0.486$ at $p < 0.01$). Finally, the preliminary experiment compared the semantic word lists of the elicited data to the semantic word lists of the Web data. For the sake of experimentation, correlation was computed in three different ways: (1) using the whole semantic frequency lists, (2) using the top 100 items in the lists; and (3) using the top 50 items. All the six cases (three for *chocolate* and three for *wine*) showed interesting positive correlation between the elicited and the Web data, the strength of the correlation decreasing from strong to medium to low-medium as the number of items considered decreased.

The current chapter banks on results of the preliminary experiment described above and expands it in the following directions: 1. expanding the number of items considered in the semantic frequency list; 2. considering highly conventionalised fields/domains and cultural associations; 3. analysing prosody; 4. comparing the results to our 'control situation' – i.e. to the results obtained with manual coding of the whole elicited datasets. Furthermore, in Section 9.4, the results of automatic coding will be compared also to manual coding of the most frequent 150 words in the wordlist.

9.2 Matching automatic tagging categories to manual coding ones

For the purpose of comparing automatic tagging to manual tagging, automatic semantic tagging was applied to the English elicited data using Wmatrix and the USAS tagset (see Chapter 5, Section 5.3.2). The semantic structure adopted in the USAS tagset is rather different from the one developed and used in the manual tagging process. However, as we shall see in the following paragraphs, comparisons are still possible, by applying a conversion process similar to that used for matching the UCREL semantic taxonomy to that of the Collins English Dictionary (CED) and described by Archer, Rayson, Piao and McEnery (2004).

To allow comparison, the USAS tags were matched to the semantic fields used in the manual coding of the elicited data. For each tag, matching was accomplished by looking at the prototypical examples provided in Archer, Wilson and Rayson (2002), imagining them in the given context (i.e. next to the words *chocolate* and *wine*, but also in the wider context of general speech), and finding a suitable semantic field in the manual tagging list. Examples of matching are provided in Table 9_1.

In the table, the words or expressions specified in the manual coding columns refer to the Codebook semantic field; double slashes (//) indicate that matching is 'one-to-many'. The word 'Other' indicates no matching. For the matching between Codebook semantic fields and conceptual domains, please see Table 2 in the Appendix.

USAS tag	USAS semantic category	*Chocolate* manual coding	*Wine* manual coding
O4.6+	Temperature: Hot/on fire	// Drink // Other	// Storage // Other
O1.1	Substances and materials: solid	// Food // Other	// Food // Other
I2.2	Business: Selling	Transaction	Transaction
X3.1	Sensory: Taste	Taste	Taste
E2-	Dislike	Passion	Passion
L1+	Alive	Existence	Existence
S3.1	Personal relationship: General	Friendship	Friendship
A2.1+	Change	Other	Other
A1.5.1	Using	Other	Other

Table 9_1. Conversion schemes: some examples

Different conversion schemes were necessary in order to account for the different fields of the two key words. For example, the elicited corpus showed that USAS tag O4.6+ (Temperature: Hot/on fire), which corresponds primarily to the word 'hot', tends to refer to different semantic fields when next to the word 'chocolate' or 'wine': if chocolate is hot, it is a drink; if wine is hot, we are talking about a storage issue. However, given that both chocolate and wine belong to the same general category of food and drinks, the two conversion schemes show a limited number of differences. A given USAS tag could match one or more categories of the manual codes, or even none of them. Matching was not sought for categories indicating logical or grammatical relations (Table 9_2). Indeed these categories were disregarded in all the analyses.

Code	Description	Code	Description	Code	Description
Z4	Discourse Bin	Z99	Unmatched	A13.3	Degree: Boosters
Z5	Grammatical bin	A7	Probability	A13.4	Degree: Approximators
Z6	Negative	A7+	Likely	A13.5	Degree: Compromisers
Z7	If	A7-	Unlikely	A13.6	Degree: Diminishers
Z7-	Unconditional	A13	Degree	A13.7	Degree: Minimisers
Z8	Pronouns	A13.1	Degree: Non-specific	A14	Exclusivisers/particularisers
Z9	Trash can	A13.2	Degree: Maximisers	N1	Numbers

Table 9_2. Categories excluded from analysis

One of the major issues in matching two different schemes of this type is how to distribute frequency in the case of 'one-to-many' matching. In this study, when the matching scheme presented 'one-to-many' mapping (about 34% of cases for semantic fields and 30% of cases for conceptual domains, in both datasets), the frequency of the USAS tag was equally distributed among all of the possible matching domains/fields. So, for example the USAS conceptual domain SUBSTANCES AND MATERIALS: SOLID (78%) was equally distributed between Codebook domain FOOD (39%), and in category OTHER (39%). Though this clearly leads to an approximation, it seemed the only possible solution, since manual tags refer to the relationship that exists between the key word (*chocolate* or *wine*) and the rest of the sentence, while automatic tags describe individual words, regardless of the key word. Manually looking at individual concordances in order to recreate the relationship to the key word was discarded in this case, as the aim of the study is precisely to investigate and assess automated procedures.

