The Turkish National Corpus (TNC): Comparing the Architectures of v1 and v2

Yeşim Aksan
Mersin University
Mersin, Turkey
yesimaksan@gmail.com

Selma Ayşe Özel
Çukurova University
Adana, Turkey
saozel@gmail.com

Hakan Yılmazer
Çukurova University
Adana, Turkey
yilmazerhakan@gmail.com

Umut Ufuk Demirhan
Mersin University
Mersin, Turkey
umutufuk@gmail.com

Abstract— Turkish National Corpus (TNC) released its first version in 2012 is the first large scale (50 million words), web-based and publicly-available free resource of contemporary Turkish. It is designed to be a well-balanced and representative reference corpus for Turkish. With 48 million words coming from the written part of it, the untagged TNC v1 represents 4438 different data sources over 9 domains and 34 different genres. The morphologically annotated, 50 million words TNC v2 with 5412 different documents compiled from written and spoken Turkish is planned for release in 2016 offers new query options for linguistic analyses. This paper aims to compare architectures of the TNC v1 and v2 on the basis of a set of queries made on both versions. Standard, restricted and wildcard lexical searches are performed. Then, the speed of two versions in retrieving the query results in concordance lines is compared. Finally, it is argued that TNC v2 performs better and faster than that of TNC v1 due to the in-memory inverted index structure. Since building language corpora is a very recent issue for Turkish, the architecture of TNC v2 would serve as a model for similar corpus construction projects.

Keywords—Turkish National Corpus (TNC); corpus building; architecture; inverted index; relational database; in-memory data structures

I. INTRODUCTION

There are at least two different kinds of corpora in Turkish today: (i) large-sized general linguistic corpora that are constructed and made available for users with proper corpus tools, (ii) NLP corpora built with no linguistic criteria in mind but rather as tools for testing algorithms devised for different applications [1]. The first electronic linguistic corpus designed to represent modern Turkish is the 2 million words, downloadable Middle East Technical University Turkish Corpus (MTC) [2]. MTC is tagged by XCES style annotation using special software developed by the members of the project group as well as its corpus query workbench. In the years following the construction of the MTC, the need for a large-scale general reference corpus of Turkish has become more and more obvious. To meet the challenge, Turkish National Corpus (TNC) is built as reference corpus of Turkish. The project team followed the best practices at all stages of corpus development. Major design principles were adopted from the experiences of the British National Corpus with minor modifications. The end product is the TNC, a well-balanced, representative, and large-scale (50 million words) free resource of a general-purpose corpus of contemporary Turkish [3].

As maintained by [14] “if the corpus in question claims to be general in nature, then it will be typically balanced with regard to genres, domains that typically represent the language under consideration”. In line with this definition, the major aim in building the TNC is to represent texts from different genres, domains and types in a balanced manner so that the conclusions drawn from quantitative and qualitative analysis of corpus data hold true for language use in general. Genre balance is an important aspect of corpus design [15]. Both versions of the TNC have data from different domains and genres set them apart from text archives or a collection of texts difficult to categorize and separate by genre, such as the Web. The number of linguistic and computational linguistic studies using the TNC as a reference corpus is increasing. While most of the linguistic and NLP studies use TNC for compiling naturally occurring language evidence and for hypothesis-testing [16, 17, 18, 19], there are still others following a corpus-driven approach and attempt to build hypotheses and describe Turkish on the basis of the TNC [20, 21]. Overall, the usefulness of the TNC as a general corpus primarily is due to the data itself. With 48 million words, the TNC v1 represents written component of the corpus which contains 4438 different data sources over 9 domains and 34 different genres, and was published as a free resource for non-commercial use in October 2012. Size of the TNC v2 is 50,997,016 running words, representing a wide range of text categories spanning a period of 23 years (1990-2013). It consists of samples from textual data representing 9 different domains (98%) with 4,978 documents and transcribed spoken data (2%) with 434 documents. The morphologically annotated, complete version of the TNC v2 is planned for release in 2016, offering new query options for linguistic analyses.

This paper is organized as follows: Section two explains the design features of the TNC. Section three describes basic features of the TNC interface. The architectures of the TNC v1 and v2 are presented in section four. Section five displays the comparative query results obtained through the two versions of the corpus. The paper finally argues that in-memory inverted index structure and relational database structure are effective in terms of speed and extension of web-based language corpora.

II. Design of The TNC

The only Turkish corpus of its kind, the TNC is constructed following the principles used to construct the British National Corpus in its basic design and implementation. The distribution of samples in written component of the corpus is determined proportionally for each text domain, time, and medium. Table I and II show the distribution of texts across domain and medium, respectively.

TABLE I. The distribution of texts across domains in the TNC

Domain	No. of words	% of words	No. of documents	% of documents
Imaginative: Prose	9,365,775	18.74 %	674	13.54 %
Informative: Natural and pure sciences	1,367,213	2.74 %	253	5.08 %
Informative: Applied science	3,464,557	6.93 %	461	9.26 %
Informative: Social science	7,151,622	14.31 %	671	13.48 %
Informative: World affairs	9,840,241	19.69 %	757	15.21 %
Informative: Commerce and finance	4,513,233	9.03 %	429	8.62 %
Informative: Arts	3,659,025	7.32 %	347	6.97 %
Informative: Belief and thought	2,200,019	4.4 %	297	5.97 %
Informative: Leisure	8,421,603	16.85 %	1,089	21.88 %
Total	49,983,288	100.00 %	4,978	100.00 %

TABLE II. The distribution of texts across mediums in the TNC

Medium	No. of words	% of words	No. of documents	% of documents
Unspecified	10,541	0.02 %	1	0.02 %
Book	31,456,426	62.93 %	2,141	43.01 %
Periodical	15,968,240	31.95 %	2,092	42.02 %
Miscellaneous: published	958,999	1.92 %	294	5.91 %
Miscellaneous: unpublished	1,589,082	3.18 %	450	9.04 %
Total	49,983,288	100.00 %	4,978	100.00 %

