Assembling the Kazakh Language Corpus

Olzhas Makhambetov, Aibek Makazhanov, Zhandos Yessenbayev,
Bakhyt Matkarimov, Islam Sabyrgaliyev, and Anuar Sharafudinov
Nazarbayev University
Research and Innovation System
53 Kabanbay batyr ave., Astana, Kazakhstan
{omakhambetov, aibek.makazhanov, zhyessenbayev, bmatkarimov,
islam.sabyrgaliyev, anuar.sharaphudinov}@nu.edu.kz

Abstract

This paper presents the Kazakh Language Corpus (KLC), which is one of the first attempts made within a local research community to assemble a Kazakh corpus. KLC is designed to be a large scale corpus containing over 135 million words and conveying five stylistic genres: literary, publicistic, official, scientific and informal. Along with its primary part KLC comprises such parts as: (i) annotated sub-corpus, containing segmented documents encoded in the eXtensible Markup Language (XML) that marks complete morphological, syntactic, and structural characteristics of texts; (ii) as well as a sub-corpus with the annotated speech data. KLC has a web-based corpus management system that helps to navigate the data and retrieve necessary information. KLC is also open for contributors, who are willing to make suggestions, donate texts and help with annotation of existing materials.

1 Introduction

This article describes theoretical and practical issues experienced during the construction of the Kazakh Language Corpus. Kazakh is an agglutinative and highly inflected language which belongs to the Turkic group. It is official state language of Kazakhstan and a mother tongue for more than 10 million people all around the world. However, up until the early 90's of 20th century, due to historical reasons of the Soviet era, Russian language was the predominant language in spoken and written communication in Kazakhstan. This fact in turn caused the problem of underrepresentation of Kazakh language in various fields such as science, entertainment, official documentation, etc. For this reason, while assembling the corpus, we had to group categories that are generally presented as separate in other corpora into five stylistic genres. Also, in contrast to other corpora (Aksan et al., 2012; Chen, 1996), we included texts as they were available, i.e we did not try to fill a predefined set of categories. Substantial part of materials was collected using source-customized web crawlers and donated texts.

KLC also contains a manually annotated sub-corpus with morpho-syntactic and structural markups encoded in XML following general notions outlined in CES (Ide, 1998). Our syntactic tagset comprises a set of syntactic categories well-defined in a classical Kazakh grammar, and the part of speech (POS) tagset is based on a positional system in which the tags are formed by concatenation of POS of a word form and a chain of encoded linguistic properties, such as number, case, voice etc. The annotations have been carried out manually by philology students specializing in morphology and syntax. Trying to make the annotation process as comfortable as possible, we have designed a web-based annotation tool with a user-friendly interface. We took a great care for the annotation quality, and to do that we (i) arranged the validation process, and (ii) equipped the tool with a recommendation system that, as we will show, improves the inter-annotator agreement.

As a part of KLC we have also compiled the annotated read-speech corpus (RSC), which includes audio recordings of words, phrases, sentences (from all genres), news articles and excerpts from books, that were carefully chosen from the primary part of the corpus. All text materials were read by volunteers who represented different age, gender, region and education backgrounds in a balanced way. Each audio file is accompanied with a label file and a corresponding text transcript. Moreover, some of the transcripts have been grammatically annotated, i.e. in addition to a word-level segmentation of audio information a portion of our data has lexical, and morpho-syntactic annotations. In total RSC contains 10GB or more than 40 hours of speech.

This paper is organized as follows. Section 2 reviews the existing work. Section 3 provides detailed information about the primary corpus. Sections 4 and 5 thoroughly describe annotated text and speech sub-corpora respectively. Finally, we draw conclusions and discuss future work in Section 6.

Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1022–1031,
Seattle, Washington, USA, 18-21 October 2013. ©2013 Association for Computational Linguistics

2 Related Work

Since the pioneering corpus of Brown University was completed in 1964 by Francis and Kučera (1979), corpus linguistics has become a thriving research field. Over the past two decades researchers all around the world released many corpora, including well known British National Corpus (BNC) (Burnard, 2007) developed between 1991 and 1994, and containing more than 100 million words of written and spoken language from a wide range of sources (Ide and Macleod, 2001; Al-Sulaiti and Atwell, 2006). All materials were selected on a basis of three independent criteria (medium, domain and time), where each criterion had predefined target proportions. The spoken part (remaining 10%) consists of orthographic transcriptions of unscripted informal conversations and spoken language collected in different contexts. BNC is tagged for part of speech (POS) using the CLAWS4 (Constituent Likelihood Automatic Word-tagging System) (Leech et al., 1994) tagging system developed at Lancaster University. BNC is generally accepted as a balanced corpus, and many researchers, such as the creators of Turkish National Corpus (Aksan et al., 2012), Korean National Corpus (Kim, 2006) etc., adopted it as a model for compiling their own corpora.