9.3 Analyses

The top 50/100/150 items in the semantic frequency list of the English elicited datasets were compared to the results of manual tagging of the same datasets (see Chapter 6). The 150 limit was arbitrarily chosen considering that a semantic category conflates one or more words in the dataset. Consequently, the top 150 items in the semantic frequency list represent a percentage of the whole dataset which is certainly higher than that of the most frequent 150 words in the wordlist. Consequently, considering that in Chapter 7 over 90% of the highly conventional semantic fields appeared with as few as about 300 words, it seemed reasonable to hypothesise that an even smaller number of the most frequent semantic categories could be enough to highlight all or most of the cultural associations of the node words.

Comparison was performed both qualitatively, and quantitatively, at the level of semantic fields and conceptual domains. In other words, the most frequent 150 USAS tags, once converted into Codebook semantic fields, were compared to semantic fields Tables 6_1 and 6_8 and to conceptual domains Tables 6_4 and 6_11 in Chapter 6. The following paragraphs summarise the results of this comparison.

The results of these qualitative and quantitative comparisons between the most frequent 150 USAS categories in the English elicited datasets and manual semantic analysis of the datasets, at the level of semantic fields, are summarised in Tables 9_3 and 9_4. Column one shows the number of most frequent (Top) semantic tags considered; columns two reports the overall percentage of fields covered (with reference to tables 6_1 and 6_8). Columns three and four show the percentage of highly conventionalised fields (H Cnv) and cultural associations (H+M Cnv) covered. Column five summarizes field increases in passing from one threshold to the next. Finally, the last column reports the results of Spearman's Rank Correlation test (for $p < 0.01$). Percentages are rounded to the second decimal.

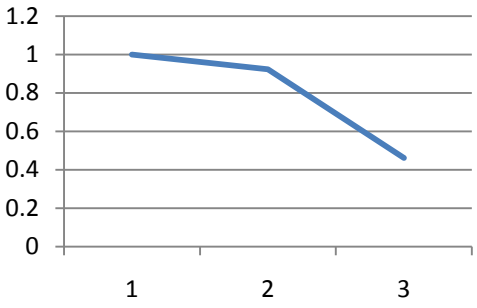
Matched USAS fields	Codebook fields (%)	H Cnv (%)	H+M Cnv (%)	Field increase	Spearman's rho
TOP 50	28.41	34.29	35.59	+ 26 fields	0.505
TOP 100	54.55	57.14	66.10	+ 24 fields	0.503
TOP 150	67.05	74.29	79.66	+ 12 fields	0.492

Table 9_3. English *chocolate* elicited dataset: semantic field comparison

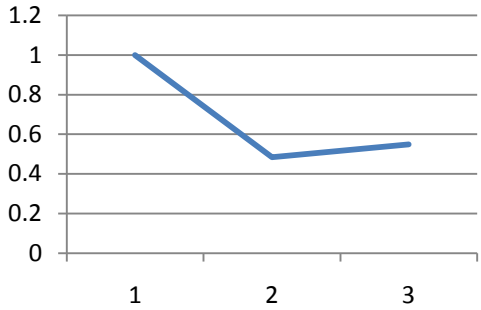
Matched USAS fields	Codebook fields (%)	H Cnv (%)	H+M Cnv (%)	Field increase	Spearman's rho
TOP 50	36.47	60.00	55.77	+ 31 fields	0.558
TOP 100	52.94	80.00	73.08	+ 15 fields	0.584
TOP 150	68.24	80.00	80.77	+ 17 fields	0.525

Table 9_4. English *wine* elicited dataset: semantic field comparison

The most frequent 150 items in the USAS frequency list – which represent 56% of each list – showed about 67-68% of the Codebook fields highlighted with manual tagging. This corresponds to 74-80% of the highly conventionalised fields and about 80% of the cultural associations (fields with high or medium conventionalisation). Furthermore, as already noticed in Chapter 8, Zipf's law does not seem to apply to field increases at different thresholds (see Graphs 9_1 and 9_2), below.



Graph 9_1. Data from Table 9_3



Graph 9_2. Data from Table 9_4

Finally, Spearman's test results are all in the modest range, a result which is similar to the one obtained in the preliminary experiment (Bianchi, 2010). Furthermore, differently from what noticed in Chapter 8, no increasing tendency can be seen when moving from one threshold to the next.

At the level of conceptual domains, the situation is summarised in Tables 9_5 and 9_6, below.

USAS fields	Overall Codebook domains (%)	H Cnv (%)	H+M Cnv (%)	Domain increase	Spearman's rho
TOP 50	66.67	100	81.82	+ 10 fields	0.810
TOP 100	86.67	100	90.91	+ 3 fields	0.881
TOP 150	93.33	100	100	+ 1 fields	0.904

Table 9_5. English *chocolate* elicited dataset: conceptual domain comparison

Matched USAS fields	Overall Codebook domains (%)	H Cnv (%)	H+M Cnv (%)	Domain increase	Spearman's rho
TOP 50	66.67	100	90.00	+ 10 fields	0.545
TOP 100	80.00	100	100	+ 2 fields	0.763
TOP 150	93.33	100	100	+ 2 fields	0.429

Table 9_6. English *wine* elicited dataset: conceptual domain comparison

The most frequent 150 items in the USAS frequency list – which represent 56% of each list – showed about 93% of the Codebook domains highlighted with manual tagging, and 100% of the highly conventionalised fields and of the cultural associations. The majority of domains entered the picture already in the top 50 items. Finally, Spearman's test results are in the strong range for *chocolate*, but in the modest range for *wine*. Furthermore, at least in the case of *wine*, Spearman's rho does not increase as the number of USAS fields considered increases.

