Corpus analysis

Alessia Cadeddu

This analysis has been carried out on a corpus of dessert recipes taken from the Internet.

Total number of words in the text corpus: 5467

I have examined the first 100[1] most frequent words in the corpus, though, in some cases, I have taken into consideration items worth noting which are not included in the first one-hundred-word group.

First, I have carried out a **quantitative analysis** aiming at identifying:
- the first 100 most frequent words
- the most significant items
- the core vocabulary
- the minimal core vocabulary.

Secondly, I have tried to interpret the quantitative data by observing in what ways the target language is different from everyday language (**qualitative analysis**). Thus, I have focused on:
- collocations
- linguistic patterns
- language functions
- verbs (form, tenses, etc.)

Thirdly, I have drawn my general conclusions (from the previous investigations) on the target language exemplified in the corpus text.

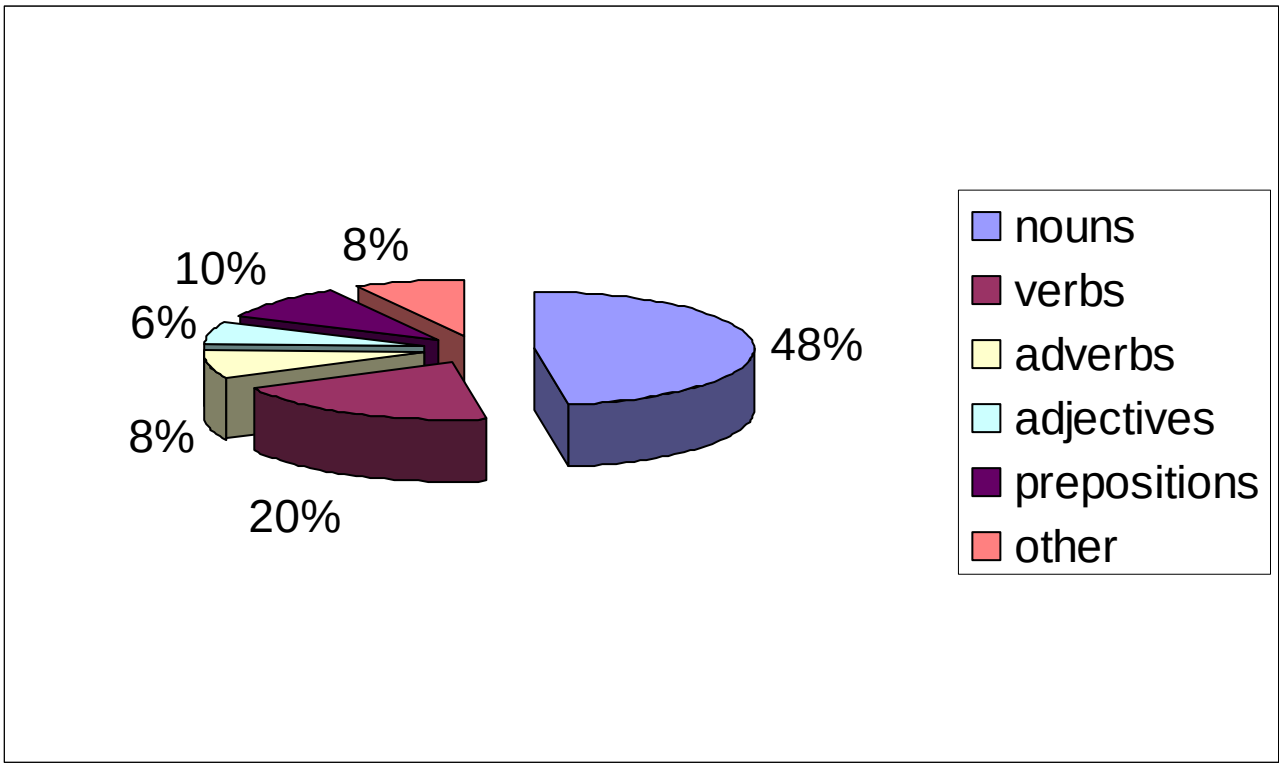
Finally, I have tried to foresee the implications of my investigations on a syllabus dealing with that particular target language and, partially, based on the corpus text.

[1] Actually I have decided to include also those words having the same number of occurrences (10) as the ninety-seventh one.
In addition, in the first one-hundred-word group, there are some words which, despite being spelt in different ways, refer to the same item (i.e. [50] oz, [54] ounces, [70] ozs; [35] t, [73] tablespoons, [108] tablespoon, and so on).

Quantitative analysis

- **The most significant items**

Here are the figures of the first 100 most frequent words in the whole corpus text divided into word classes and expressed as percentages.