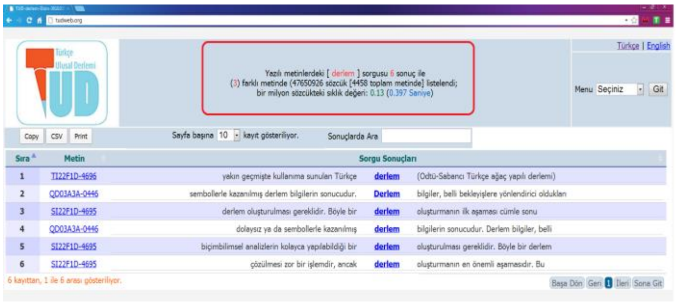
The representativeness of the TNC is secured through balance and sampling of varieties of contemporary language use. The selection of written texts is done via the criteria of text domain, medium, and time. The criterion of domain means that texts are distributed along two major types, namely imaginative and informative. While the imaginative domain is represented by texts of fiction, the informative domain is represented by texts from the social sciences, the arts, commerce-finance, belief-thought, world affairs, applied sciences, natural-pure sciences, and leisure. The criterion of medium refers to text production. The texts collected to represent the written medium are carefully selected from books, periodicals, published or unpublished documents, and texts written-to-be-spoken such as news broadcasts and screenplays, among others. The criterion of time defines the period of text production. Here, the distribution of the size of the texts for each year is decided in terms of relative representation of each domain in the medium.

Transcriptions from authentic spoken language constitute 2% of the TNC's database, which involve everyday conversations recorded in informal settings such as conversations among friends, talk among family members and friends, etc., as well as speeches collected in particular communicative settings, such as meetings, lectures, and interviews. The spoken component of the TNC contains a total of 1,013,728 running words. Of these words, 439,461 of them come from orthographic transcriptions of everyday conversations and their relevant medium, and 574,267 of them are orthographic transcriptions of context-governed speeches.

Part-of-speech annotation, morphological tagging, and lemmatization of the TNC are done by developing a natural language-processing (NLP) dictionary based on the NooJ_TR module [13]. The unique, semi-automatic process of developing the NLP dictionary includes the following steps: (i) automatically annotating the type list with the NooJ_TR module, which follows a root-driven, non-stochastic, rule-based approach to annotating the morphemes of the given types using a graph-based, finite-state transducer; (ii) manually checking and revising the output and eliminating artificial/non-occurring ambiguities and theoretically possible multi-tags. After these stages, the entries of the NLP dictionary and actual running words of the corpus are matched via the software which has been developed by using PHP and MySQL.

III. Features of the TNC Interface

Web-based interface of the TNC provides for multitude of features for the analysis of corpus texts including concordance display (Fig. 1), sorting concordance data (Fig. 2), creating descriptive statistics for query results over the language-external restriction categories of texts via distribution (Fig. 3), and compiling lists of collocates (Fig. 4) for query terms on the basis of several statistical methods.



Yazılı metinlerdeki [derlem] sorgusu 6 sonuç ile (3) farklı metinde (47650926 sözcük [4458 toplam metinde] listelendi; bir milyon sözcükteki sıklık değeri: 0.13 (0.397 Saniye)

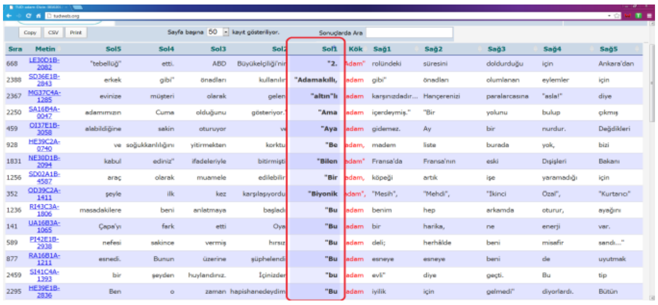
Sayfa başına 10 kayıt gösteriliyor. Sonuçlarda Ara

Sıra	Metin	Sorgu Sonuçları		
1	TI22F1D-4696	yakın geçmişte kullanıma sunulan Türkçe	derlem	(Odtü-Sabancı Türkçe ağaç yapılı derlemi)
2	QD03A3A-0446	sembollerle kazanılmış derlem bilgilerin sonucudur.	Derlem	bilgiler, belli bekleyişlere yönlendirici oldukları
3	SI22F1D-4695	derlem oluşturulması gereklidir. Böyle bir	derlem	oluşturmanın ilk aşaması cümle sonu
4	QD03A3A-0446	dolaysız ya da sembollerle kazanılmış	derlem	bilgilerin sonucudur. Derlem bilgiler, belli
5	SI22F1D-4695	biçimbilimsel analizlerin kolayca yapılabildiği bir	derlem	oluşturulması gereklidir. Böyle bir derlem
6	SI22F1D-4695	çözülmesi zor bir işlemdir, ancak	derlem	oluşturmanın en önemli aşamasıdır. Bu

6 kayıttan, 1 ile 6 arası gösteriliyor.

Fig. 1. TNC v1 concordance results page

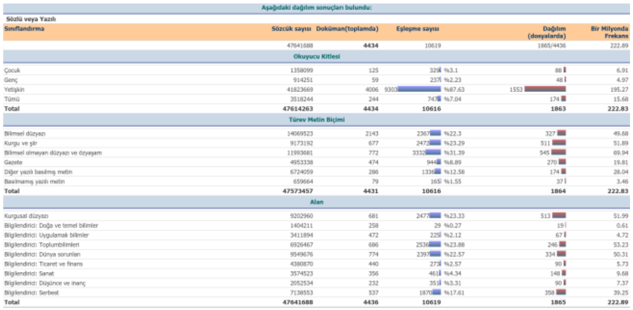
Fig. 1 shows the query results in the TNC which are given as concordance display (key word in context-KWIC). "A concordance is a list of all the occurrence of a particular search term in a corpus presented within the context in which they occur-usually a few words to the left and right to the search term" [22]. A search term in TNC can be a single word, multiword phrases and words containing wildcards. Concordances can be sorted alphabetically not only according to the node word but also the context up to 5 words to the left or right of the node word. This function of the TNC help users find linguistic patterns easily.



