Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2016"

Moscow, June 1–4, 2016

DEVELOPING A POLYSYNTHETIC LANGUAGE CORPUS: PROBLEMS AND SOLUTIONS[1]

Arkhangelskiy T. A. (tarkhangelskiy@hse.ru),
Lander Yu. A. (yulander@hse.ru)

National Research University Higher School of Economics,
Moscow, Russia

Although there exist comprehensive morphologically annotated corpora for many morphologically rich languages, there have been no such corpora for any polysynthetic language so far. Developing a corpus of a polysynthetic language poses a range of theoretical and practical challenges for corpus linguistics. Some of these challenges have been partly addressed when developing corpora for languages with extensive morphological inventories and numerous productive derivation models such as Turkic or Uralic, while others are unique for this kind of languages. As we are currently working on a corpus of the polysynthetic West Circassian language, we had to identify these challenges and propose theoretical and practical solutions. These include the tokenization problem, which involves delimiting morphology from syntax, the problem with lemmatization and part-of-speech tagging, and a number of glossing and search issues. The solutions proposed in the paper are partly implemented and will be available for public testing when the preliminary version of the corpus is released.

Keywords: polysynthesis, Adyghe, West Circassian, language corpora, morphology

[1] This paper is based on our research supported by Russian Foundation for Basic Research (grant No. 15-06-07434a).

РАЗРАБОТКА КОРПУСА ПОЛИСИНТЕТИЧЕСКОГО ЯЗЫКА: ПРОБЛЕМЫ И РЕШЕНИЯ

Архангельский Т. А. (tarkhangelskiy@hse.ru),
Ландер Ю. А. (yulander@hse.ru)

Национальный исследовательский университет
«Высшая школа экономики», Москва, Россия

Несмотря на то, что в настоящее время существует множество морфологически размеченных корпусов для языков с богатой морфологией, до сих пор не было создано ни одного корпуса полисинтетического языка, который бы учитывал необходимую морфологическую информацию. Разработка корпуса для таких языков ставит перед корпусным лингвистом ряд нетривиальных теоретических и практических задач. Некоторые из них в меньшем объёме встречались и частично решались ранее при создании корпусов языков с большими морфологическими системами и обилием продуктивных деривационных моделей, например, тюркских или уральских языков. Однако многие из этих проблем уникальны для полисинтетических языков. В ходе работы над созданием корпуса полисинтетического адыгейского языка мы обрисовываем эти проблемы и предлагаем ряд теоретических и практических решений. Описываемые проблемы включают в себя токенизацию (связанную с нечёткой границей между синтаксисом и мофологией), лемматизацию и морфологическую разметку, а также ряд вопросов, связанных с глоссированием и поиском в корпусе. Предлагаемые решения частично реализованы и будут доступны для тестирования в пилотной версии корпуса.

Ключевые слова: полисинтетизм, адыгейский язык, языковые корпуса, морфология

1. Introduction

The main feature that distinguishes a corpus of language from a mere collection of texts is its annotation. While it is possible to add various levels of annotation to a corpus, including syntactic parsing, semantic labeling, anaphora resolution, etc., what is absolutely necessary for morphologically rich languages is morphological annotation. Traditionally, this kind of annotation is split into lemmatization and tagging. Lemmatization means annotating words with their lemmata (dictionary forms). The term tagging generally means annotating words with grammatical tags, such as “noun” or “genitive case”. While tagsets of moderately morphologically complex languages often only include part-of-speech tags, tagsets for more complex ones usually cover all morphological categories. The annotation is normally searchable with the help of an online or offline search interface.

Since corpora of written languages are normally too large for manual annotation to be feasible, their compilation includes developing an automatic tool for morphological annotation. Naturally, the more complex the morphology, the more difficult it is to develop such a tool. However, that when the complexity reaches certain point, we come face to face with completely new challenges beyond the increase in size of the formalized description of the morphology. It turns out that the very concepts of lemmatization and tagging have to be redefined to embrace the complexity of the morphological system. While these problems appeared to a certain degree and were partially addressed in corpora of such languages as Turkish, Tatar or Udmurt, they primarily manifest themselves in polysynthetic languages. In this paper, we outline these challenges, using the data from West Circassian (also known as Adyghe), a polysynthetic language belonging to the Circassian branch of the Northwest Caucasian (Abkhaz-Adyghe) family. To the best of our knowledge, no publicly available morphologically annotated corpora of polysynthetic languages have been developed to date, which makes our corpus unique.

