Identification of Basic Phrases for Kazakh Language using Maximum Entropy Model

Gulila Altenbek*,+ Xiaolong Wang* Gulizhada Haisha+

*School of Computer Science and Technology, Harbin Instituteof Technology,150001,China.

+College of Information Science and Engineering, Xinjiang University,830046, China.
+The Base of Kazakh and Kirghiz Language of National Language Resource Monitoring and Research Centre Minority Languages, Xinjiang, 830046, China.

gla@insun.hit.edu.cn, gla@xju.edu.cn, wangxl@insun.hit.edu.cn

Abstract

This paper proposes the definition, classification and structure of the Kazakh basic phrases, and sets up a framework for classifying them according to their syntactic functions. Meanwhile, the structure of the Kazakh basic phrases were analyzed; and the determination of the Kazakh basic phrases collocation and extraction of the Kazakh basic phrases based on rules were followed. The Maximum Entropy (ME) model uses for the identification of the phrases from texts and achieved a result of automatic identification of Kazakh phrases with an accuracy of 78.22% based on rules System and additional artificial modification. Design feature of this ME model join rely on templates of Kazakh Word, part of speech, affixes. Experimental results show that the accuracy rate reached 87.89%.

1 Introduction

Automatic phrase identification is an important task in natural language processing. A phrase is a group of words that work together. Phrase recognition is a grammatical unit agent between words and sentences in natural language processing. Phrase identification Parser has been developed for different languages, for example, the Church's Base NP Recognition for English (Church, 1988). The rule-based Model and Maximum Entropy Model (ME) are the most commonly used technology for phrase representation and parsing.

Kazakh Language belongs to the Turkish Language group in the Altaic language family. It is an agglutinative language with word structures formed by adding derivational or inflectional affixes to root words. Phrase identification is also an indispensable part for Kazakh information processing. In the past a few year, we have put forward methods for Kazakh morphological analysis, which includes stem extraction, part of speech(POS) tagging, spelling check, etc. Recently, we are working on syntax parsing, analysis of phrase structure, automatic identification of phrase and in-depth analysis of sentence structure.

Kazakh phrases are syntactic units consisting of two or more than two words. The phrases can be classified into two categories, which are free phrase and fixed phrase. We are exploring methods which are more suitable for shallow syntactic parsing of Kazakh according to the nature of Kazakh language. The research includes a systematic study on information regularity and disambiguation of the Kazakh phrase, and automatic recognition of basic phrases of Kazakh language. We have developed a rule-based method for the automatic recognition of Kazakh basic phrases, and automatic identification of verb phrase, noun phrase and adjective phrase based on maximum entropy in Kazakh language at the same time. Moreover, the ambiguity of structures is also resolved based on rules.
This study solves the problem of Kazakh phrase recognition by providing some effective methods. This sets up a basis for further syntactic process and tree bank building. This research also provides a way to build database for various fields like knowledge acquisition, syntactic understanding, Chinese-Kazakh machine translation, the process of large-scale corpus, etc.

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 1007–1014, Dublin, Ireland, August 23-29 2014.

In this paper, our work focuses on identifying noun phrases, adjective phrase and verb phrases, which are the most difficult aspects of Kazakh phrase recognition analysis. This is achieved by using rules are ME method.

2 Related work

There are a variety of techniques used for phrase recognition, which include rule-based technique, statistical technique, and a combination of them. Church's (1988) approach used manual or semi-automatic annotation phrase corpus as a training corpus. Another popular method is to use a Chunk parsing for statistics model to determine the boundary (Koeling, 2000). Chunk parsing was first introduced by Abney (1991), which is one of the most widely used syntactic parsing methods. The main idea of chunk parsing lies in seeking the appropriate breakthrough point, and decomposing the full parsing problems into a syntax topology statistical structure and syntactic relations. Zhao and Huang (1998) are pioneers in Chinese phrase studies; Tsinghua University had also completed its TCT (Tsinghua Chinese Treebank) for Chinese (Zhou, 2004). The method has been also applied into studies of other languages, such as Kazakh Base NP recognition (Altenbek et al, 2009), and Uyghur Base VP Recognition by CRF (Mamatmin et al, 2012).