The Russian National Corpus (RNC) has been released by the group of specialists from different organizations led by the Institute of Russian language, Russian Academy of Sciences (Ruscorpora, 2003). The corpus covers primarily a period from the middle of the XVIII to the early XXI centuries. It includes both written texts (fiction, memoirs, science, religious literature and others) and recorded spoken data (public speeches and private conversations). Currently RNC contains over 350 million word forms that are automatically POS-tagged and lemmatized. The corpus also includes semantic tags for words and texts (Apresjan et al., 2006). Along with its main part, RNC contains such subcorpora as: Deeply Annotated Corpus, that contains sentences with a complete morphological and syntax structure markup, where the syntax structure is largely based on the Meaning-Text Theory introduced by Aleksandr Žolkovskij and Igor Mel'čuk; English – Russian, German – Russian, Ukrainian – Russian, Belorussian – Russian parallel corpora; Dialect corpus; Poetry corpus and others.

Unfortunately, up until now, not too much work has been accomplished in developing a corpus that will represent Kazakh language. To the best of our knowledge there has been a limited number of attempts to compile one, but resulting corpora are too small in size and scope, or not available to the public. A Kazakh corpus has been initiated by the Committee on Languages of the Ministry of Culture of the Republic of Kazakhstan (CLMCRK, 2009). This corpus is small in size and not annotated, as it remains in its very early stage of development. The new sub-corpus for Kazakh has been recently built by Baisa et al. as a part of larger corpus of Turkic languages (Baisa and Suchomel, 2012). This corpus was compiled using a web crawler that selected texts based on a language model trained on Wikipedia texts. Although the obtained corpus is relatively large in size, the data was not categorized by genres. Also, since a crawler was not source-customized, the corpus may contain some noise coming in the form of text in Russian or other languages. We also could not find enough information about a Kazakh corpus that has been developed at Xinjiang University and used in their research (Altenbek and Xiao-long, 2010). The absence of an available corpus that will be large enough to represent Kazakh language decelerates many research activities (Mukan, 2012). We believe that building an open Kazakh corpus will have a significant impact and it will be very useful tool in the analysis of Kazakh.

3 KLC Primary Corpus

KLC is one of the first attempts to build a large scale, general purpose corpus that represents the present state of Kazakh language. Currently, the size of the primary corpus is more than 135 million words and it contains approximately more than 400 000 documents classified by genres into the following five sections: (1) *literary* section contains Kazakh literary texts that were published in the range from the beginning of the XX century till present; (2) *official* section includes mainly official statutes, orders, acts and other materials produced by the governmental organizations within the period of 2009-2012; (3) *scientific* section includes books, research monographs, dissertations, articles and essays from various fields (informatics, biology, chemistry, etc.); (4) *publicistic* section contains periodicals and articles from online sources, i.e. newspapers and magazines published over the last ten years; (5) *informal language* section includes documents with colloquial Kazakh texts extracted from the popular blog platforms starting from 2009. We have to note that while compiling this corpus we intentionally relaxed the document selection criteria by not restricting the collected data to particular domains, media, and time. This was mainly dictated by the lack of materials, and partially due to the reasons mentioned in the introduction.

Our main sources of data were Internet websites as well as digitized forms of books, dissertations and articles from public and personal libraries. For each website we designed a source-specific crawler, thereby increasing the precision of the meta data (e.g. authors, news categories, etc.) extraction. Additionally, we filtered out documents with a high consistency of Russian texts by aligning them to a language model trained on pure Russian texts. We also filtered out all documents with the size

Genre	# docs	# all words	# unique words
Literary	8 255	7 733 456	423 445
Publicistic	404 884	79 302 154	951 659
Official	25 302	44 670 856	335 264
Scientific	527	2 227 878	153 877
Informal	6 110	1 337 953	162 074
TOTAL	**445 078**	**135 272 297**	**1 365 202**

Table 1: A quantitative description of the corpus.

documents, total	1213
documents, %	0.3
all words, total	613 511
all words, %	0.4
unique words, total	80 368
unique words, %	5.9
lemmata, total	42 901

Table 2: A quantitative description of the annotated data

less than 1kB. It took about 7 months to grow the corpus to its current size. Table 1 provides a general quantitative description of the corpus.

We release the data under a license that in accordance with Kazakhstan‘s law allows distribution of some materials in whole (official documents, news articles) and some only in part (literature, scientific texts, analytics) provided that sources are properly cited. This license does not allow printed or electronic publications or similar use of substantial portions of text drawn from the corpus without the permission of its original publisher(s) or copyright holder(s).

3.1 Text Documents Description

Each document is stored in a plain text format in the UTF-8 encoding. Documents contain both the content and the meta-data in a single file, and have the following simple structure:

- TITLE – the title of a document;
- SOURCE – the source of a document
- AUTHOR – the author(s) of a document;
- DATE – the date when a document was published;
- META – additional information;
- TEXT – the content of a document.

Provided that the corresponding information is present in a source, the <META> tag contains both the name of the section of the corpus to which a document belongs and a further categorical sub-division, such as the type of a literary work, e.g. a poem. That is, whenever possible such categories are assigned automatically, e.g. some websites provide this information. For sources that lack meta data, such as the digitized books, dissertations and scientific papers, the corresponding categories (informatics, biology, chemistry, etc.) are assigned manually.