Polysynthetic languages can be informally described as languages that may convey morphologically much of the information that in standard synthetic languages like Russian is conveyed by syntax. Consider the following West Circassian example:

(1)	*…тыгъужъыми*	*ыцэхэр*	*къыфыӀуигъэпсыгъэх*
	təʁ^{w}əẑə-m-jə	ə-ce-xe-r	qə-fə-ʔ^{w}-jə-ʁe-psə-ʁe-x
	wolf-OBL-ADD	3SG.PR-tooth-PL-ABS	DIR-BEN-LOC-3SG.ERG-CAUS-shine-PST-PL
	‘and the wolf	made its teeth	shine for him’

The verb in (1) simultaneously contains not only the causative and cross-reference affixes but also a locative preverb and the benefactive applicative, which here introduces the null cross-reference affix of the beneficiary.

Not surprisingly, such languages differ from Standard Average European in many respects. For example, much of their morphology is highly productive and shows syntactic properties (e.g., recursion, semantically based variation in order, etc.); cf. de Reuse (2009) who coined the term “productive non-inflectional concatenation” (PNC) for this kind of morphology. In addition, polysynthetic morphology sometimes is not at all selective and can attach to stems belonging to various lexical classes. For example, in West Circassian, tense markers appear not only on verbs, but also on adjectives, nouns and even postpositions. These properties of morphology pose multiple problems for tagging polysynthetic texts, as will be shown below.

2. West Circassian corpus

Developing West Circassian corpus is an ongoing project that started in 2015. As West Circassian is a written language with standard orthography based on the Cyrillic alphabet, the corpus will mostly consist of written texts, however, a certain number of manually annotated spoken texts collected during fieldwork will also be added. The corpus is being built in line with the general principles of medium-scale corpus design developed within the framework of Russian Academy of Sciences Corpus Program which was in effect in 2011–2014[2]. The workflow adapted in this program includes collecting written texts in standard orthography, developing an automated morphological analysis tool, annotating the texts with it and placing the corpus in the search engine with publicly available search interface. Morphological tagging in this framework is usually rule-based, being carried out with the help of a formalized description of the inflection and productive derivation of a language together with a grammatical dictionary containing the description of its lexis. The search engine which was used for most of these corpora and which we are going to use in the West Circassian corpus, was originally developed for the Eastern Armenian National Corpus and by default allows search by wordform, lemma or stem, translation, a combination of grammatical tags, as well as complex search involving a combination of the aforementioned properties (see Arkhangelskiy et al. 2012). However, the search capabilities which were sufficient for non-polysynthetic languages, proved to be insufficient for the West Circassian data, and have to be enhanced. We replaced the "bag of tags" principle, according to which the morphological tagger assigns each token grammatical tags without specifying relative order of tags within the set, with a mechanism that allows specifying relative position of morphemes in a search query. This enhancement is discussed in detail in section 3.4.

Currently, we are testing the solutions proposed below with a pilot version of grammatical dictionary. A publicly available preliminary version of the corpus is expected to be released in 2016.

3. Problems and solutions

For any West Circassian token, the following types of morphological and lexical annotation are included in our corpus:

(i) lemmatization,
(ii) part-of-speech attribution,
(iii) the presence of productive morphemes,
(iv) the order of productive morphemes.

Here productive morphemes comprise both inflection and PNC but not non-regular derivation which should be covered by the lexicon. The discussion will cover these topics in that order.