Sıra	Metin	Sol5	Sol4	Sol3	Sol2	Sol1	Kök	Sağ1	Sağ2	Sağ3	Sağ4	Sağ5
668	LE30D1B-2082	"tebellüğ"	etti.	ABD	Büyükelçiliği'nin	"2.	Adam"	rolündeki	süresini	doldurduğu	için	Ankara'dan
2388	SD36E1B-2843	erkek	gibi"	önadları	kullanılır	"Adamakıllı,	adam	gibi"	önadları	olumlanan	eylemler	için
2367	MG37C4A-1285	evinize	müşteri	olarak	gelen	"altın"lı	adam	karşınızdadır...	Hançerenizi	paralarcasına	"asla!"	diye
2250	SA16B4A-0047	adamımızın	Cuma	olduğunu	gösteriyor."	"Ama	adam	içerdeymiş."	"Bir	yolunu	bulup	çıkmış
459	OI37E1B-3058	alabildiğine	sakin	oturuyor	ve	"Aya	adam	gidemez.	Ay	bir	nurdur.	Değdikleri
928	HE39C2A-0740	ve	soğukkanlılığını	yitirmekten	korktu	"Be	adam,	madem	liste	burada	yok,	bizi
1831	NE30D1B-2094	kabul	ediniz"	ifadeleriyle	bitirmişti	"Bilen	adam"	Fransa'da	Fransa'nın	eski	Dışişleri	Bakanı
1256	SD02A1B-4587	araç	olarak	muamele	edilebilir	"Bir	adam,	köpeği	artık	işe	yaramadığı	için
352	OD39C2A-1411	şeyle	ilk	kez	karşılaşıyordu	"Biyonik	adam",	"Mesih",	"Mehdi",	"İkinci	Özal",	"Kurtarıcı"
1236	RI43C3A-1806	masadakilere	beni	anlatmaya	başladı	"Bu	adam	benim	hep	arkamda	oturur,	ayağını
141	UA16B3A-1065	Çapa'yı	fark	etti	Oya	"Bu	adam	bir	harika,	ne	enerji	var.
589	PI42E1B-2938	nefesi	sakince	vermiş	hırsız	"Bu	adam	deli;	herhâlde	beni	misafir	sandı..."
877	RA16B1A-1211	esnedi.	Bunun	üzerine	şüphelendi	"Bu	adam	esneye	esneye	beni	de	uyutmak
2459	SI41C4A-1393	bir	şeyden	huylandınız.	İçinizden	"bu	adam	evli"	diye	geçti.	Bu	tip
2295	HE39E1B-2836	Ben	o	zaman	hapishanedeydim	"Bu	adam	iyilik	için	gelmedi"	diyorlardı.	Bütün

Fig. 2. TNC v1 sorting function

Users can also view distributional information of the query result based on pre-defined meta-textual categories. The distribution page allows users to access descriptive statistics concerning the distribution of the query result of without performing multiple queries.

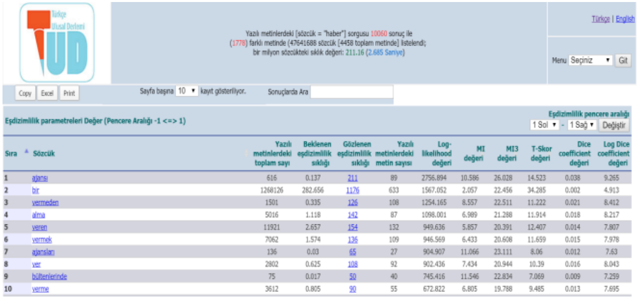


Aşağıdaki dağılım sonuçları bulundu:

Sözlü veya Yazılı

Sınıflandırma	Sözcük sayısı	Doküman(toplamda)	Eşleşme sayısı		Dağılım (dosyalarda)	Bir Milyonda Frekans
	47641688	4434	10619		1865/4436	222.89
Okuyucu Kitlesi						
Çocuk	1358099	125	329	%3.1	88	6.91
Genç	914251	59	237	%2.23	48	4.97
Yetişkin	41823669	4006	9303	%87.63	1553	195.27
Tümü	3518244	244	747	%7.04	174	15.68
Total	**47614263**	**4434**	**10616**		**1863**	**222.83**
Türev Metin Biçimi						
Bilimsel düzyazı	14069523	2143	2367	%22.3	327	49.68
Kurgu ve şiir	9173192	677	2472	%23.29	511	51.89
Bilimsel olmayan düzyazı ve özyaşam	11993681	772	3332	%31.39	545	69.94
Gazete	4953338	474	944	%8.89	270	19.81
Diğer yazılı basılmış metin	6724059	286	1336	%12.58	174	28.04
Basılmamış yazılı metin	659664	79	165	%1.55	37	3.46
Total	**47573457**	**4431**	**10616**		**1864**	**222.83**
Alan						
Kurgusal düzyazı	9202960	681	2477	%23.33	513	51.99
Bilgilendirici: Doğa ve temel bilimler	1404211	258	29	%0.27	19	0.61
Bilgilendirici: Uygulamalı bilimler	3411894	472	225	%2.12	67	4.72
Bilgilendirici: Toplumbilimleri	6926467	686	2536	%23.88	246	53.23
Bilgilendirici: Dünya sorunları	9549676	774	2397	%22.57	334	50.31
Bilgilendirici: Ticaret ve finans	4380870	440	273	%2.57	90	5.73
Bilgilendirici: Sanat	3574523	356	461	%4.34	148	9.68
Bilgilendirici: Düşünce ve inanç	2052534	232	351	%3.31	90	7.37
Bilgilendirici: Serbest	7138553	537	1870	%17.61	358	39.25
Total	**47641688**	**4436**	**10619**		**1865**	**222.89**