3.2 Writing System of Kazakh language

Kazakh adopts different writing systems depending on the regions where it is spoken (Cyrillic alphabet in Kazakhstan, Arabic and Latin graphics in other countries). Recently the government of Kazakhstan has decided to adopt Kazakh alphabet to a Latin graphic. In this regard we believe that KLC could become a valuable tool. Indeed, we have already provided a group working on this problem with statistical information about letter distributions in Kazakh texts. This information could also aid in designing various speech corpora as well as a proper Kazakh keyboard layout. It can be stated that the latter was done rather carelessly just as a simple adjustment to a Russian keyboard (Wikipedia, 2012). Current Kazakh Cyrillic alphabet consists of 42 letters, whereas 9 of them are pure Kazakh letters and the others adopt the Russian symbolic. Figure 1 shows the distribution of Kazakh letters in the corpus. It can be seen that there is a small non-zero distribution of pure Russian letters (underlined). This can be explained by the ineluctable use of Russian words due to the lack of a proper translation or inheritance of Russian vocabulary.

4 The Annotated Sub-corpus

In order to enhance the effectiveness of the corpus as a research tool, we have annotated a portion of the data for syntactic and POS tags, lemmata, and for morpheme types and boundaries. Table 2 provides net amount and the percentages (with respect to the current size of the corpus) of the annotated data in terms of documents, words, unique words, and lemmata.

The annotation process has been carried out completely manually. We favored a manual annotation over a semi-automatic one, for the following two reasons: (i) finding language independent tools (not to mention Kazakh-specific) which support a fine grained level of annotation that we employ turned out to be rather challenging; (ii) though we refused and partially could not afford a semi-automatic annotation we provided the annotators with a *semi-automatic-like annotation* experience by equipping our annotation tool with a fairly advanced recommendation system. The annotation was performed mainly by the undergraduate students majoring in Kazakh philology. As a quality control measure, two validators (a graduate student majoring in Kazakh philology and one of the authors) were assigned to check a random sample of about 10% of the annotated data. Validators did not just fix errors, we also held regular "work-through-errors" sessions in an attempt to synchronize annotations. Our analysis of validated data suggest that the annotation

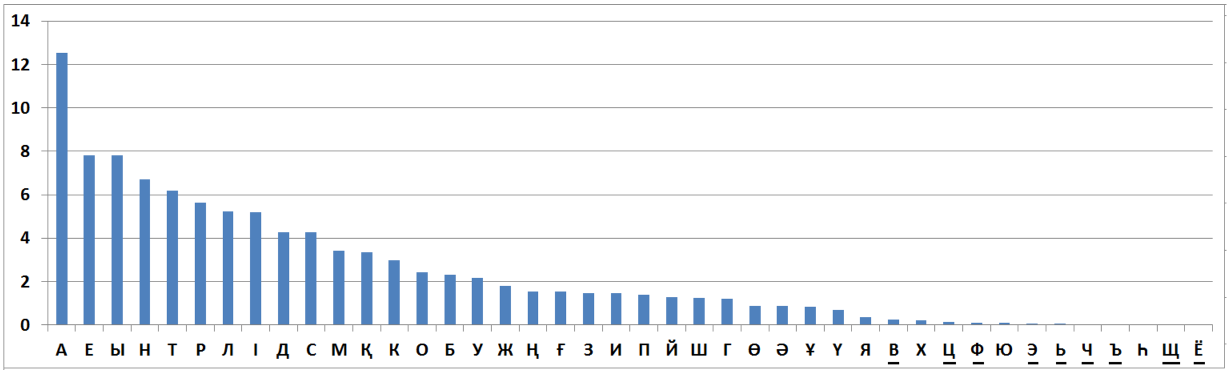


Figure 1: The distribution of letters across the corpus, in %.

Tag	Description	PTB equivalents
S	Simple declarative clause	S
BSS	Independent clause	S
BGS	Dependent clause	SBAR(Q)
BAS	Subject	NP
BND	Predicate	VP
TOL	Object	(WH)NP
ANT	Modifier	ADJP
PYS	Adverbial	(WH)ADVP (WH)PP
X	Void, unknown, uncertain	X

Table 3: The syntactic tagset description

#	Linguistic property	Code	Cardinality
1	Animacy	A	2
2	Number	N	2
3	Possessiveness	S	10
4	Person	P	8
5	Case	C	7
6	Negation	G	2
7	Tense	T	3
8	Mood	M	4
9	Voice	V	5

Table 4: Linguistic properties considered in the POS tagset design

quality was fairly high, as roughly only 6% of annotated tokens were fixed.

To the best of our knowledge this is the first attempt to annotate Kazakh texts with various linguistic markups. Given this, in the following subsections we would like to describe the tagsets (syntactic and POS), the annotation scheme (the format in which the annotated data is stored and distributed), and the annotation tool itself.

4.1 Designing the Tagsets

The syntactic tagset. At the initial stage of the corpus development we did not plan to build a detailed treebank, leaving this task for the future work. Therefore, our syntactic tagset comprises a compact set of syntactic categories well-defined in a classical grammar. Table 3 contains the tagset description along with the equivalent tags defined in a widely used Penn Treebank (Marcus et al., 1993) tagset[1]. In addition to that, we also label proverbials which are rather common elements of Kazakh language. We do not treat them as a separate syntactic category, for they typically serve as a single syntactic unit (e.g. predicate, adverbial, clause, etc.) Instead each syntactic tag has a corresponding binary property that marks the proverbial case.