This chart clearly shows that almost half of the first 100 most frequent words are **nouns (48%)**.
Another main word class is that of **verbs (20%).**
Then, there are **prepositions**, **adverbs** and **adjectives**.
Despite the large number of articles and conjunctions either in the first 100-word group or in the whole corpus, I have decided to leave them out of consideration and to include them in the category 'OTHER' because they don't represent a peculiar feature of the target variety of language.
Similarly, I have included pronouns in the category 'OTHER', because the only examples of pronouns in the first 100-word group are: *you* which occurs 13 times and *all* 10 times.

▪ Core vocabulary

Here are the lists of the first 100 most frequent words in the corpus divided into word classes. As already said, I have examined just the major word classes (see above) in my corpus text.

Nouns: 47%

1.	[7]	cake	96
2.	[8]	sugar	86
3.	[11]	cream	64
4.	[12]	butter	54
5.	[13]	c[ups]	49
6.	[15]	chocolate	45
	[16]	cup	45
7.	[18]	mixture	44
8.	[21]	flour	39
9.	[22]	minutes	38
10.	[23]	egg	37
11.	[27]	top	34
12.	[29]	oven	33
13.	[30]	pan	32
14.	[31]	bowl	31
15.	[33]	salt	27
	[34]	eggs	26
16.	[35]	t[ablespoons]	26
17.	[36]	water	26
18.	[39]	vanilla	25
19.	[41]	pie	24
20.	[52]	crust	18
21.	[53]	filling	18
	[54]	ounces	18
22.	[57]	whites	17
23.	[60]	inch	16
24.	[61]	layer	15

25.	[62]	rack	15
26.	[64]	teaspoon	15
27.	[66]	yolks	15
28.	[69]	ingredients	14
29.	[70]	ozs	14
30.	[71]	pastry	14
	[73]	tablespoons	14
31.	[74]	tin	14
32.	[77]	orange	13
33.	[78]	sides	13
34.	[81]	almonds	12
35.	[82]	cheese	12
36.	[83]	coffee	12
37.	[84]	lemon	12
38.	[86]	paper	12
39.	[87]	powder	12
40.	[88]	saucepan	12
41.	[91]	almond	11
42.	[93]	cocoa	11
43.	[98]	bottom	10
44.	[100]	cinnamon	10
45.	[103]	grams	10
46.	[104]	ground	10
	[108]	tablespoon	10
47.	[109]	time	10
	[110]	tsp	10

Verbs: 20%

1.	[19]	add	43
2.	[28]	beat	33
3.	[32]	mix	29
4.	[37]	bake	25
5.	[38]	is	25
	[43]	baking	23
6.	[51]	remove	19
7.	[55]	pour	18
8.	[59]	heat	16
9.	[63]	remaining	15
10	[68]	fold	14

11	[72]	spread	14
12	[76]	make	13
	[92]	be	11
13	[95]	stir	11
14	[96]	whisk	11
15	[99]	chopped	10
16	[101]	cover	10
17	[102]	cut	10
18	[105]	let	10
19	[106]	melted	10
20	[107]	put	10

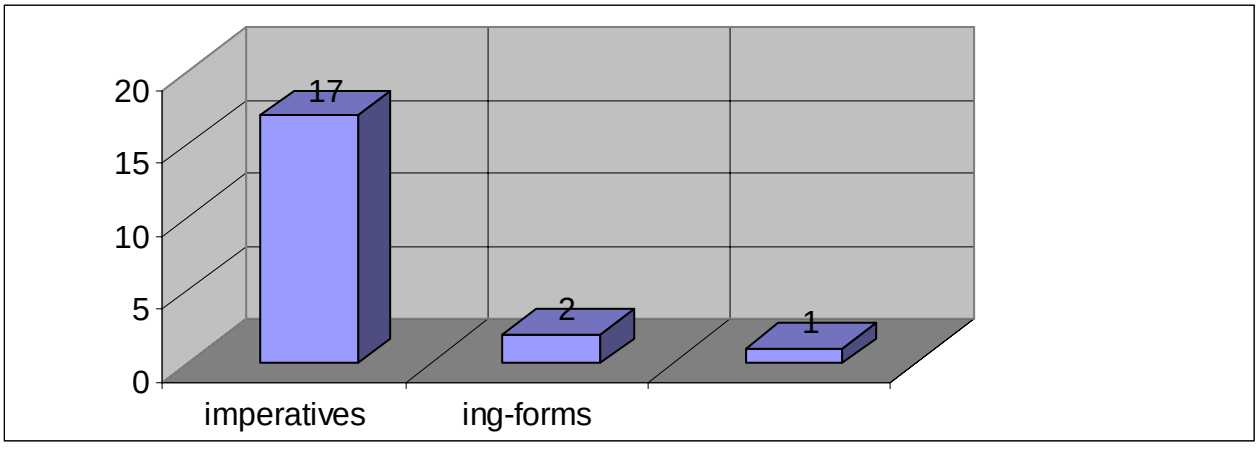
However, it is to be said that most of these words reoccur frequently as a different verbal tense in the whole corpus text. Here are some examples:

