Building a specialised corpus in Turkish

Gülsüm Atasoy*

1. Introduction

Corpora are designed to investigate a given language as a whole and answer specific research questions (Hunston 2008: 154). In designing corpus, it is important to plan and consider on the size, text types, population, domain (the subject matter of the text) and medium (e.g. book, periodicals, written to be spoken) of the corpus (Meyer 2002: 30; Hunston 2008: 155).

As Hunston (2008: 155) mentions it, "...how a corpus is designed depends on what kind of corpus it is and how it is going to be used". If the research is on sociolinguistics, the variables such as age, sex, region gain importance. For example, COLT corpus (Corpus of London Teenage Corpus) is such a corpus, which is limited by the teenage language; moreover, ICLE corpus (The International Corpus of Learner English) is a corpus, which consist of expository essays written in English by university students who are learners of English. That is to say, if the corpus intended to facilitate research on a single register, then the corpus should contain texts representing that register (Hunston 2008: 156).

It is not always possible to build the desired, planned corpus as there can be practical constraints on corpus building such as, software limitations, copyright, ethical issues and the text availability (Hunston 2008: 156–157). In collecting the written texts, there are three methods. These are keying (writing by hand), scanning and obtaining texts electronically.

Moreover, three issues should be taken into account when designing a corpus. These are representativeness, balance, and size. Representativeness is the relationship between the corpus and the body of language it is being used to represent. In order to be representative in corpus, the compiler should use equally sized samples and also view the texts as having beginnings, middles and ends (Baker 2006: 27). Balance refers to the consistency in proportions of the texts in a corpus. Size is related with representativeness (McEnery et al. 2006: 13; Hunston 2008: 160).

If a range of topics is to be included in the corpus, it must be of a sufficient size to allow this. In order to achieve representativeness, a corpus should include texts from different categories of writing and speech. The categories should include (Hunston 2008: 161; Baker 2006: 27):

- topic areas (books and magazines on various subjects),
- modes of publication (books, newspapers, leaflets),
- social situation (casual conversation, interviews, lessons), (spoken corpus criterion)
- interactivity (monologue, dialogue, multi-party conversation), (spoken corpus criterion).

* Mersin University.

For example, a specialised corpus dealing with telephone calls with an operator service should be balanced by including a variety types of operator conversations so that it can be representative and the size should be arranged considering how many conversations to be included (McEnery et al. 2006: 15–16).

2. Methodology

2.1. The purpose of the study

Our main purpose is to design and build a specialised Magazine Texts Corpus, which covers the years 1990–2009. We will sort the data acquired from our corpus and analyse the frequency distribution of the discourse connector *ama* in terms of semantic, syntactic and pragmatic features.

2.2. Data collection technique

We have taken our texts from the databases of TNC (Turkish National Corpus Project*), which is under construction. These texts are all computerised and available in a usable form. In designing the Magazine Texts Corpus, we have paid attention to the corpus in order to be balanced and large enough to be representative. This is summarised in the grid below:

Table 1. Design Features of the Turkish Magazine Texts Corpus

	Magazines	Aksiyon	Birikim	Gonca	Şebnem
Represent ativeness	Subject Matter/ Topic Areas	technology, economy, politics, cinema, shopping, sports, etc.	socialism	children and family	religion and belief
	Medium	Periodical			
Balance	Per Magazine	20000 words			
Size		80000 word corpus			

We analyse our data by the help of the software NooJ, which is a corpus processor that can launch sophisticated queries over large corpora in order to produce various results (concordances, statistical analyses, information extraction, etc.).

* Acknowledgment to Scientific and Technological Research Council of Turkiye (TÜBİTAK 108K242).

2.3. Method of analysis

The data is analysed in the light of the methodology of the corpus linguistics by the help of the software, NooJ. We follow results of the study of Ruhi (1998) on the semantic and syntactic features and the study of Sekali (2007) on the pragmatic features of the discourse connector *ama*, then we search for the frequency distributions of the attained results through the constructed corpus via NooJ.

2.4. The limitation of the research

"Newspapers are not homogeneous. In addition, a whole year of any one particular newspaper is not a sample but the whole population of possible texts from that newspaper and particular year" (Hundt 2008: 179), we claim that magazines are not also homogeneous, they contain text categories and represent the particular years. Moreover, as Hunston (2008: 156–157) mentions that there can be practical constraints on corpus building, we have faced with some of these constraints. We do not have every issue of the given magazines between the years 1990–2009. The only available years of the magazines are shown in the table below.