The POS tagset. Kazakh is an agglutinative Turkic language, in which word forms are generated by means of the affix inflection. Different affix types mark different linguistic properties. For instance, consider a translation of a simple Kazakh sentence:

Mektepke bardym. - school.Dat go.Past.1sg - I went to school.

In this example pronoun *"I"* and preposition *"to"* are "hidden" in the affixes of case and person, i.e.:

Mektep(NN = *a school*) + *ke* (dative case = *to school*)

bar(VB, imperative = *go*) + *dy* (past tense = *he/she went*) + *m* (1st person = *I went*)

As the example shows, inflected affix chains contain important information that is not always present in the context, hence a tagset should be designed in a way to capture this information to the extent possible. For this reason, we design a positional tagset (Oflazer et al., 2003; Hajič and Hladká, 1998; Hana and Feldman, 2010), in which the final tags are constructed by the concatenation of the basic tag (often POS of a word form) and the en-

[1] For ease of presentation we used bracketing instead of listing, i.e. SBAR(Q) should be read as SBARQ, SBAR; (WH)NP as WHNP, NP; etc.

#	Tag	Description	LPs	Cap.	#	Tag	Description	LPs	Cap.
		Noun:					**Pronoun:**		
1	ZEP	*non-personal*	ANSPC	314	20	SIMZ	*personal*[3]	NSPC	229
2	ZEQ	*personal*	ANSPC	314	21	SIMU	*demonstrative*	NSPC	157
		Verb:			22	SIMS	*interrogative*	NSPC	157
3	ET	*regular*	GTMVP	840	23	SIMD	*reflexive*	NSPC	157
4	ETU	*infinitive*	GSC	196	24	SIMB	*indefinite*	NSPC	157
5	ETK	*auxiliary*	P	8	25	SIMY	*indefinite, negative*	NSPC	157
6	ETB	*auxiliary, negative*	P	8	26	SIMP	*indefinite, universal*	NSPC	157
7	KEL	*auxiliary, desiderative*	GT	6			**Adposition:**		
8	ESM	*present participle*	GNSPC	314	27	KOM	*auxiliary nominal*	C	7
9	KSE	*past participle*	G	2	28	SHS	*preposition*	-	1
		Adjective:			29	SHZ	*conjunction*	-	1
10	SE	*regular*	P	8	30	SHD	*particle*	-	1
11	SES	*comparative*	P	8			**Interjection:**		
12	SEA	*superlative*	P	8	31	OSP	*vocative*	-	1
		Numeral:			32	OSQ	*thought*	-	1
13	SN	*cardinal*	NSPC	157	33	OSO	*emotion*	-	1
14	SNR	*ordinal*	NSPC	157					
15	SNZ	*collective*	NSPC	157	34	ELK	**Onomatopoeia**	-	1
16	SNB	*fraction*	NSPC	157	35	MOD	**Modal word**	-	1
		Adverb:							
17	US	*regular*	-	1	36	BOS	*foreign word*	-	1
18	USS	*comparative*	-	1					
19	USA	*superlative*	-	1			**Total capacity:**	-	**3844**

Table 5: The POS tagset description

coded chains of linguistic properties (LPs). Table 4 contains main LPs defined in Kazakh grammar along with their codes and *cardinalities*, i.e. a number of values they accept. Although integrating a rich set of LPs may considerably enlarge the size of a tagset, we tried to consider as many LPs as possible for the following two reasons: (i) previous research shows that increasing the size of a tagset does not necessarily decrease the tagging accuracy (Elworthy, 1995) and that for agglutinative languages omitting grammatical aspects may hurt the accuracy of *n*-gram tagging (Feldman, 2008); (ii) it is easier to reduce a detailed tagset than to re-annotate data for the missed information. Table 5 provides a detailed description of the designed tagset (not including punctuation) both qualitatively and quantitatively. The table contains a list of tags grouped by the ten major POS (in bold). For each tag we provide a set of LPs it accepts and *generative capacities*, i.e. the upper bound on a number of possible tags that can be generated from a given basic tag and the different combinations of the corresponding LPs[2]. The list of 36 basic tags was compiled following the best practices of Penn tagset design (Marcus et al., 1993), and bearing in mind the specifics of Kazakh grammar. Particularly, we broke down the major POS categories in sub-categories, in order to capture semantic distinctions and various usage patterns. For instance, negative (tag #6) and desiderative (tag #7) auxiliary verbs in conjunction with main verbs are used to mark uninflected negation (via *no* and *not*) and desiderative mood construction (via usage of *to come* in the meaning of *to want*) respectively. Similarly, auxiliary nominals (tag #27) are used as prepositional phrases such as, *in front of*, *at the top of*, etc. Also, apart from the ordinal and cardinal numerals we distinguish *collectives* (tag #15), that are used to emphasize completeness of quantities as in *both*, *all three*, etc.; and *fractions* (tag #16) as in *half*, *quarter*, etc. Finally, following classical Kazakh grammar, we treat onomatopoeias (tag #34), i.e. sound imitations as in *tic-tac* or *knock-knock*, as a distinct part of speech.