3.1. Tokenization: the subtle boundary between syntax and morphology

Tokenization, which is the first task in the text processing pipeline, already poses a problem specific for West Circassian. There admittedly exist difficulties for tokenization even in non-polysynthetic languages, e.g. annotation of multiword named

[2] Most middle-scale corpora developed within the framework of this program are available at http://web-corpora.net.

entities (such as “New York”), contractions, hyphenated words or text-based emoticons, as well as ways for dealing with these difficulties (cf. Grana et al. 2002, Bocharov et al. 2012). Most existing corpora assume for the sake of technical simplicity that a token cannot contain a whitespace, thus disregarding named entities (or annotating them at a separate level) and offering solutions for other problems within the limits of this constraint. Indeed, splitting the text into pieces delimited by whitespaces before further processing makes the tokenization step relatively fast and easy.

Although West Circassian normally does distinguish between syntactic relations and relations between morphemes, there are certain problems in demarcating morphology and syntax which lead to another kind of tokenization difficulties. Consider the following Adyghe example (2):

(2) *иджэнэ шхъонтIэ дахэхэр*
jə-ǯene-šχʷenṭe-daxe-xe-r
POSS-dress-blue-beautiful-PL-ABS
‘her beautiful blue dresses’

This example consists of three graphical tokens separated by whitespaces in standard orthography. Although it looks like an ordinary noun phrase, phonetic and morphologic criteria (specifically the absence of e/a alternation in the nominal stem, see Arkadiev, Testelets 2009) indicate that this *nominal complex* behaves as a single wordform (see Lander, to appear (a) for details). The reason why this is so problematic for corpus construction is the following. When attached to a nominal complex, prefixes and suffixes normally go to the left and the right edges of the whole complex, respectively. For instance, in the example (2) above, the plural marker modifies the whole complex apparently headed by the noun ‘dress’. However, if only graphical tokens are taken into account when performing morphological analysis, search queries like “dress in plural” or “a combination of a possessive and a plural marker in one word” will miss this example.

The nominal complex problem has no simple solution. If we do not recur to machine learning or other statistical methods which require a manually tagged golden standard corpus, all rule-based methods will not provide accurate results, as most words do not have alternations, and in most cases not having any prefixes or suffixes is perfectly normal for a West Circassian word. Even if we can identify such complexes accurately enough, annotating the whole complex as a single token has its drawbacks. For example, a simple query like “the token *daxexer*” would not find this graphical token inside a complex. At the current stage, we are not including nominal complexes recognition in our tokenization module. However, in the process of morphological analysis we are tagging tokens with no expected alternation, which can help in recognizing complexes in the future.

3.2. Lemmatization

The idea that a lemma can be attributed to every or almost every word is usually taken for granted in contemporary corpus linguistics. While this statement undoubtedly holds for all major languages for which corpora have been created, the situation is much less clear for languages with productive derivational morphology. Turkic or some of the Uralic languages provide "light" versions of such challenge which have been addressed in corpora and in bilingual dictionaries. In these languages, multiple derivational affixes, specifically, verbal markers such as causatives or iteratives, or nominalizations, may attach to the stem. Although these affixes are very productive and mostly semantically regular, in some cases they add to the meaning of the word in a non-compositional way. One of the existing solutions to this problem is annotating roots instead of lemmata. Another option is providing two alternative variants of tagging, which allows users to search for both derived and non-derived lemmata. Although this leads to some morphological ambiguity, its scale is limited in these languages: for example, according to Khakimov et al. (2014), this kind of ambiguity accounts for only 7.2% of all ambiguously tagged tokens in Tatar National Corpus.

In polysynthetic languages, however, this problem is much more pervasive and profound. In West Circassian, there are plenty of PNC affixes which are so productive that it is infeasible to include any new item derived with them into the lexicon. Nevertheless, the derived items often have non-compositional meanings, with the meanings themselves being often far less predictable than in Turkic languages.

Consider, for example, the applicative derivation, which adds an indirect object to the subcategorization frame of a word (see Smeets 1984; Lander, to appear (b); Lander & Letuchiy, to appear for details). West Circassian possesses a dozen of applicative affixes which may be added to roots and stems in a straightforward manner, as in (3) where the benefactive complex translates in English as 'for them':

(3) *афэтшIыщт*
[a-fe]-t-ṣ̂ə-š’t
3PL.IO-BEN-1PL.ERG-do-FUT
'We will do this for them.'