Fig. 3. TNC v1 distribution function



Yazılı metinlerdeki [sözcük = "haber"] sorgusu 10060 sonuç ile (1778) farklı metinde (47641688 sözcük [4458 toplam metinde] listelendi; bir milyon sözcükteki sıklık değeri: 211.16 (2.685 Saniye)

Eşdizimlilik parametreleri Değer (Pencere Aralığı -1 <=> 1)

Sıra	Sözcük	Yazılı metinlerdeki toplam sayı	Beklenen eşdizimlilik sıklığı	Gözlenen eşdizimlilik sıklığı	Yazılı metinlerdeki metin sayısı	Log-likelihood değeri	MI değeri	MI3 değeri	T-Skor değeri	Dice coefficient değeri	Log Dice coefficient değeri
1	ajansı	616	0.137	211	89	2756.894	10.586	26.028	14.523	0.038	9.265
2	bir	1268126	282.656	1176	633	1567.052	2.057	22.456	34.285	0.002	4.913
3	vermeden	1501	0.335	126	108	1254.165	8.557	22.511	11.222	0.021	8.412
4	alma	5016	1.118	142	87	1098.001	6.989	21.288	11.914	0.018	8.217
5	veren	11921	2.657	154	132	949.636	5.857	20.391	12.407	0.014	7.807
6	vermek	7062	1.574	136	109	946.569	6.433	20.608	11.659	0.015	7.978
7	ajansları	136	0.03	65	27	904.907	11.066	23.111	8.06	0.012	7.63
8	ver	2802	0.625	108	92	902.436	7.434	20.944	10.39	0.016	8.043
9	bültenlerinde	75	0.017	50	40	745.416	11.546	22.834	7.069	0.009	7.259
10	verme	3612	0.805	90	55	672.822	6.805	19.788	9.485	0.013	7.695

Fig. 4. TNC v1 result of a collocation analysis of *haber* 'news'

Collocation function allows users to list collocates (the words that the query-term occurs most frequently with) by offering six statistical association measures for calculating collocational strength: Log-likelihood, MI, MI3, T-score, Dice coefficient and Log Dice coefficient.

TNC v2, on the other hand, offers new features and query options. Since v2 is morphologically annotated, lemma form searches, morphemes and morpheme sequences and PoS-tag restricted searches (Fig. 5 and Fig. 6) can be conducted. As for some of the new features, users can save query history and they can search spoken component of the corpus by using meta-textual categories such as genre, domain, interaction type, speakers' age, sex.

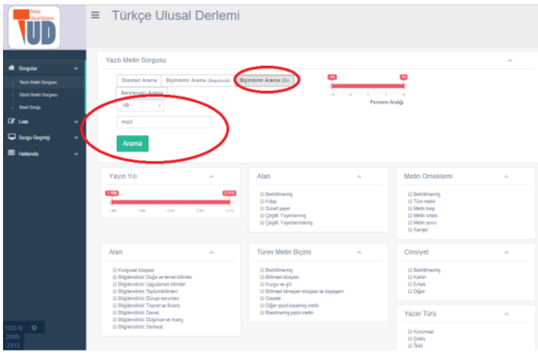
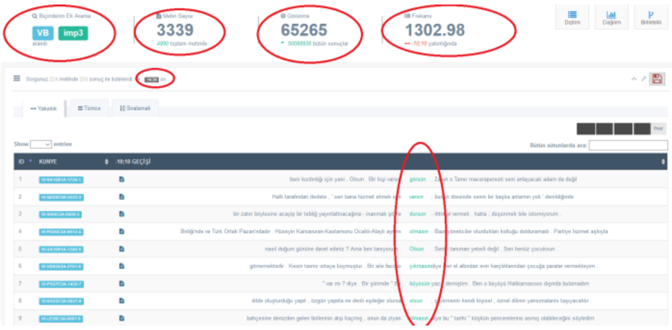


Fig. 5. TNC v2 PoS-tag query



Sorgunuz 224 metinde 250 sonuç ile listelendi.

ID	KÜNYE		-10;10 GEÇİŞİ		
1	[illegible]		beni kızdırdığı için yani . Olsun . Bir kişi varsa	görsün	. Zaten o Taner maceraperesti seni anlayacak adam da değil
2	[illegible]		Halk tarafından devlete , " sen bana hizmet etmek için	varsın	; bunun dışında senin bir başka anlamın yok " dendiğinde
3	[illegible]		bir zatın böylesine acayip bir tebliğ yayınlattıracağına - inanmak şöyle	dursun	- ihtimal vermek , hatta ; düşünmek bile istemiyorum .
4	[illegible]		Birliği'nde ve Türk Ortak Pazarı'ndadır . Hüseyin Karsavuran-Kastamonu Ocaklı-Alaylı ayrımı	olmasın	: Bazı yöneticiler oturdukları koltuğu dolduramadı . Partiye hizmet aşkıyla
5	[illegible]		nasıl doğum gününe davet ederiz ? Ama ben tanıyorum .	Olsun	. Sen tanıman yeterli değil . Sen henüz çocuksun .
6	[illegible]		gitmemektedir . Kesin tanını ortaya koymuştur . Bir aile facia	çıkmasın	diye ben el altından evin harçlıklarından çocuğa paralar vermekteyim .
7	[illegible]		" var mı ? diye . Bir şiirimde " Bir	büyüsün	yaz demiştim . Ben o büyüyü Halikarnassos dışında bulamadım
8	[illegible]		dilde oluşturduğu yapıt , özgün yapıtla ne denli eşdeğer olursa	olsun	, çevirmenin kendi kişisel , öznel dilinin yansımalarını taşıyacaktır .
9	[illegible]		bahçesine denizden gelen bitlerinin atıp kaçmış , onun da ziyan	olmasın	diye bu " tarihi " köşkün pencerelerine asmış olabileceğini söyledim