As regards semantic prosody, i.e. when the semantic categories adopted for analysis fall into evaluative categories (see Chapter 3, Section 3.6.4), the USAS tagset includes a specific category (A5) subdivided into 4 subcategories: 'A5.1 Evaluation: Good/bad', 'A5.2 Evaluation: True/False', 'A5.3 Evaluation: Accuracy', and 'A5.4 Evaluation: Authenticity'. Within each category, plus (+) or minus (-) signs indicate positive or negative evaluation, respectively. In the most frequent 150 semantic items,

this category appeared with a clear predominance of positive evaluation. In quantitative terms, *chocolate* showed 70 positive words vs. 30 negative ones, i.e. a positive evaluation which is about 2.3 times bigger than the negative one. *Wine* showed 184 positive words vs. 30 negative ones, with positive evaluation being about 6 times bigger than the negative one. These results are comparable to manual tagging of the two elicited datasets (our control situation) in qualitative terms, but not in quantitative ones (Chapter 6, Table 6_15). In fact, in the whole manually coded datasets, positive assessment was 2.8 times bigger than negative assessment for *chocolate*, and 2.4 times bigger for *wine*.

9.4 Concluding remarks

In the current chapter, the English elicited datasets were automatically tagged with Wmatrix, and the most frequent 150 items in the resulting semantic frequency lists were compared to the results of manual coding of the entire datasets, at the level of semantic fields, conceptual domains, and semantic prosody. Since the semantic structure adopted in the USAS tagset is rather different from the one developed and used in the manual tagging process, a conversion scheme was applied which matched the USAS tags to the semantic fields used in the manual coding of the elicited data.

At a qualitative level, the results are encouraging. In fact, comparison showed that the most frequent 150 items in the USAS frequency list – which represent 56% of each list – showed about 67-68% of the Codebook fields highlighted with manual tagging, and about 93% of the conceptual domains, including 74-80% of the highly conventionalised fields and about 80% of the cultural associations, and 100% of the highly conventionalised and cultural domains. Furthermore, the most frequent 150 USAS categories in the semantic frequency list showed marked preference for positive, rather than negative assessment, as was the case in the control situation.

From a quantitative perspective, correlation results assessed using Spearman's test showed modest correlation for semantic fields and modest/strong correlation for conceptual domains. We must not forget, however, that the conversion procedure adopted introduced quantitative approximations. In fact, in about 34% and 30% of the cases, for semantic fields and conceptual domains, respectively, the frequency of the USAS tags considered was equally (and not proportionally) distributed among two or more Codebook semantic fields, which obviously influenced Spearman's results.

Finally, the most frequent 150 USAS items in the semantic frequency list were compared to manual coding of the most frequent 300 words in the wordlist (see Chapter 7). At the level of semantic fields, manual tagging of the top 300 words in the wordlist provided better results than the procedure experimented in the current chapter, at both qualitative and quantitative levels. At the level of conceptual domains and semantic prosody, the two procedures seem comparable in terms of results at the qualitative level, but not at the quantitative one. At the level of conceptual domains, manual coding of the top items in the wordlist showed not only about 100% of highly conventionalised fields and of the cultural associations, but also strong/very strong correlations with the whole datasets. On the other hand, the top 150 items in the semantic frequency list recovered 100% of the highly conventionalised and cultural domains, but showed inconsistent correlation results (modest correlation for *wine* and strong for *chocolate*). Finally, the ASSESSMENT field is characterised in all the cases

under analysis by prevalence of positive vs. negative assessment, but the proportion between the two types of assessment is markedly different (4.2 times bigger in the *chocolate* manually coded top 300 words; 2.4 times bigger in the *wine* manually coded top 300 words; 2.3 times bigger in the *chocolate* top 150 USAS tags; and 6 times bigger in the *wine* top 150 USAS tags).

It seems clear from the current results that the approximations involved in the application of the conversion scheme have variably influenced the quantitative comparisons. It is noticeable, however, that, despite approximations, the most frequent 150 semantic categories were able to retrieve over 70% of the high conventionalisation fields/domains and cultural associations, with the already noticed 'improvement' in the number of semantic categories when passing from a more detailed coding scheme to a less detailed one.

Finally, comparison of the most frequent 150 USAS items in the semantic frequency list to manual coding of the most frequent 300 words in the wordlist suggests that, at least for small corpora, such as the elicited ones used in the current work, using an automatic semantic tagging tool is worth only if the tagging semantic categories can be used without further conversion. The case is likely to be different when using larger corpora. In fact, if we consider that both procedures are sensitive to corpus size, when working with very large corpora, the top N items in the semantic frequency list would be more representative of the overall corpus that the top N words in the frequency list.