The *maximum* size of the tagset equals to the total generative capacity, or 3844 tags. However, depending on the

[2] The multiplication of cardinalities of LPs does not always give the exact number of possible tags, for there are rules that restrict certain combinations of LPs. Moreover, some LP combinations may be technically valid but semantically incorrect as they would make no sense, e.g. *bala + m + myn - I am a son of my son.* Where possible we tried to account for such exceptions, checking the combinations and providing exact numbers.

[3] Unlike any other part of speech that accepts the NSPC LP chain and must be in the third person (singular or plural) to be in any case other than nominative, personal pronouns can be in any case for any person, thus having a larger capacity.

morpheme → token	*direct speech → sentence*
token → syntactic unit	*sentence → direct speech*
syntactic unit → sentence	*list item → sentence*
sentence → paragraph	*sentence → list item*
paragraph → chapter	*dialog → sentence*
chapter → document	*sentence → dialog*

Figure 2: Structural markup hierarchy

level of granularity required for an application, some or even all LPs may be dropped or added back in, providing additional flexibility. Even the minimal tagset of 36 basic tags can be further reduced to a universal tagset (Petrov et al., 2011) that consists of 11 tags, with the first seven major POS groups being mapped to their direct equivalents, and the latter four (Interjection through Foreign word) being mapped to the *catch all* category.

Lastly, given the designed tagset the aforementioned Kazakh sentence can be tagged as follows:

Mektepke/ZEP_A0N0S0P3C3 (ZEP - non-personal noun; A0 - inanimate; N0 - singular; S0 - no possessor; P3 - 3rd person; C3 - dative case) ***bardym***/ET_G0T3M1V0P1 (ET - regular verb; G0 - not negated; T3 - past tense; M1 - indicative mood; V0 - active voice; P1 - 1st person) ***./.***

4.2 The Annotation Scheme

We have developed an XML-based annotation scheme that follows paradigms of the CES (Ide, 1998) and is convertible into the XCES standard (Ide et al., 2000). The main difference with the latter is that in our scheme the raw text and all markup types (i.e. lexical, syntactic and structural annotation; cf. Section 4) are stored in a single document. For the morpho-lexical and syntactic markups we have corresponding tags, i.e. <TOK> - *token* and <SU> - *syntactic unit*, respectively. Main linguistic characteristics, such as POS, lemmata, morpheme segmentation and syntactic labels are marked through the corresponding sub-tags and properties. All the aforementioned tags have their place in the global hierarchy of the structural markup. In turn, this hierarchy is integrated into the structure of an XML document itself. Figure 2 shows the schematic representation of the developed structural markup. A statement *A* → *B* represents *"A is contained by B"* relation.

4.3 The Annotation Tool

To ease the process of annotation we have developed a special tool that was designed as a web application with a logging and a document management system. The tool allows for (auto)saving current work and reviewing and revising the already annotated documents.

Functionality-wise the tool consists of the following three modules: (1) the syntactic module is designed to parse sentences using the syntactic tagset described in subsection 4.1, and to simultaneously mark sentence boundaries; (2) the morpho-lexical module is designed to perform a morphological analysis, and to comprise such functionalities as morpheme segmentation, POS tagging and lemma identification; (3) finally, the structural module is designed to mark up the logical structure of a document, i.e. paragraphs, dialogues, direct speech, lists, etc. Annotation of a given document is performed in the order in which we described the modules. The decision on such an order, as many other major design decisions, was made accounting for the annotators' feedback, suggestions and requests, thus making the annotation experience as convenient as possible. The validators have almost identical interface with additional functionality, such as a quick look up and correction of word-level (morphology) and sentence-level (syntax) markups. Also, both the validators and the annotators have means to correct orthography and punctuation. However, the originals of each and every annotated document are kept. In fact we have already collected data on misspellings to use in our ongoing research in spelling correction.

	Before	After
inter-annotator agreement	0.81	0.84
average MAE	0.08	0.07
average speed, words/hour	212.1	322.6

Table 6: Various characteristics of the annotation process before and after introducing the recommendation system

We have also developed a recommendation system for morphological analysis based on the already annotated data. For a given word or a given morpheme, the system generates a list of recommended markups ordered by the decreasing frequency of the previous usage. While this approach arguably has a potential to propagate errors, our experiments suggest the opposite. We have measured the inter-annotator agreement, average mean absolute error (MAE), and the annotation speed with and without the recommendation system. All measurements were taken for five annotators, who had been working with the tool for two weeks. For the experiments the annotators were given a randomly chosen news article containing about 300 words. The agreement was calculated using Fleiss' kappa (Fleiss and others, 1971). The average MAE was calculated as

$$\Delta MAE = \frac{1}{|A|}\sum_{a\in A}\frac{W - C_a}{W}$$

where A is a set of annotators, W is a number of words in a test document, and C_a is a number of words correctly annotated by the annotator a. The golden truth annotation was provided by the validators. The comparison of the measurements is given in Table 6. As we can see