Fig. 6. TNC v2 PoS-tag query results

IV. THE ARCHITECTURES OF TNC V1 AND TNC V2

TNC is a user-friendly, platform independent, Web-based corpus developed for Turkish language. HTML [12], CSS [7], PHP [5] [6], and JavaScript [8] languages, and MySQL [4] database management system are used for implementation of the TNC. The main architecture of TNC version 1 is presented in Fig. 7. To develop TNC v1, text documents in the written component of the corpus are first pre-processed to extract metadata such as author, year, source, domain etc. that describe each document in the collection. Metadata of each document are stored in a MySQL table on disk. After metadata extraction step, each token, which is a character string separated by white space characters, in each document is identified and unique token list is formed from all documents in the collection. Each token is given a unique identifier and while unique tokens are found from documents, their frequencies in each document are also counted. Unique tokens, their ids, and frequencies are stored in another MySQL table. For each unique token found from the document collection, a kind of inverted index structure is formed. In the index structure position of each unique token are stored for each document in the collection. This index structure is stored over disk by using MyISAM file structure of MySQL. By using the inverted index structure, concordance data, descriptive statistics, and lists of collocates for unique tokens in the corpus are computed and they are stored as compressed files over disk by applying IGBinary [9] compression method of PHP. IGBinary applies binary data compression and storage therefore reading and decompression of the data are performed faster with respect to other compression methods. The unique token list and names of its compressed data files including concordance data are then loaded to memory as a hash table to improve performance of user searches. When a user sends a query by using the TNC GUI, the queried token is searched from the hash table and the name of the compressed concordance file of the token is found. After that the compressed concordance file is read from disk to memory, then this file is decompressed and if the user gives some filtering options in his query these filters are applied over the decompressed file, then the computed results are randomly shuffled and displayed to the user.

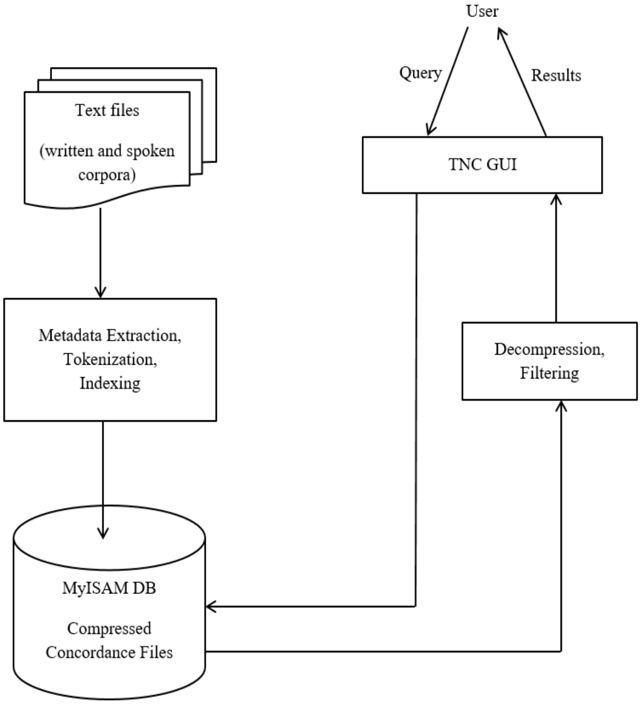

User
Query
Results
Text files
(written and spoken corpora)
TNC GUI
Metadata Extraction, Tokenization, Indexing
Decompression, Filtering
MyISAM DB
Compressed Concordance Files

Fig. 7. Architecture of the TNC v1

The TNC v2 is an updated and improved version of the TNC v1. Metadata extraction, tokenization and indexing steps are similar to that of the TNC v1. Metadata are stored over disk as a MySQL table. Unique token list including frequencies for each document are loaded to memory instead of storing over disk. Only document collection and metadata for the documents are stored on disk. For all unique tokens in the collection, a kind of inverted index structure is constructed in which the positions of the token in each document are stored. This inverted index structure is located in memory by using Redis [10] which is an open source (BSD licensed), in-memory data structure store and supports data structures such as strings, hashes, lists, sets, sorted sets, etc. When a user sends a query by using the TNC GUI, the queried token is searched from the in-memory inverted index and unique types forming the concordance output of queries, descriptive statistics for query results, and lists of collocates are computed in real time. If the user gives some filtering in his query, these filters are searched from metadata table stored in the database, and the results of this search are used to filter unique type lists for the given token. Finally, the computed concordances are shuffled and a random number of results are displayed to the user. The architecture of the TNC v2 is presented in Fig. 8. As the inverted index structure is stored in memory, all computations are performed very fast as it is shown in the next section.