Maximum Entropy was first introduced to NLP area by Berger et al (1996) and Della Pietra et al. (1997). Maximum Entropy is an extremely flexible technique for linguistic modelling. It can use a virtually unrestricted and rich feature set in the framework of a probability model. It is a conditional, discriminative model and allows mutually dependent variables (Ratnaparkhi, 1999).

3 Kazakh Phase Parsing

3.1 Kazakh Morphology

Morphological analysis is an important task in natural language processing research. It was developed for different languages, included English (Porter, 1980), Finnish (Karttunen, 1983), Turkish (Oflazer, 1994; Gülşen, 2004), and Arabic (Beesley, 1996).
Comparing with other languages, the Kazakh morphological system uses a large number of suffixes and a small number of prefixes. Every word has a root, or a stem (Milat, 2003;Zhang 2004). The basic Kazakh phrase is an adjacent and non-nested phrase which does not contain recursive structure.

3.2 The Categories of Kazakh Phrase

Parsing is one of the most basic and fundamental components in natural language processing. Chunk parsing intends to obtain a fragment without thinking deeply.

A Kazakh phrase is composed of two or more than two words which connected with meaning and grammatical structure. There is only a core word in a Kazakh phrase. In the case of Kazakh, Kazakh phrases can be divided into fixed phrases and temporary phrases by the meanings of the phrases.

Abney propose the first complete description of lexical chunks system. In this study the basic phrase chunks base was found according to Abney's system. The five most common phrase in Kazakh are

NO.	Category	Explanation	Example (Kazakh)	Example (English)
1	NP	noun phrase	التىن كۇز	The golden autumn
2	VP	verb phrase	مۇراتقا جەتۋ	Achieve dreams
3	ADJP	adjective phrase	تاپ – تازا	Very clean
4	NUMP	Numeral phrases	سەگىز توعىز مىڭ	Eight & nine thousand
5	ADVP	Adverb phrase	مڭ الدىندا عى	The front of

Table 1. Part of Kazakh phrase categories.

noun phrase, verb phrase, adjective phrase, Numeral phrases, Adverb phrase as shown in table 1. Kazakh language is rich in the external morphology which shows prominent in phrase structure.

3.3 The Basic Kazakh phrase mark specification

Basic Kazakh phrase marks both its own attribute, for example part of speech, stems and affixes, and types of phrase. We used IOB Tagging to mark the start and end of chunks.

Basic Kazakh phrase	start of chunks	Inner tag of chunks	Out tag of chunks
noun phrase	B-NP	I-NP	O
verb phrase	B-VP	I-VP	
adjective phrase	B-ADJP	I-ADJP	
Adverb phrases	B-ADVP	I-ADVP	
Numeral phrase	B-NUMP	I-NUMP	

Table 2. The Basic Kazakh phrase IOB Tagging.

4 Statistics and Analysis of Kazakh Phrase Structure

Referring to modern Kazakh grammar (Milat, 2003; Dingjing Zhong. 2004), the basic rules of phrase structure of Kazakh language was summarized. The phrase structures are extracted from the corpus, and a set of rules are created based on it as well.

In the representation of basic phrase structures, the following part of speech tagging symbols are used in XML documents of Kazakh corpus: v (verb), n. (noun), adj. (adjective), num. (number), adv. (adverb), pron. (pronoun), ono. (onomatopoeia), int.(interjections), conj. (conjunction), part. (partical). The Kazakh phrases Structure divided by the function of phrases in our system are shown below.

Kazakh verb phrase structure:

1) n+v; 2) v+v; 3) adv+v; 4) adj+v; 5) v+adv; 6) v+v+v; 7) pron+v; 8) n+part+v; 9) n+conj+v; 10) ono+v; 11) int+v; 12) v+part+v; 13)v+part; 14) v+conj+v; 15)pron+part+v.

Kazakh noun phrase structure:

1) n+n; 2) n+conj+n; 3) pron+conj+pron; 4) pron+n; 5) adj+conj+adj; 6) adj+n; 7) adj+adv+n;
8) num+n; 9) v+n; 10) [ ]+n.

Kazakh adjective phrase structure:

1) adj+n; 2) adj+v; 3) adj+n+v; 4) pron+adj; 5) adv+adj+n; 6) adj+adj+n; 7) num+adv+n;

Collocations, like v+adv, n+part+v, pron+adv, v+part+v, v+part, also exist in other phrase except verb phrase. These conditions easily cause ambiguity.