Age group	I		II		III		IV		
Region	F1	M1	F2	M2	F3	M3	F4	M4	**Sum**
1	3	3	2	1	2	1	2	1	**15**
2	2	3	2	1			2	1	**11**
3	1	1	2	3	2	1	1		**11**
4	3	2		1		1			**7**
5	2	2	2	1	2	2	2	1	**14**
6	2	2	2	2	2		1	2	**13**
7	2	2	1	2	2		2	1	**12**
8	2	1	1	2	1	1	2	1	**11**
9	3	2	2	1	3	1	1	1	**14**
10	1	1	2	2	1	1	2	1	**11**
11	2	1	2	1	1		2		**9**
12	2	2	2		2	1	2	1	**12**
13	2	2	2	1	1	1	1	1	**11**
14	2	1	1	1	1	2	1	2	**11**
15	1	3		1	2				**7**
Total	**30**	**28**	**23**	**20**	**22**	**12**	**21**	**13**	**169**
Age group, %	**35%**		**25%**		**20%**		**20%**		

Table 7: The distribution of the speakers.

the inter-annotator agreement improves with the incorporation of the recommendations, while, in contrast to the error propagation assumption, the error rate slightly decreases. Moreover, we get more than 100 words/hour increase in the labeling speed. Thus, we conclude that as long as the quality of the already annotated data is high, the recommendation system will help to produce quality annotations at a higher speed. One can argue that we used a rather small sample of data to evaluate our recommendation system. However, we drew conclusions not only from the experimental results but also from opinions of validators, who confirmed that they noticed that after integrating the recommendation system annotations grew more coherent and synchronized.

Finally, let us provide a brief technical description of the tool. The design and structure of the front end is based on HTML5 and CSS3. We also use JQuery for HTML elements manipulation and various event handling. The tool can be tried out at `http://kazcorpus.kz/klcweb/annotated/#annotsample`, and the detailed information about it can be found at `http://kazcorpus.kz/klcweb/annotated/#annotdemo`.

5 Read Speech Corpus

Most of the modern speech processing systems require a large amount of audio and text data for training acoustic and language models. Depending on the type of an application required data varies from high quality microphone read speech (Garofalo et al., 2007) to conversational telephone speech (Godfrey and Holliman, 1997; Canavan and Zipperlen, 1996), from continuous speech (Garofolo et al., 1993) to connected (Leonard and Doddington, 1993) and isolated words (Pitrelli et al., 1995). In our current work, we collected a corpus of more than 40 hours of high quality microphone read Kazakh speech of 169 native speakers for the large vocabulary continuous speech recognition tasks.

5.1 Text Materials

The text materials to be uttered were carefully selected from the primary section of the corpus and divided into two parts: *sentences* and *stories*. The "sentences" part has more than 12 000 different sentences randomly and equally extracted from all of the five genre specific sections of the corpus. The sentences are chosen so that in total they contain more than 120 000 words which belong to the set of the most frequent words that cover the 95% of all the texts in the corpus. Additionally, the sentences were grouped according to their length in words. Thus, we have ten groups of sentences, so that the first group contains the sentences of length six, the second – of length seven, and so on up until the length of 15.

The "stories" part contains short online news extracted from publicistic genre section of the corpus. Each story consists of up to 300 words. All the materials were subdivided into non-intersecting sets of texts and distributed among the speakers in the following manner. Each speaker was assigned exactly 75 sentences and one story. Of the 75 sentences 50 belonged to the first five “short-sentenced” groups (10 sentences per each group), and the remaining 25 belonged to the last five “long-sentenced” groups (5 sentences per each group).

5.2 Speakers

The main criteria of a speaker selection were the following: a region where (s)he learned Kazakh or spent most of his/her life; age; gender; and the ability to read Kazakh.

The first criterion helped us to capture various accents attributed to speakers’ settlement both local and external. From the regional perspective we divide the speakers into 15 groups: 14 domestic (one per each administrative region, i.e. “oblast”, of Kazakhstan) and one abroad (all foreign countries). Furthermore, the speakers are divided into the following four age groups (not including children and school students): (i) 18-27 years, (ii) 28-37 years, (iii) 38-47 years, (iv) 48 years and above. We did not strictly balance the speakers by their gender due to the difficulties in finding the volunteers, but still tried to choose no more than three speakers of the same gender per one age-regional group. A female-to-male distribution of speakers is 57% to 43%, respectively.

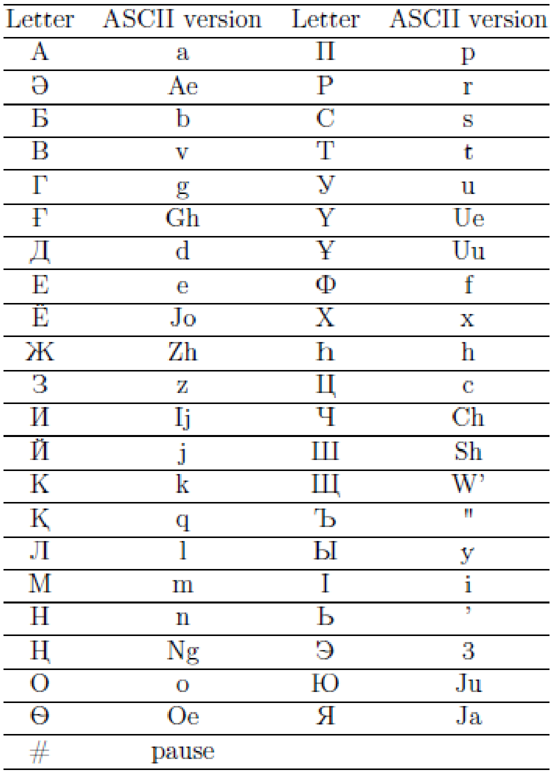
The other important criterion is the ability to read Kazakh, since not all of the interviewees could read in Kazakh sufficiently fluent, which is a common issue in a bilingual country such as Kazakhstan. Additionally, we kept a record of the speakers’ education, i.e. whether they attended and graduated from a university, or graduated from a school or a college without attending any universities.