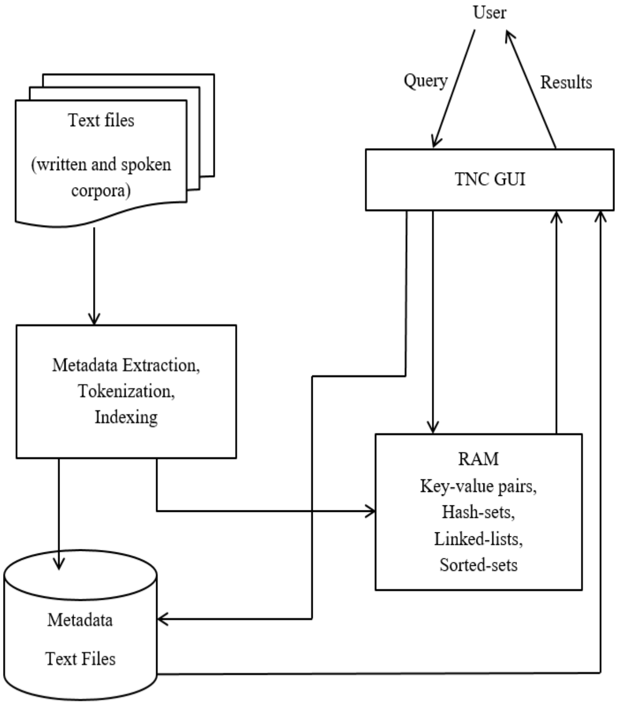


Fig. 8. Architecture of the TNC v2

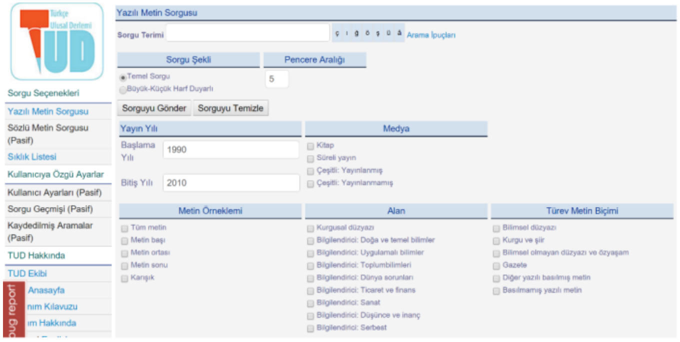
On the other hand, the system specifications of the computer running the TNC v1 interface are prominently different from the TNC v2. The system properties of the server running the TNC v2 interface seems sufficient enough to process and store huge amount of data in memory. Table III briefly presents the major hardware specifications of both versions.

TABLE III. HARDWARE SPECIFICATIONS OF COMPUTERS RUNNING TWO VERSIONS OF THE TNC

	OS	RAM	CPU	Disk
TNC v1	FreeBSD 9.0	16 GB	1 X Intel Xeon x3440 2.53 GHz 4 cores	500 GB SATA 2
TNC v2	Ubuntu Server 14.04 (Virtual machine running on FreeBSD host)	64 GB	2 X Intel Xeon E5-2630v2 2.60 GHz 2 cores	350 GB Virtual Disk

V. QUERIES ON TNC V1 AND TNC V2

In what follows the speed of two versions of the TNC are compared on the basis of standard, restricted and wildcard queries conducted on the written component of the TNC v1 and written and spoken components of TNC v2. Fig. 9 and Fig. 10 respectively show the main pages of the both versions.



● Temel Sorgu
○ Büyük-Küçük Harf Duyarlı

Pencere Aralığı: 5

Başlama Yılı: 1990

Bitiş Yılı: 2010

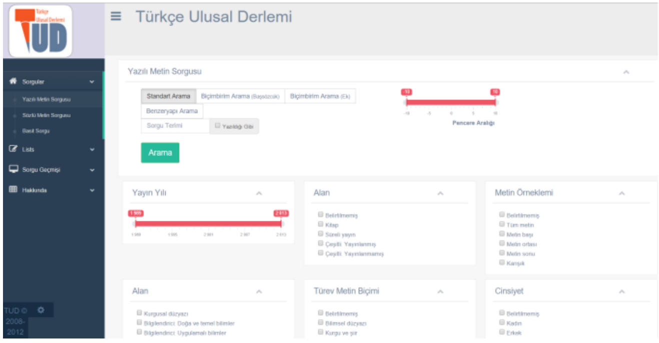
Medya: □ Kitap □ Süreli yayın □ Çeşitli: Yayınlanmış □ Çeşitli: Yayınlanmamış

Metin Örneklemi: □ Tüm metin □ Metin başı □ Metin ortası □ Metin sonu □ Karışık

Alan: □ Kurgusal düzyazı □ Bilgilendirici: Doğa ve temel bilimler □ Bilgilendirici: Uygulamalı bilimler □ Bilgilendirici: Toplumbilimleri □ Bilgilendirici: Dünya sorunları □ Bilgilendirici: Ticaret ve finans □ Bilgilendirici: Sanat □ Bilgilendirici: Düşünce ve inanç □ Bilgilendirici: Serbest

Türev Metin Biçimi: □ Bilimsel düzyazı □ Kurgu ve şiir □ Bilimsel olmayan düzyazı ve özyaşam □ Gazete □ Diğer yazılı basılmış metin □ Basılmamış yazılı metin

Fig. 9. TNC v1 main page



Alan: □ Belirtilmemiş □ Kitap □ Süreli yayın □ Çeşitli: Yayınlanmış □ Çeşitli: Yayınlanmamış

Metin Örneklemi: □ Belirtilmemiş □ Tüm metin □ Metin başı □ Metin ortası □ Metin sonu □ Karışık

Alan: □ Kurgusal düzyazı □ Bilgilendirici: Doğa ve temel bilimler □ Bilgilendirici: Uygulamalı bilimler

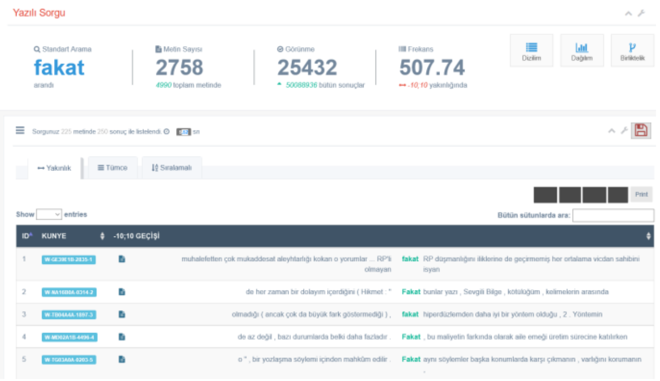
Türev Metin Biçimi: □ Belirtilmemiş □ Bilimsel düzyazı □ Kurgu ve şiir

Cinsiyet: □ Belirtilmemiş □ Kadın □ Erkek

Fig. 10. TNC v2 main page

A. Standard Queries

Standard search in the TNC offers users to make searches in the whole of the corpus without filtering the queries on the basis of written or spoken parts of the corpus. Users type the search term in the form labeled query term and send it. Just on top of the results page, users can view frequency information of the node word. A normalized frequency of a 1-million-word scale is also stated. Query results are displayed in a KWIC view by default. Each column in the result page displays the ID of the concordance line, the text where the node word is found and the concordance line, respectively. Users can display the further context to the left and right of the node word by clicking search term in the concordance lines. When such a query is made for exact form of the node word *fakat*, it takes just about 5.52 seconds to compute concordance lines among 2758 different corpus text in the TNC v2 (Fig. 11), while it takes 14.57 seconds for the same query word in the TNC v1 (Fig. 12).