[162]	stirring	7
[212]	cook	5
[226]	mixed	5
[266]	makes	4
[269]	melt	4
[297]	baked	3
[301]	beating	3

[313]	covered	3
[338]	mixing	3
[534]	whisking	2
[540]	adding	1
[699]	mixes	1
[754]	putting	1
[809]	stirred	1

Thus, for example, the verb *to bake* occurs 48 times, the verb *to add* 44, the verb *to beat* 36, the verb *to mix* 35, and so on.

As for moods and tenses, the frequency of the **imperative** is striking: there are 17 imperatives out of 20 verbs.



Prepositions: 10%

1.	[4]	of	134
2.	[5]	in	109
3.	[6]	to	104
4.	[9]	with	84
5.	[14]	for	49
6.	[17]	on	45
7.	[20]	into	41
8.	[25]	over	37
9.	[42]	at	23
10.	[47]	from	19

Adverbs: 8%

1.	[45]	about	20
2.	[49]	out	19
3.	[56]	together	18
4.	[58]	before	16
5.	[65]	well	15
6.	[75]	completely	13
7.	[89]	then	12
8.	[90]	when	12

Adjectives: 6%

1.	[40]	cool	24
2.	[44]	small	22
3.	[46]	white	20
4.	[48]	large	19
5.	[79]	smooth	13
6.	[94]	cold	11

▪ Minimal core vocabulary

At this point it is possible to identify the minimal core vocabulary, that is, the essential items in the target specialised language.

Nouns

Nouns referring to food		Nouns referring to preparation	Nouns referring to utensils	Nouns referring to measurement
cake	eggs	mixture	oven	minutes
sugar	yolks	pie	pan	ounces
cream	almond	crust	bowl	inch
butter	cocoa	filling	tin	tablespoon
chocolate	coffee	layer	saucepan	teaspoon
flour	lemon	ingredients	cinnamon	grams
egg	orange	pastry		
water	powder	top		
vanilla	salt	sides		
whites				

Verbs

add
beat
mix
bake
remove
pour
heat
fold
spread
make

be
stir
whisk
chop
cover
cut
let
melt
put

Qualitative analysis

▪ Collocations

The most striking feature in the target variety of language is clearly the co-occurrence of technical words. Associations are mostly made contiguously: VERB + NOUN, ADJECTIVE + NOUN, VERB + ADVERB. Frequent collocations are:

'beat the eggs'	*'melted butter'*	*'mix well'*
'bake at … degrees'	*'melted chocolate'*	*'cool completely'*
'bake for … minutes / hours'	*'brown sugar'*	
'remove from / oven / pan / the heat'	*'electric mixer'*	*'finely chopped'*
	'white chocolate'	
'pour into a pan'	*'large eggs'*	
'whisk the eggs'	*'heavy cream'*	
'put the cake into the oven'		
'let the pie cool'		

▪ Linguistic patterns

The need for concision and economy of expression frequently make it necessary the use of:

- **Ellipsis of articles**

'*combine flour, baking powder, soda and salt*'
'*beat egg yolks, castor sugar and cream*'
'*slice bananas into large bowl. Add remaining ingredients*'

- **Ellipsis of parts of the utterance** as in the construction VERB + UNTIL + ADJECTIVE

'*mix until combined*' instead of 'mix until it is combined'
'*stir until smooth*'
'*whisk until creamy*'

- **Fronting of adverbs of manner**

'gradually add sugar' instead of 'add sugar gradually'
'gently fold the egg whites'
'thinly slice the third banana'

▪ Language functions

The sequencing of the technical process of food preparation is expressed through:

- **Juxtaposition of imperative forms**

*'**Pour** butter into …. **Sprinkle** coconut …. **Bake** at ….'*
*'**Mix** flour, …. **Add** grated orange …. **Add** grated cooking apple. **Mix** well. **Add** eggs, …. **Add** sugar, …. **Mix** well.'*
*'**Add** the Mascarpone and beat for …. **Set** aside. **Beat** the egg whites…'*