5 Rule-based phrase tagging

Kazakh language has two characteristics that have to be taken into account: agglutinative morphology and rather free word order with explicit case marking.

The corpus we used in this process has been already segmented. The way we extracted stem and affix was briefly mentioned in the paper. In this paper we used the segmented results of early work, as it is not the core part of the algorithm.

Input: word segmentation (extraction stem and affix) and POS tagged corpus (test.xml);

Output: First: Phrase tagged file; Second: Phrase file;

Based on the basic rules of phrase, we have done extraction of phrases from POS tagged Kazakh corpus. The extraction process is as follows:

(a) First roughly segmented XML corpus. The common segmentation marks include semicolon, comma, full stop, exclamation mark, question mark.

(b) For the segmented data, we extract the three elements of basic phrase: part of speech (POS), affix, and the word.

(c) Look for the matched rule in the rule set. If found, save the basic phrase. Otherwise go back step 1. According to combination rules of basic Kazakh phrase, basic phrase was extracted from corpus and modified by manual work. The correct combination of basic Kazakh phrase was marked.

6 Analysis of Kazakh phrase structure ambiguity

Ambiguity computer analysis of language structure has been one of the difficulties problems. This article from the delimitation ambiguity and structural relationship is to study two aspects of phrase structure ambiguity.

One of the difficulties in Kazakh phrase research is the phrase disambiguation problem. Ambiguous reasons is word POS ambiguity, phrase boundaries is not easy to determine, POS with the same sequence, E.g. there are five ambiguous forms:

(1) VD form (v + adv)

Eg.1a : adv/تومەندەۋ v/قابىلداۋى is verb phrase. (Admission to reduce)

Eg.1b : adv/مەكشە v/قابىلدۇى is adverb phrase. (Admission to more than)

(2) ND for (n+adv, pron+adv)

Eg.2a : adv/جاڭالاۋ n/كيىمىن is verb phrase.(Change a new clothes)

Eg.2b : adv/كەرەمەت n/ناتيجەسى is adverb phrase.(Good record)

(3) NPV form （ n+part+v, pron+part+v)

Eg.3a : v /ۇيرەنۋ part /تۇرالى n /بىرلىك is verb phrase.(Learn about unity)

Eg.3b : v /دەى part / عانا n / اسان is noun phrase.(only Ashan)

(4) VPV form (v+part+v)

Eg.4a : v /كەتتى part /دە v / كەلدى is verb phrase.(came then left)

Eg.4b : v / تۇسىنۋ part /تۇرالى v / زەرتتەۋ is adverb phrase. (Relevant research to understand)

(5) VP form (v+part)

Eg.5a : part / بۇرىندىاۋ v / سويلەۋدەن is verb phrase.(Speaking before)

Eg.5b : part / جونىندە v /رەتتەۋ is verb phrase.(Organize the relevant)

For these ambiguities, we can't simply use the rules to match ways to eliminate, but rather to use maximum entropy model to solve the problem.

7 Kazakh Phrase Identification based Maximum Entropy Model

Maximum Entropy Model is an effective machine learning model which is proposed to solve the POS tagging problem, it using ME model is the ability to incorporate various features into the conditional probability. The Kazakh phrase recognition task is presented as follow.

The entropy model P: $H(p) \equiv -\sum_{x,y} p(x,y)\log(x,y)$ (1)

Note: X represents the environmental context words to be marked and y is the output.

Maximum Entropy Model： Such a model can be shown to have the following form:

$$p^* = \arg\max_{p \in C} H(p) \quad (2)$$

Goal: select a distribution p from a set of allowed distributions that maximizes H(y|X).

7.1 Feature defined

Kazakh language is an agglutinative language with word structures formed by adding derivational, inflectional affixes or suffixes to root words. The features include words, part of speech (POS), inflectional affixes of the training corpus. It seems that the features are naïve. However, these three kinds of features are the most important components of Kazakh language, and they reflect the characteristic of Kazakh language.

According to its own characteristics of a Kazakh, this feature space is defined as follows:

(1) *the word*, including the current word, the previous word and next word.

(2) *part of speech(POS)*, including the part-of-speech types of the current word, previous word and next word.
(3) *Affix ingredients*, including the current word and the word about the additional ingredient infomation.
(4) *Phrase tag* that contains the current word and the words to the right and the left two words Phrase marker.