fakat arandı

2758 — 4990 toplam metinde

25432 — 50088936 bütün sonuçlar

507.74 — -10; 10 yakınlığında

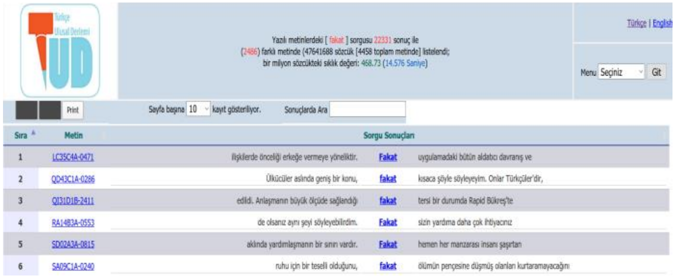
Sorgunuz 225 metinde 250 sonuç ile listelendi. sn

Show entries

Bütün sütunlarda ara:

ID	KUNYE	-10;10 GEÇİŞİ		
1	W-GE39E1B-2835-1	muhalefetten çok mukaddesat aleyhtarlığı kokan o yorumlar ... RP'li olmayan	fakat	RP düşmanlığını ilklerine de geçirmemiş her ortalama vicdan sahibini isyan
2	W-NA16B0A-0314-2	de her zaman bir dolayım içerdiğini (Hikmet : "	Fakat	bunlar yazı , Sevgili Bilge , kötülüğüm , kelimelerin arasında
3	W-TB04A4A-1897-3	olmadığı (ancak çok da büyük fark göstermediği) ,	fakat	hiperdüzlemden daha iyi bir yöntem olduğu , 2 . Yöntemin
4	W-MD02A1B-4496-4	de az değil , bazı durumlarda belki daha fazladır .	Fakat	, bu maliyetin farkında olarak aile emeği üretim sürecine katılırken
5	W-TG03A0A-0203-5	o " , bir yozlaşma söylemi içinden mahkûm edilir .	Fakat	aynı söylemler başka konumlarda karşı çıkmanın , varlığını korumanın .

Fig. 11. TNC v2 query results-*fakat* ‘but’



Yazılı metinlerdeki [fakat] sorgusu 22331 sonuç ile (2486) farklı metinde (47641688 sözcük [4458 toplam metinde] listelendi; bir milyon sözcükteki sıklık değeri: 468.73 (14.576 Saniye)

Sayfa başına 10 kayıt gösteriliyor. Sonuçlarda Ara

Sıra	Metin	Sorgu Sonuçları		
1	LC35C4A-0471	ilişkilerde önceliği erkeğe vermeye yöneliktir.	Fakat	uygulamadaki bütün aldatıcı davranış ve
2	QD43C1A-0286	Ülkücüler aslında geniş bir konu,	fakat	kısaca şöyle söyleyeyim. Onlar Türkçüler'dir,
3	QI31D1B-2411	edildi. Anlaşmanın büyük ölçüde sağlandığı	fakat	tersi bir durumda Rapid Bükreş'te
4	RA14B3A-0553	de olsanız aynı şeyi söyleyebilirdim.	Fakat	sizin yardıma daha çok ihtiyacınız
5	SD02A3A-0815	akında yardımlaşmanın bir sınırı vardır.	Fakat	hemen her manzarası insanı şaşırtan
6	SA09C1A-0240	ruhu için bir teselli olduğunu,	fakat	ölümün pençesine düşmüş olanları kurtaramayacağını

Fig. 12. TNC v1 query results-*fakat* ‘but’

On the other hand, while the TNC v1 does not allow the search of one of the most frequent word *kadar* ‘until’, which ranks 45 with 142693 frequency of occurrence in the frequency list of the TNC, the architecture of TNC v2 allows its search by displaying random in 10.82 seconds to users.

TABLE IV. THE STANDARD QUERY OF *FAKAT* ‘BUT’ AND *KADAR* ‘UNTIL’ WITHIN WRITTEN COMPONENT OF THE TNC

Query item	TNC version	Word count	Text count	Hits	Different text	Time
fakat ‘but’	TNC v1	47641688	4458	22331	2486	14.57 sec
	TNC v2	50088936	4990	25432	2758	5.52 sec
kadar ‘until’	TNC v1	47641688	4458	N/A	N/A	> 60 sec
	TNC v2	50088936	4990	133807	4252	10.82 sec

B. Restricted Query

Restricted queries can be performed in the written component of TNC with the criteria of publication date, media, sample, domain, derived text type, author information, audience and genre. Table V demonstrates such a sample query performed by restricting the node word *büyük* ‘big’ in terms of the publication date (between 1995-2005), medium (books) and sample (whole text) of the corpus documents. Once again the TNC v2 is fast in the restricted query search. It only takes 3.52 seconds to produce concordance lines in the v2, while the same query lasts 9.31 seconds in the v1.

TABLE V. THE RESTRICTED STANDARD QUERY OF BÜYÜK ‘BIG’ IN TERMS OF PUBLICATION DATE (1995-2005), MEDIUM (BOOKS) AND SAMPLE (WHOLE TEXT) WITHIN WRITTEN COMPONENT OF THE TNC

Query item	TNC version	Word count	Text count	Hits	Different text	Time
büyük	TNC v1	47641688	4458	3476	168	9.31 sec
‘big’	TNC v2	50088936	4990	3079	170	3.52 sec

C. Wildcard Queries

Wildcards are also used in standard and restricted queries in the TNC. Special character * permits users to search word forms starting with *kol*, such as *kolay* ‘easy’, *kollarına* ‘to his arms’, *koltuğa* ‘to the armchair’, as is seen in Table VI the TNC v2 is slightly faster than that of v1 in displaying query results.

The wildcard query aims to obtain word forms containing both /b/ and /p/ as the final sound of *kitap* is only permitted in the TNC v2 and 41,098 hits are found in across the corpus documents in 22.25 seconds.