- **Listing**

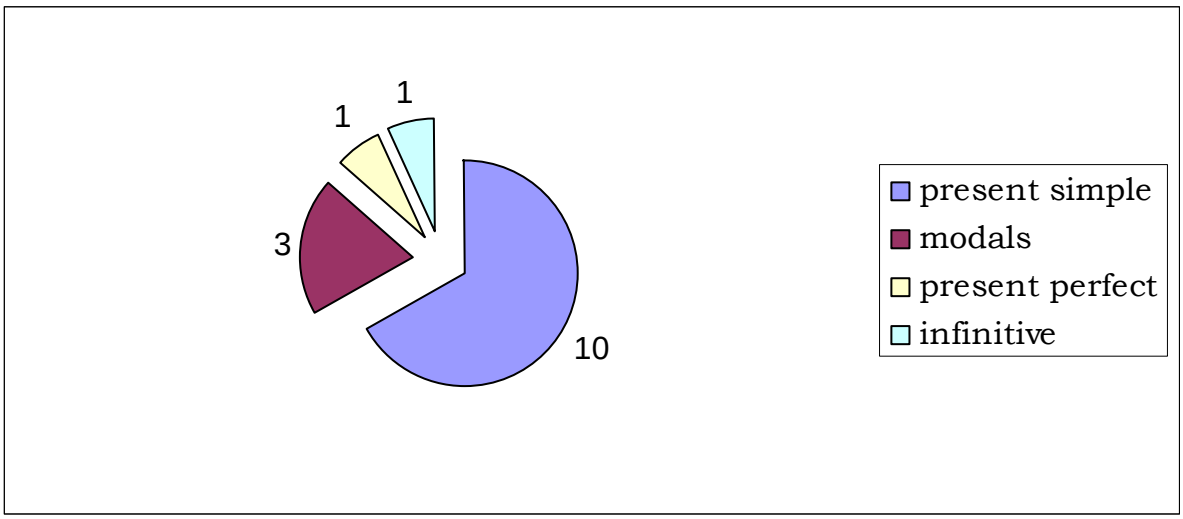
1 Heat oven to ….
2 Peel, core and slice the apples …
3 Mix sugar, flour ….

- **Use of adverbs of time**

*'**When** the cake is baked, remove from …. **Then** carefully turn …'*
*'Whisk egg yolks, **then** add lemon juice'*
*'**Meanwhile**, combine …'*

▪ Passive forms

Passive forms are almost rare in the whole corpus text. Here are the figures of all the occurrences per tenses:



▪ Verbal tenses

As already said (see chart page 5), the most frequent verbal tense in the corpus text is the **imperative**[2]. This is clearly due to the fact that recipes are nothing but sets of instructions for preparing food dishes. Among the other verbal tenses occurring in the corpus are: -ing forms and past participles.

Ing-forms have different functions throughout the corpus:

- Sometimes, they are nominalized in order to summarize ideas that would be, otherwise, too long to express.

'***baking***' instead of 'how to bake the cake'
*'… before final **serving**'* instead of 'before the cake is finally served'
*'while **preparing** …'*

- Sometimes, they are mere qualifying adjectives:

*'**baking** powder / oven'*
*'**cooking** spray'*
*'a **serving** plate'*

- Sometimes, they describe the technical process:

*'**beating** well …'*
*'**stirring** occasionally'*
*'**whisking** constantly …'*

Past participles generally have a qualifying function:

*'**chopped** butter'*
*'**melted** chocolate'*
*'**whipped** cream'*
*'**beaten** eggs'*
*'**granulated** sugar'*
*'**softened** butter'*

▪ Format of a recipe

[2] This is the reason why personal pronouns are very rare in the corpus text.

Each recipe follows a **typical format**:

- **Title**
- **List of ingredients and quantities**[3]
- **Description of the technical process**
- **Baking instructions**
- **Suggestions about serving**

Conclusions

According to the previous investigations on the corpus text, the most important features of the target specialised language are:

- Very high frequency of nouns related to food, food preparation, kitchen utensils and systems of measurement.
- High frequency of technical verbs related to food preparation.
- Habitual collocations
- Recognizable linguistic patterns (ellipsis and fronting) aimed at concision
- Use of verbal juxtaposition, listing and time adverbs to describe the sequences of the technical process of food preparation
- Very high frequency of the imperative
- Recognizable textual format.

This outcome would lead me to design a syllabus which would include the following objectives:

- Enlarging and/or acquiring everyday and technical vocabulary
- Talking about a process
- Identifying the typical structure of a recipe
- Writing for concision
- The use of the imperative
- The use of adverbs

[3] In some recipes quantities are expressed according to the GB and US systems of measurement, in others according to the European systems.