Table 2. The Magazines and Available Date Distributions

Magazines	Aksiyon	Birikim	Gonca	Şebnem
Available date distributions	1995–1999	1991–2008	2002–2009	2002–2009

The corpus designer should keep in mind that "The compiler of a corpus should be willing to change his initial corpus design if the circumstances arise requiring such changes to be made" (Meyer 2002: 32) and may confront constraints. In building Turkish Magazine Texts Corpus (TMTC), we have confronted with two constraints. Initially, we have planned to build a 100.000 word Magazine Texts Corpus and decided on 5 different kinds of magazines to include in the corpus. However, we had to discard one of the magazine from the corpus as we did not have the publication of the magazine even though we had the texts. Hence, we changed our corpus design and build a 80.000 word corpus from 4 different kinds of magazines. The other constraint is the year inconsistency of the magazines. Not all the magazines are published through the years 1990–2009. Unfortunately, we can not have the text availability and the year consistency under these circumstances. As we study on function words, it would not change the results.

We consider that "a balanced corpus would consist of the same amount of text from each newspapers concerned" (Hunston 2008: 163). In designing TMTC, we have made a mathematical calculation in order to provide representativeness and balance. That is to say, we plan to build a 80.000 word corpus to be representative enough and we have 4 magazines available. Then we should take 20.000 words from each magazine so that we would attain to 80.000 words. We have built a 80.098 word corpus from magazines, which has provided us with 101449 tokens. The grid and the sampling frames are given as below:

Table 3. The Corpus Design Grid of TMTC

Magazine	Available date distribution	Approximate number of words to be taken per year	Sum
Aksiyon	1995–1999 (5 years)	4000	20000
Birikim	1991–2008 (18 years)	1111,11	20000
Gonca	2002–2009 (8 years)	2500	20000
Şebnem	2002–2009 (8 years)	2500	20000

3. Analysis of the data

This section discusses the frequency distribution of the discourse connector *ama* in terms of its semantic, syntactic and pragmatic features attested in TMTC.

3.1. Semantic features of *ama*

All the studies on the discourse connector *ama* in Turkish prove that *ama* has two semantic functions (Altunay 2007:174; Altunkaya 1987: 106; Doğan 1994: 201–204; Göksel and Kerslake 2005: 519; Halliday and Hasan 1976: 237, 250–255; Ruhi 1998: 139)

1. *ama* negates the expectation created by the first clause of an utterance.
2. *ama* signals the contradiction between the first and the second clause of an utterance.

In our study, considering these two semantic functions of *ama*, we have focused on the frequency information of *ama*. In the 80.098 word Turkish Magazine Texts Corpus, we have totally 192 utterances containing the discourse connector *ama*. 99 of them signal conflict; that is, negates the expectation created by the first clause while 93 of the utterances signal the contradiction between the first and the second clause of an utterance. To give examples for negating the expectation from TMTC:

(1) *Heyecanla yatağa girdim, ama çok zor uyuyabildim.*
'In excitement, I went to bed, *but* I could hardly sleep.'

(2) *Çoğu kralı yüksek sesle eleştirmiş. Halkından bu kadar vergi alıyor ama yolları temiz tutamıyor, demiş.*
'Many of them criticised the king loudly. He takes that much tax *but* cannot keep roads clean, he said.'

In the example (1), we see that the writer says he went to bed in excitement and afterwards it is expected something parallel. However, he finishes his sentence with negating the expectation with the sentence '*but* I could hardly sleep'. In the example (2), 'the king takes that much tax *but* can not keep roads clean', which negates the expectation intro-duced in the first sentence. In the following examples, we see *ama* signalling contradiction.

(3) *Bu kaç keredir oluyor, ama ben hala alışamadım.*
'This happens many times, *but* I haven't still been able to get used to it.'

(4) *Sandığından eski ama kullanılmamış çatal-kaşıkları ve mis gibi elma kokulu temiz bir havlu çıkardı.*
'From her chest, she took out old *but* not used forks and spoons and a clean towel smelling like a fresh apple.'

In the example (3) it says that 'this happens many times, but I haven't still been able to get used to it'. We see the contradiction of happening many times and not able to get used to it. Likewise, another contradiction appears in the example (4) 'old but not used forks and spoons'.

3.2. Syntactic features of *ama*

In addition to the semantics of *ama*, negating the expectation and signalling the contradiction between the two clauses, we also see that the discourse connector *ama* has syntactic characteristics which may appear in clause inital position, clause medial position and clause final position.