Multi-unit search pattern where *beyaz* 'white' or *peynir* 'cheese' is queried across the corpus documents. The speed of the TNC v2 is again better than v1. The query in written and spoken parts of the corpus returned 12,212 hits in 2,085 different texts in 1.73 seconds.

Owing to in-memory index structure of the TNC v2 it is possible to search lexical items used frequently in Turkish such as *ama* 'but' (ranking 43 among 73,383 lemmas in the NLP Dictionary of TNC) and *bu* 'this' (ranking 6 among 73,383 lemmas in the NLP Dictionary of TNC) in a reasonable fastness. *Ama* 'OR' *bu* wildcard query returned relevant strings within 15.66 seconds in the TNC v2 but the same query takes more than 60 seconds in the v1. As a final remark, the speed of TNC v2 concerning some other wildcard query options needs to be optimized.

TABLE VI. THE WILDCARD QUERIES IN THE TNC

Query item	TNC version	Word count	Text count	Hits	No. of diff. text	Time
kol*	TNC v1	47,641,688	4,458	53,041	3,523	30.78 sec
	TNC v2	50,088,936	4,990	58,154	3,864	22.95 sec
kita[b,p]*	TNC v1	47,641,688	4,458	N/A	N/A	N/A
	TNC v2	50,088,936	4,990	41,098	2,687	22.25 sec
beyaz|peynir	TNC v1	47,641,688	4,458	10,881	1,894	6.46 sec
	TNC v2	50,088,936	4,990	12,212	2,085	1.73 sec
ama|bu	TNC v1	47,641,688	4,458	N/A	N/A	N/A
	TNC v2	50,088,936	4,990	836,838	4,565	15.66 sec

VI. CONCLUSION

This paper describes the design principles, interface features and the architecture of the TNC. Then it compares the architecture of the TNC v1 and v2. On the basis of the standard, restricted and wildcard corpus queries, it is shown that in-memory inverted index structure of the TNC v2 computes better and faster than that of v1 which is designed as disk-based compressed concordance data files for each unique term. In terms of speed, the v2 architecture allows users to perform searches across many corpus files (5,412 data files of the TNC) very rapidly, but such architecture needs more memory to display query results fast. We should also note that the relational database structure used in both versions of the TNC has its advantages to process large corpus files such that it allows for a "modular structure in which any number of features can be incorporated in to the architecture" [11]. For future work any extension in the features of the TNC would be possible via relational database and inverted index structures.

ACKNOWLEDGMENT

This work is supported by TÜBİTAK (Grant No: 115K135, 113K039).

REFERENCES

[1] M. Aksan and Y. Aksan, "Linguistic corpora: A view from Turkish," in *Studies in Turkish Language Processing*, K. Oflazer and M. Saraçlar, Eds. Berlin: Springer Verlag, (forthcoming).

[2] B. Say, D. Zeyrek, K. Oflazer and U. Özge, "Development of a corpus and a treebank for present-day written Turkish," in *Proceedings of the 11th International Conference of Turkish Linguistics*, 2004, pp 183–192.

[3] Y. Aksan, M. Aksan, A. Koltuksuz, T. Sezer, Ü. Mersinli, U. U. Demirhan, H. Yılmazer, Ö. Kurtoğlu, G. Atasoy, S. Öz and İ. Yıldız, "Construction of the Turkish National Corpus (TNC)," in *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC)*, 2012, pp. 3223-3227.

[4] MySQL 5.5 Release Notes, http://dev.mysql.com/doc/relnotes/mysql/5.5/en/

[5] PHP 5.4.21, http://www.php.net/releases/5_4_21.php

[6] PHP 5.6.10, http://www.php.net/releases/5_6_10.php

[7] CSS, http://www.w3schools.com/css/

[8] Javascript, http://www.w3schools.com/js/

[9] PHP PECL IGBinary Extension, http://codepoets.co.uk/2011/php-serialization-igbinary/

[10] Redis, http://redis.io/

[11] M. Davies, "The 385+million word Corpus of Contemporary English (1990-2008+)," *International Journal of Corpus Linguistics*, vol. 14, no. 2, pp. 159-160, 2009.

[12] HTML, http://www.w3schools.com/html/

[13] M. Aksan and Ü. Mersinli, "A corpus based Nooj module for Turkish," in *Proceedings of the NooJ 2010 International Conference and Workshop*, 2011, pp. 29-39.

[14] T. McEnery, R. Xiao, and Y. Tono, *Corpus-based Language Studies*, London: Routledge, 2006.

[15] M. Davies, "The Corpus of Contemporary American English as the first reliable monitor corpus of English," Literary and Linguistic Computing, vol. 25, no. 4, pp. 447-464, 2010.

[16] S. Akşehirli, "Dereceli karşıt anlamlılarda belirtisizlik ve ölçek yapısı," *Journal of Language and Linguistic Studies*, vol. 10, no. 1, 49-66, 2014.

[17] G. İşgüder Şahin and E. Adalı, "Using morphosemantic information in construction of a pilot lexical semantic resource for Turkish," in *Proceedings of the 21st International Conference on Computational Linguistics*, 2014, pp. 929-936.

[18] S. Demir, "Generating valence shifted Turkish sentences," in *Proceedings of 8th INLG*, 2014, pp. 128-132.

[19] O. Yılmaz, "Tag-based semantic website recommendation for Turkish language," *International Journal of Scientific and Engineering Research*, vol. 4, no. 3, pp. 1-7, 2013.

[20] A. Uçar and Ö. Kurtoğlu, "A corpus-based account of polysemy in Turkish: A case of ver-'give'," in *Proceedings of the 15th International Conference on Turkish Linguistics*, 2012, pp. 539-551.

[21] Ü. Mersinli, "Associative measures and multi-word unit extraction in Turkish," *Journal of Language and Literature*, vol. 12, no. 1, pp. 43-61, 2015.

[22] P. Baker, A. Hardie and T. McEnery, *A Glossary of Corpus Linguistics*, Edinburg: Edinburg Press, 2006.