This rule-based approach was applied to generate the maximum entropy model training corpus. Based on Kazakh linguistics, the atomic feature space is as shown in table 3.

Feature tag	Feature explanation	Feature tag	Feature explanation
W(-1)	previous one word	POS (-2) POS (-1)	POS of previous two word and POS of previous one word
W(0)	the current word	POS (-1) POS (0)	POS of previous one word and POS of the current word
W(+1)	next one word	POS (0) POS (+1)	POS of the current word and POS of next one word
W(-1) W(0)	previous one word and the current word	POS (+1) POS (+2)	POS of next one word and POS of next two word
W(0) W(+1)	the current word and next one word	POS (-2） POS (-1) POS (0)	POS of previous two word and POS of previous one word and POS of the current word
W(-1) W(0） W(+1)	previous one word and the current word and next one word	POS (-1) POS (0) POS (+1)	POS of previous one word and POS of the current word and POS of next one word
POS (-2)	POS of previous two word	POS (0) POS (+1) POS (+2)	POS of the current word and POS of next one word and POS of next two word
POS (-1)	POS of previous one word	Affix(-1)	affix of previous word
POS (0）	POS of the current word	Affix(0)	affix of current word
POS (+1)	POS of next one word	Affix(1)	affix of next one word
POS (+2)	POS of next two word		

Table 3. Atomic feature templates.

7.2 Feature selection

Basic phrases with statistical model recognition need to select a high correlation, and the Kazakh language features to train with good effect. Establish model based on rule of the language, this work selected feature through templates. After several rounds of experimental debugging, then used artificial selection, twenty one templates were selected for Kazakh verb phrase, only considered important features. According to each one's feature, templates were defined as follow.

No.	template	No.	template	No.	template
1	LPos,Cpos,RPos	8	CVP,RVP,RRVP	15	CWord,RWord
2	LLPos,Lpos,CPos	9	LVPCPosRVP	16	LPos,LVP
3	CPos,Rpos,RRPos	10	LPos, LAffix, LVP	17	RWord,RPos
4	CPos,CAffix,RPos	11	Cpos, CAffix, CVP	18	RPos,RVP
5	LPosLAffixCPos	12	CWord,RWord,RAffix	19	CPos,RPos
6	LVP,CVP,RVP	13	CWord,CPos	20	LPos,CPos
7	LLVP,LVP,CVP	14	LWord,LPos	21	LWord,LVP

Table 4. Combined feature of Kz Base VP.

In order to get the best template, this work structured and processed six template based on Table 4.

Each information function valued in the context of current word, combine the various function values into the premise of features, got the characteristics of the movement through the word tag, then it can extract features.

Template A: [RRPos, RRVP, RWord, RAffix, RPos, RVP, CPos, CVP, CWord, CAffix, LLPos, LLVP, LWord, LAffix, LPos, LVP] Observation of effects of all the words in the feature space on the result of the experiment.

Template B: [CPos, CVP, CWord, CAffix, LLPos, LLVP, LWord, LAffix, LPos, LVP] Observation of effects of left side two words of the candidate word on the result of the experiment.

Template C:[RRPos, RRVP, RWord, RAffix, RPos, RVP, CPos, CVP, CWord,CAffix] Observation of effects of right side two words of the candidate word on the result of the experiment.

Template D:[RWord, RAffix, RPos, RVP, CPos, CVP, CWord,CAffix, LWord, LAffix, LPos, LVP] Observation of effects of each side one word of the candidate word on the result of the experiment.

Template E:[RWord, RAffix, RPos, RVP, CPos, CVP, CWord, CAffix, LLPos, LLVP, LWord, LAffix, LPos, LVP] Observation of effects of left side two words and right side one word of the candidate word on the result of the experiment.

Template F:[RRPos, RRVP, RWord, RAffix, RPos, RVP, CPos, CVP, CWord, CAffix, LWord, LAffix, LPos, LVP] Observation of effects of left side one word and right side two words of the candidate word on the result of the experiment.

We selected some corpus from *Xinjiang Daily* tested on six features above, we got different influences of different characters. It shows that the C and F template give us the most highest result, namely the two words on the right have the biggest influence to the result. It proves Kazakh verb phrases are commonly at the end of the sentence.