The speakers were encoded using the following scheme: <Region><Gender><Year of birth><Initials><Education>, where “Region” holds the values in the range of [1-15], “Gender” – F or M, “Year of birth” – the last two digits of a year of birth, ”Initials” – initials of a name followed by a surname, “Education” – 1 for school, 2 for college, and 3 for university, e.g. 06F70ZK3.

In total, we have recorded 169 speakers. Table 7 presents a distribution of the speakers across the age, gender and regional groups. The blank spots show the speaker profiles that we could not recruit. Mostly, these cases correspond to the distant regions and elder male groups.

5.3 Recording Setup

The actual recording sessions took place in a sound-proof studio of the university with the assistance of a sound operator. Before the recordings, the speakers were instructed, documented and given some time to prepare, as well as asked to fill in the copyright transfer form for the audio data with their voice. They were not constrained on the manner, speed or time except for the correctness of reading. The average time for a recording session per speaker was about 40-45 minutes, though there were cases that lasted for two hours. Audio data was captured using a professional vocal microphone Neumann TLM 49 and digitized by LEXICON I-ONIX U82S sound card. The format of the recorded audio files is 44.1 kHz 16-bit PCM-encoded mono WAVE file format. All the recorded audio files were manually post-processed to have each utterance (sentences and stories) in a separate file and in the corresponding directories. The size of the speech corpus is about 8.5 GB on disk. A collective duration of the audio files is more than 40 hours long.



Letter	ASCII version	Letter	ASCII version
А	a	П	p
Ә	Ae	Р	r
Б	b	С	s
В	v	Т	t
Г	g	У	u
Ғ	Gh	Ү	Ue
Д	d	Ұ	Uu
Е	e	Ф	f
Ё	Jo	Х	x
Ж	Zh	Һ	h
З	z	Ц	c
И	Ij	Ч	Ch
Й	j	Ш	Sh
К	k	Щ	W’
Қ	q	Ъ	"
Л	l	Ы	y
М	m	І	i
Н	n	Ь	’
Ң	Ng	Э	3
О	o	Ю	Ju
Ө	Oe	Я	Ja
#	pause		

Figure 3: ASCII version of the Kazakh letters.

5.4 Transcription and Annotation

Each audio file is provided with its corresponding orthographic transcription and TIMIT-style word-level segmentation, as well as morpho-syntactic annotation files. Both the transcript generation and the annotation were performed manually by trained linguists. The transcription files contain the exact orthographic transcriptions of the utterances, which may differ from the original text. For example, the numbers, abbreviation, foreign words and dates are expanded depending on how they were uttered by the speakers. In addition, the transcription of the stories have the sentence boundaries labeled with <s> and </s> tags. For the segmentation we used WaveSurfer (2013), an open-source tool for sound visualization and manipulation, which supports TIMIT word-level transcription format. Although, it supports Unicode, it does provide a proper support for Kazakh symbols. Therefore, we used an ASCII version of the Kazakh letters depicted on Figure 3. Also, we used the # symbol for the pauses and silence, and ˆ symbol for other non-speech events.

6 Conclusion and Future Work

In this work we have described the design and compilation process of the Kazakh Language Corpus. KLC is oriented for a wide range of users and we believe that it will be a valuable tool for research communities, especially given that a portion of the data has been labeled with multiple levels of annotation, including word-level segmentation of audio information. We are already using the annotated data in our initial experiments in morpheme segmentation and error correction.

One can explore the corpus through the website (`http://kazcorpus.kz`) that was designed to provide the best experience in the analysis of data.

For the future work we plan to use the corpus as a research tool to tackle the following problems: (i) automatic part of speech tagging, (ii) morphological disambiguation, (iii) statistical machine translation. For the latter we have already started collecting parallel text in Russian and English.

Acknowledgments

We would like to thank the Ministry of Education and Science of the Republic of Kazakhstan for supporting this work through a grant under the 055 research program.

We express our gratitude to Dr. A. Sharipbayev for his valuable advice on methodology of constructing the read-speech corpus.

We also would like to sincerely thank our validators and annotators: Bobek A., Asemgul R., Aidana Zh., Nazerke G., Nazym K., Ainur N., Sandughash A., Zhuldyzai S., Dinara O., Aigerim Zh. The annotation and validation work they have done helped a great deal in designing tagsets. This work would not be possible without their contribution.