On this topic, Ruhi (1998: 141) remarks that *ama* semantically marks conflict and adversative relations external to the ongoing topic. We come across with this usage of *ama* in our corpus in the clause final position, as in the examples given below:

(5) *Buna "Star strateji Türkiye'de de tuttu" denir mi bilinmez ama, son yıllarda Türkiye'nin starları da reklamlarda boy göstermeye başladı.*
'It is not known whether the star strategy have worked in Turkey *but*, lately the stars of Turkey have started to appear in advertisements.'

(6) *Panço reklamları pek öyle ahım şahım reklamlar değil ama, bütün Tarkan hayranları Doritos Panço yiyorlar.*
'The Panço ads are not favourable advertisements *but*, all the fans of Tarkan eat Doritos Panço.'

Out of 192 concordance lines consisting of *ama* in TMTC, we have *ama* in clause final position in 13 lines. Interestingly, *ama* in all 13 lines occurs in clause final position; all ending with comma, which give way to the following clause. In other words, we do not have any lines ending with period in our corpus data. We can comment that this can be a coincidence in occurance of *ama* with comma in clause final position, which may result from the random choice of the data.

We also see the discourse connector *ama* in clause medial position, that is to say, *ama* occurs in between two clauses in 89 concordance lines in TMTC. To illustrate some of these lines:

(7) *Televizyonlardan belki para alabilirler ama yazılı basından para alabileceklerini pek sanmıyorum.*
'They may take money from Televisions *but* I don't think that they would be able to get money from the written press.'

(8) *Şimdi deniliyor ki; bütün bu fenalıklar olmasın, hepimiz aleyhindeyiz ama bunu önleyecek kanun yapmayın. Bunun manasını ben anlayamadım doğrusu.*
'Now it is said that all these evils may not happen, we are all against *but* don't make law preventing this. I couldn't understand its meaning actually.'

The most frequent occurance of *ama* is in the clause inital position with the 90 concordance lines. Even though there is not a big difference from the clause medial position, the corpus data reveals these quantitive results. To give examples:

(9) *Sizin ramazanınız nasıl geçti bilmiyorum, ama benimki harikaydı.*
'I don't know how was your ramadan, *but* mine was great.'

(10) *Ne partim adına ne de devlet adına size söz verebilirim. Ama şahşım adına çalışacağıma söz veriyorum.*
'I can promise on behalf of neither my party nor the state. *But* I proimse to work on my own behalf.'

We can conclude that out of 192 concordance lines, *ama* occurs in 13 concordance lines in the clause final position, in 89 concordance lines in clause medial position and in 90 concordance lines in clause inital position.

3.3. Pragmatic features of *ama*

To analyse diversified pragmatic usage of *ama*, we have focused on the study of Sekali (2007). She points out that there are four pragmatic values of the coordinator *but*, which are stated as below.

3.3.1. but: The linguistic construction of an intermediary representation

Sekali states "indirect meanings or implicatures are not encoded in the utterances prior to their connection, *but* are linguistically constructed through the association of the enunciative operations marked by *but*" (2007: 157). We came across with this usage of *but* mostly in dialogues. As we do not have dialogues in our data, we have found out that *but* as an intermediary respresentation is not used in TMTC.

3.3.2. but: The inner structure of P and Q on the retrieval of the implicit utterance

Sekali claims that the main forms structuring the implicit utterance i: A form in which Q takes up the grammatical structure of P with a change of one of its lexical entries or with different modality (Q=P'). In this structuring, *but* introduces stronger argument, versus status and refutation.

In TMTC, we see *ama* as the implicit utterance in 96 concordance lines out of 192. This is the most frequent occurance of *ama* in terms of pragmatic value in our data. To show some examples from the corpus:

(11) *1639'dan sonra Türkiye ile İran arasında bir daha savaş çıkmadı ama gerçek bir dostluk da kurulamadı.*
'After 1639, there is no war between Turkey and Iran *but* a real peace isn't able to be established, too.'

(12) *Konutların reklamlarında hep aynı iki özelliğinin vurgulandığını görüyoruz. Hepsi, "İstanbul'un dışında" ama "çok yakın"; otoyol üzerinden otomobille 'birkaç dakikada ulaşılabilecek' mesafede.*
'In the ads of houses, we see that always the same two features are mentioned. All of them are out of İstanbul *but* very close to İstanbul; on the road by car it is on the distance of reach in few minutes.'

Ama introduces refutation to the first statements. In (11), there is no war between Turkey and Iran after 1639, in the following it is expected to be peace. However, the following statement refutes the first statement by saying '*but* a real peace is not able to be established'. In the example (12), *ama* introduces the stronger argument in 'The houses are out of İstanbul but very close to İstanbul'. Here *ama* helps to consider that being out of İstanbul does not mean the houses are away from İstanbul.