7.3 General threshold selection

There are two general feature selection methods: incremental feature selection and feature selection of based on frequency threshold. The frequency is greater than a threshold value equal to a characteristic. Through repeating them many times, the frequency threshold value was characterized k = 5, characterized in that the use of the frequency characteristic is greater than 5.

8 Kazakh Phrase Recognition System

Kazakh phrase recognition system, which based on Maximum Entropy Model, consists of four modules, namely, pre-processing module, training module, Feature selection module, identification module. System training process as shown flow as figure 1.

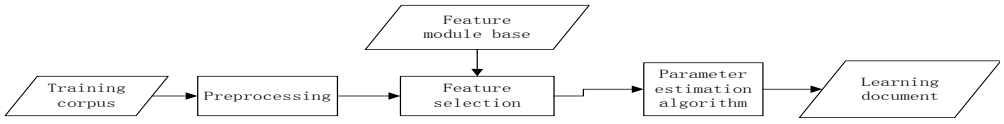

Figure 1. Training data flow diagram.

System testing process as shown flow as figure 2.

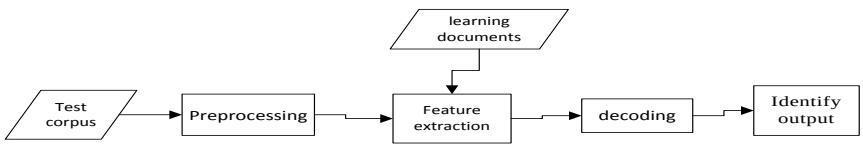

Figure 2. Testing data flow diagram.

The Kazakh basic verb phrase recognition results such as shown figure 3:

```
‹kaza_xml›
 ‹article id="&lt;*1b_1jN01001*&gt;"›
  ‹title›التاي شارۋالارى شەتەلگە شعسپ جۇمىس ىستەپ، كىرىس تاپتى ‹/title›
  ‹paragraph id="1"›
   ‹word pos="n" stem="ٴتىلشى" affix="/ىمىز" var="0" vp="O"›ٴتىلشىمىز‹/word›
   ‹word pos="n" stem="چىن" affix="" var="3" vp="O"›چىن‹/word›
   ‹word pos="n" stem="چىن" affix="" var="3" vp="O"›چىن‹/word›
   ‹word pos="n" stem="التاي" affix="/دان" var="4" vp="B"›التايدان‹/word›
   ‹word pos="v" stem="حابارلايدى" affix="" var="0" vp="I"›حابارلايدى‹/word›
   ‹punction›·‹/punction›
   ‹word pos="n" stem="قۇربان" affix="" var="5" vp="O"›قۇربان‹/word›
   ‹word pos="n" stem="ايت" affix="" var="1" vp="O"›ايت‹/word›
   ‹word pos="v" stem="كەل" affix="/ۋ" var="0" vp="B"›كەلۋ‹/word›
   ‹word pos="prep" stem="جونىندە" affix="" var="0" vp="I"›جونىندە‹/word›
   ‹punction›،‹/punction›
```


Figure 3. The Kazakh language basic verb phrase recognition.

By following a comprehensive analysis of Kazakh words, the following is the Kazakh shallow parsing process:

（1） Sentence :

قوڭىر كۇز كەلىپتى، قامبار سول جەرگە، قوي باعىپ كەلسە، ازىناعان كۇزدىڭ جەلى سوعىپ تۇرىپتى.

Golden autumn is coming, Hambar came to the place whish has very strong winds together with sheep.

（2） POS:

قوڭىر/n كۇز/n كەلىپتى/v، قامبار/n سول/pron جەرگە/n، قوي/n باعىپ/v كەلسە/v، ازىناعان/adj كۇزدىڭ جەلى/n سوعىپ/v تۇرىپتى/v.

（3） Phrase POS:

جەرگە pron/اسول n/ (Hamubar(قامبار]ا VP،[(is coming) v/كەلىپتى NP[(autumn) n/كۇز (Golden) n/قوڭىر]]

AP[(the جەلى n/اكۇزدىڭ adj/ازىناعان]]]ا VP،[(came) [v/كەلسە [v/VPباعىپ (sheep) n/اقوي]ا NP[RP[،n/

.VP (blowing.[v/تۇرىپتى VP[v/سوعىپvery strong winds) NP[n/

9 Experiment Results and Analysis

9.1 Data set

In this paper, according to the data set, we used the data of January 2008 of the *Xinjiang Daily* (Kazakh version) corpus. The corpus consists of the raw texts and the POS tagged XML format texts, experiments were done for phrase extraction.