References

Yesim Aksan, Mustafa Aksan, Ahmet Koltuksuz, Taner Sezer, Umit Mersinli, Umut Ufuk Demirhan, Hakan Yilmazer, Gulsum Atasoy, Seda Oz, Ipek Yildiz, and Ozlem Kurtoglu. 2012. Construction of the turkish national corpus (tnc). In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

L. Al-Sulaiti and E.S. Atwell. 2006. The design of a corpus of contemporary arabic. *International Journal of Corpus Linguistics*, 11:135–171.

Gulila Altenbek and WANG Xiao-long. 2010. Kazakh segmentation system of inflectional affixes. In *Joint Conference on Chinese Language Processing*, pages 183–190. CIPS-SIGHAN.

Juri Apresjan, Igor Boguslavsky, Boris Iomdin, Leonid Iomdin, Andrei Sannikov, and Victor Sizov. 2006. A syntactically and semantically tagged corpus of russian: State of the art and prospects. In *The fifth international conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy.

Vt. Baisa and Vt. Suchomel. 2012. Large corpora for turkic languages and unsupervised morphological analysis. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 28–32, Istanbul, Turkey. European Language Resources Association (ELRA).

Lou Burnard, editor. 2007. *Reference Guide for the British National Corpus*. Research Technologies Service at Oxford University Computing Services, February.

Alexandra Canavan and George Zipperlen. 1996. Callhome japanese speech.

K.-j. Chen. 1996. Sinica corpus: Design methodology for balanced corpora. In B.S. Park and J.B. Kim, editors, *Proceeding of the 11th Pacific Asia Conference on Language, Information and Computation*, pages 167–176, Seoul. Kyung Hee University.

CLMCRK. 2009. The corpus of kazakh language. [visited 29/08/2012].

David Elworthy. 1995. Tagset design and inflected languages. In *In EACL SIGDAT workshop iFrom Texts to Tags: Issues in Multilingual Language Analysis*, pages 1–10.

Anna Feldman. 2008. Tagset design, inflected languages, and n-gram tagging. *Editors: Paul Robertson and John Adamson*, 3(1):151.

J.L. Fleiss et al. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Winthrop Nelson Francis and Henry Kučera. 1979. *Manual of information to accompany a standard corpus of present-day edited American English, for use with digital computers*. Brown University, Department of Lingustics.

John Garofalo, David Graff, Doug Paul, and David Pallett. 2007. Csr-i (wsj0) complete linguistic data consortium, philadelphia.

John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallett, Nancy L. Dahlgren, and Victor Zue. 1993. Timit acoustic-phonetic continuous speech corpus.

JJ Godfrey and E Holliman. 1997. Switchboard-1 release 2. *Linguistic Data Consortium, Philadelphia*.

Jan Hajič and Barbora Hladká. 1998. Tagging inflective languages: prediction of morphological categories for a rich, structured tagset. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*

- *Volume 1*, ACL '98, pages 483–490, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jirka Hana and Anna Feldman. 2010. A positional tagset for russian. *Proceedings of LREC-10. Malta.*

Nancy Ide and Catherine Macleod. 2001. The american national corpus: A standardized resource of american english. In Paul Rayson, Andrew Wilson, Tony McEnery, Andrew Hardie, , and Shereen Khoja, editors, *Proceedings of the Corpus Linguistics 2001 Conference*, pages 274–280. Lancaster University (UK).

N. Ide, P. Bonhomme, and L. Romary. 2000. Xces: An xml-based standard for linguistic corpora. In *Proceedings of the Second Annual Conference on Language Resources and Evaluation*, pages 825–830, Athens.

Nancy Ide. 1998. Corpus encoding standard: Sgml guidelines for encoding linguistic corpora. In *Proceedings of the First International Language Resources and Evaluation Conference*, pages 463–70. Citeseer.

H. Kim. 2006. Korean national corpus in the 21st century sejong project. language corpora:their compilation and application. In *Proceedings of the 13th NIJL International Symposium*, pages 49–54, Tokyo, March.

Geoffrey Leech, Roger Garside, and Michael Bryant. 1994. Claws4: the tagging of the british national corpus. In *Proceedings of the 15th conference on Computational linguistics-Volume 1*, pages 622–628. Association for Computational Linguistics.

R. Gary Leonard and George Doddington. 1993. Tidigits.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: the penn treebank. *Comput. Linguist.*, 19(2):313–330, June.

Akmaral Mukan. 2012. *A Learner's Dictionary of Kazakh Idioms*. Georgetown University Press.

Kemal Oflazer, Bilge Say, Dilek Zeynep Hakkani-Tür, and Gökhan Tür. 2003. Building a turkish treebank. In *Treebanks*, pages 261–277. Springer.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2011. A universal part-of-speech tagset. *Arxiv preprint ArXiv:1104.2086*.

John F. Pitrelli, Cynthia Fong, Suk H. Wong, Judith R. Spitz, and Hong C. Leung. 1995. Phonebook: A phonetically-rich isolated-word telephone-speech database. In *ICASSP*, volume 1, pages 101–104.

Ruscorpora. 2003. Russian national corpus. [visited 29/08/2012].

WaveSurfer. 2013. `http://www.speech.kth.se/wavesurfer/`. Accessed: 2013-03-30.

Wikipedia. 2012. Kazakh alphabet. [visited 29/08/2012].