3.3.3. but: The notion of argumentative force

By the help of *but*, the speaker sets himself in the command of the discourse and takes the control of its progression and thematic direction (Sekali 2007: 166). Sekali argues that in the argumentative force, the speaker using *but* can both break the co-speaker's direction, imposing his own and make emphasis on the topic.

There are 47 concordance lines concerning *ama* as the notion of argumentative force in TMTC.

(13) *Mahsusa'nın faaliyetleri, Yakup Cemil'in idamı vb. macera romanları gibi dedim, ama doğrusunu isterseniz, hiç bir roman, Devlet-i Aliyye'nin son yirmi yılı kadar heyecan verici ve dramatik değildir.*
'I said the adventure novels like the activities of Mahsusa, the suicide of Yakup Cemil, etc., *but to tell the truth no other novel is as much exciting and dramatic as* the last twenty years of Devlet-i Alliye. '

(14) *Filmde eleştirilebilecek pek çok şey var. Herşeyden önce – Mustafa bu lafa köpürüyor ama İstanbul Kanatlarımın Altında bir sanat filmi değil; aksine oldukça popülist.*
'There are lots of things to be criticised in the movie. First of all – Mustafa gets angry at this word *but* İstanbul Kanatlarımın Altında is not an art movie; on the contrary, it is quite populist.'

When we look at these examples, in (13), the speaker breaks the topic by *ama* and introduces his own argument stating 'No novel is as much exciting and dramatic as Devlet-i Alliye's'. The example (14) is a clear example for making emphasis; to put it this way, in this example the speaker both breaks the topic by *ama*, imposes his own and then makes emphasis on his point mentioning 'the film is a populist one'.

3.3.4. but: The notion of explanation and condition

Explanation and condition can be retrieved in *but* compound-utterances. As Sekali explains, "...the connective *but* will interact with other linguistic operations within the connected utterances in the complex process of meaning construction" (2007: 172).

In TMTC, we have 56 concordance lines of *ama* as the notion of explanation, we do not have *ama* signaling condition. To designate examples from the Corpus:

(15) *Bu üç grup düşünce sahibinin nüfusumuza göre kesin oranlarını belirlemek güç olabilir ama, 1. gruptakilerin ezici çoğunlukta olduğu aşikardır; Türkiye'yi müslüman bir ülke yapan da.*
'It can be difficult to determine the exact ratios of these three groups of idea owners according to our population but, it is obvious that the first group is the majority; what makes Turkey muslim.'

(16) *Sonia'nın becerileri ve ayrıca kaygısının keskinliğiyle açıklanabilecek bir mucizeydi; ama toplantının amacı da, bir "mucize duası" gibi bir şeydi.*
'It was a miracle that could be explained with the skills of Sonia and also pungency of her worry but the purpose of the meeting is like "a miracle pray".'

All these examples illustrate that the clause following the connector *ama* makes explanation on the topic. In other words, in the pragmatic value the discourse connector *ama* is used to present explanation on the previously mentioned topic in the first clause.

3.4. Comparison and Contrast of ama Between the Years 1990s and 2000s

Referring to the Table 2, the magazines and available date distributions, the discourse connector *ama* can also be analysed as *ama* in 1990s magazines and *ama* in 2000s magazines because of the nature of TMTC. The magazines representing 1990s are Aksiyon and Birikim while the magazines representing 2000s are Gonca and Şebnem, which are still representative, balanced and large enough to be compared and contrasted in one another. Here is the table summarising the analysis:

Table 4. Comparison and Contrast of ama Between the Years 1990s and 2000s

		Aksiyon-Birikim	Gonca-Şebnem
Semantic features	negating the expectation	56	43
	signalling contradiction	36	57
Sum		92	100
Syntactic features	clause final position	10	3
	clause medial position	46	43
	clause inital position	36	54
Sum		92	100
Pragmatic features	*ama*: the linguistic construction of an intermediary representation	0	0
	ama: The inner structure of P and Q on the retrieval of the implicit utterance	45	51
	ama: The notion of argumentative force	28	19
	ama: The notion of explanation	19	30
Sum		92	100

In Table 4, we see that from 1990s, that is, from Aksiyon and Birikim we have 92 discourse connector *ama* in total while from 2000s, that is, from Gonca and Şebnem we have 100 discourse connector *ama* in total. In 2000s in the semantic use of *ama*, signalling contradiction is used more than negating the expectation. In the syntactic use of *ama*, we come across with a fall in the clause final position use and a rise in the clause inital position use while there is not any significant change in the clause medial position use. In terms of pragmatic features *ama* is more preferred in the use of implicit utterances and in the notion of explanation in 2000s while *ama* is more preferred as the notion of argumentative force in 1990s.