9.2 Experiment results

The experiments of the accuracy rates are evaluated using as follow standard evaluation measures:

Precision: $$P = \frac{a}{b} \times 100\% \qquad (3)$$

Recall $$R = \frac{c}{d} \times 100\% \qquad (4)$$

F-measure $$F = \frac{2 \times R \times P}{R + P} \qquad (5)$$

Note: a is number of correctly identified phrases. b is number of identified phrases. c is number of all phrases, d is number of should correct identify.

In the test corpus, there are 3000 correct tagged sentences as training data, and other 1000 sentences are for the test.

Method	Precision (%)	Recall (%)	F-measure (%)
Rule	78.22	70.01	85.25
ME	87.89	83.13	87.46

Table 5. Phrase recognition test.

10 Conclusion

This paper provided solution for identifying Kazakh basic phrases. We have tried rule-based and the maximum entropy methods. The Kazakh words, part of speech, affixes context information are used to design template of features for maximum entropy model. Based on statistical methods, higher accuracy could be obtained in the test, but it was requires more training data.

The recognition of basic Kazakh phrase could simplify sentence structure, reduce the difficulty of syntactic analyzer. This work put maximum entropy model into recognition of basic Kazakh phrase. However, there are still space for improvement on scale and accuracy rate comparing to English and Chinese. In the future, our work will focus on completing of corpus and other models.

Acknowledgments

This work is funded by the Natural Science Foundation of P.R. China (NSFC)(No.61363062, No. 61063025 and No.61272383), Science and Technology Research and Development Funds of Shenzhen City (No. JC201005260118A).

Reference

Church K. *A stochastic parts program and noun phrase parser for unrestricted text*. 1988. In Proceedings of the Second Conference on Applied Natural Language Processing. Texas, USA. 19(8):136-143.

Rob Koeling . *Chunking with Maximum Entropy Models*. 2000. Proceedings of CoNLL-2000 and LLL-2000. 109(15):139-141.

Steven Abney. *Parsing by chunks*. 1991. Dordrecht: Kluwer Academic Publishers. 257-278.

Zhao Jun and Huang Changning. 1999. *Chinese basic noun phrase structure analysis model*, Computer science . 22(2):141-146.

Qiang Zhou. 2004. Annotation scheme for Chinese Treebank, Journal of Chinese Information Processing. Vol 18(4):1-8.

Gulila Altenbek, Ruina-Sun. 2010. *Kazakh Noun Phrase Extraction based on N-gram and Rules*, International Conference on Asian Language Processing (IALP2010). Harbin, China. 305-308.

Gulila A. and Dawel, A. and Muheyat, N. 2009. *A Study of Word Tagging Corpus for the Modern Kazakh Language*, Journal of Xinjiang University. 26(4):394-401.

Zulpiya Mamatmin et al, 2012.*Uyghur Base Verb phrases Recognition* . A master's degree thesis, Beijing university of posts and telecommunications.

Adam Berger, Stephen Della Pietra, and Vincent Della Pietra. 1996. *A Maximum Entropy Approach to Natural Language* ,Processing Computational Linguistics, 22(1):39-71.

Adwait Ratnaparkhi. 1999. *Learning to parse natural language with maximum entropy models*. Machine Learning, 341(3):151-176

Porter, M.F. 1980. An algorithm for suffix stripping, Program, 14(3):130−137.

Karttunen, Lauri. 1983. KIMMO: A general morphological processor. Texas Linguistic Forum, 22:163–186.

Kemal Oflazer. 1994. Two-level description of Turkish morphology. Literary and Linguistic Computing, 9(2):137-148.

Gülşen, E. and Eşref, A. 2004. An affix stripping morphological analyzer for Turkish, Proceedings of the International Conference on Artificial Intelligence and Application, Austria, 299-304.

Beesley, K.R. 1996. Arabic finite-state morphological analysis and generation. In COLING-96, Copenhagen, 89-94.

Milat, A. 2003. Modern Kazakh language, Xinjiang People's press, China.

Dingjing Zhang. 2004. Practical Grammar of Modern Kazakh Language. Beijing: Central University for Nationalities Press.