4. Conclusion

In this study, our basic aim is to show the steps in building a specialised corpus in Turkish, and then analyse the frequency distribution of the discourse connector *ama* in terms of its semantic, syntactic and pragmatic features through the constructed corpus. Summarising the discussions above, we can draw the table below showing the features of the discourse connector *ama* in Turkish:

Table 5. The Features and Frequency Information of the Discourse Connector *ama* in Turkish

		Semantic Features	Syntactic Features	Pragmatic Features
ama		negating the expectation	clause final position	*ama*: the linguistic construction of an intermediary representation
	Frequency out of 192 lines	99=51.56%	13=6.77%	0
		signalling contradiction	clause medial position	*ama*: the inner structure of P and Q on the retrieval of the implicit utterance
	Frequency out of 192 lines	93=48.43%	89=46.35%	96=50%
			clause inital position	*ama*: the notion of argumentative force
	Frequency out of 192 lines		90=46.87%	47=24.47%
				ama: the notion of explanation
	Frequency out of 192 lines			49=25.52%

In the semantic analysis of the discourse connector *ama* in terms of frequency, the negating feature exceeds the signalling contradiction feature in between the propositions in the Turkish Magazine Texts Corpus. In the syntactic analysis of the discourse connector *ama* in terms of frequency, the occurance of clause final position is the least frequent while the occurance of clause inital position is the most frequent in TMTC. In the pragmatic analysis of the discourse connector *ama*, in terms of frequency the least occurred pragmatic value is the argumentative force while the most ocurred one is the inner structure of P and Q on the retrieval of the implicit utterance.

Comparison and contrast of *ama* between the years 1990s and 2000s

2000s plays great role in the feature of signalling contradiction in the total number. Likewise in the syntactic features 1990s has impact on the total number with the clause final position use while 2000s has impact on the total number with the clause inital position use. In terms of pragmatic features, 1990s makes the difference by the notion of argumentative force; on the other hand, 2000s makes the difference by the notion of explanation.

References

Altunay, D. 2007. Neden-etki ilişkisi bağlaçları ve metindeki bağdaşıklık. In: Aksan, Y. & Aksan, M. (eds.) *XXI. Ulusal dilbilim kurultayı bildirileri.* Mersin: Mersin Üniversitesi. 172–179.

Altunkaya, F. 1987. *Cohesion in Turkish: A survey of cohesive devices in prose literature.* [Unpublished PhD. Dissertation, Ankara: Hacettepe University.]

Baker, P. 2006. *Using corpora in discourse analysis.* London-New York: Continuum.

Doğan, G. 1994. *ama* bağlacına edimbilimsel bir bakış. *Dilbilim Araştırmaları* 1994, 195–205.

Göksel, A. & Kerslake, C. 2005. *Turkish: a comphresensive grammar.* London-New York: Routledge.

Halliday, M. A. K. & Hasan, R. 1976. *Cohesion in English.* London: Longman.

Hundt, M. 2008. Text corpora. In: Lüdeling, A. & Kytö, M. (eds.) *Corpus Linguistics: An international Handbook.* Berlin: Mouton de Gruyter. 168–186.

Hunston, S. 2008. Collection strategies and design decisions. In: Lüdeling, A. & Kytö, M. (eds.) *Corpus Linguistics: An international Handbook.* Berlin: Mouton de Gruyter. 154–168.

McEnery, T. & Xiao, R. & Tono, Y. 2006. *Corpus-based language studies.* Oxon: Routledge.

Meyer, C. F. 2002. *English corpus linguistics: An introduction.* Cambridge: Cambridge University Press.

NooJ. *www.nooj4nlp.net*

Sekali, M. 2007. He's a cop *but* he isn't a bastard: An enunciative approach to some pragmatic effects of the coordinator *but.* In: Celle, A. & Huart, R. (eds.) *Connectives as discourse landmarks.* Amsterdam: Benjamins. 155–175.

Ruhi, Ş. 1998. Restrictions on the interchangeability of discourse connectives: A study on *ama* and *fakat.* In: Johanson, L. et al. (ed.) *The Mainz meeting: Proceedings of the seventh international conference on Turkish linguistics,* August 3–6, 1994. Wiesbaden: Harrassowitz. 135–153.

Turkish National Corpus: *www.tnc.org.tr*