Since the applicative *fe-* is highly productive, its semantic contribution is purely compositional here, and it can easily be omitted (resulting in the form *tṣ̂əš’t* 'we will do this'), it makes no sense to lemmatize the form with this prefix.

The situation is different in (4), though.

(4) *фэмышIыгъэ*
fe-mə-ṣ̂ə-ʁe
BEN-NEG-do-PST
'not prosperous'

In this negative form of the word *fe-ṣ̂ə-ʁe* BEN-do-PST 'prosperous', only the negative prefix is used compositionally. The contribution of the benefactive applicative prefix and the past tense suffix is, on the other hand, idiomatic, despite the fact that both affixes are fully productive and are usually not likely to construct new lexemes. In languages like West Circassian, this kind of idiomatic lexicalization of morpheme combinations is quite widespread. It is evident that this situation requires consistent treatment that would go beyond the ambiguous analysis solution discussed above. Apart from search-related concerns, treating such combinations as having multiple ambiguous analyses with different stems or lemmata leads to difficulties during morphological tagging, as in West Circassian combinations of the root and derivational affixes can be split by inflectional morphemes. This would necessarily require adding disjointed stems to the dictionary and dealing with non-concatenative morphology, which makes morphological tagging a much more difficult task, although not completely impossible.

To address this problem, we use two different levels of annotation which are filled one after the other. During the main stage of tagging, the tokens are split into morphemes and glossed. Thus, every successfully analyzed token is assigned a lemma coinciding with its root, the description of which is stored at the first level. Then, the annotated token is passed to the second-level annotation module. This module loads a YAML file with rules that look like "if a token has root X together with affixes X, Y and Z, it should be assigned secondary lemma L". After applying the rules to the first level of annotation, all possible lemmata are written to the second level. For the word in the example (4), the first level will contain only information about the primary lemma *ṣ̂ən* 'do'. At the second level, it will be also associated with the lemma *feṣ̂əʁe* 'prosperous'. The search interface, correspondingly, was adapted to perform queries on both primary and secondary lemma at the same time.

3.3. Parts-of-speech (POS) tagging

The same kind of problems we face in lemmatization leads to challenges for POS-tagging as well. As with lemmatization, these challenges are present in Turkic and Uralic languages, to a much lesser extent. Specifically, these languages often have productive nominalization suffixes which can be used to derive a noun from virtually any verbal stem. Within the ambiguous analyses framework described above, the problem can be solved by assigning different POS tags to different analyses: the analysis that has the bare stem as its lemma will be assigned the tag "Verb", and the one where lemma includes the nominalization affix, the tag "Noun". Another way of addressing this issue, offered, for example, by Sak et al. (2008) for Turkish, is treating POS tags just like ordinary morpheme tags. In this approach, the stem and every POS-changing morpheme is annotated with the corresponding POS tag and, consequently, the analysis of one token can have more than one POS tag.

The situation is much more difficult in polysynthetic languages. Because of low selectivity of many affixes, the word class distinction itself is a serious problem for such languages[3]. In West Circassian, for example, tense affixes may attach to clearly nominal stems. The question is, then, whether this tense marker derives a new verb (see Lander and Testelets 2006 for some evidence) or it is simply not associated with any specific POS. Since both decisions are not theoretically fully justified in this case, we prefer to abstain from attempting to determine the POS tag of the word as a whole and rather only specify the POS of its primary lemma.

[3] For different views on the issue see, for example, Baker 2004 and several papers in Rijkhoff and van Lier (eds.) 2013. For Circassian see Lander and Testelets 2006.

Note that many wordforms with derivational affixes still are likely to be analyzed as belonging to one of the parts of speech, due to the presence of affixes that may be considered as clearly defining the class of the derived item. Examples of such affixes include the causative prefix and the agentive nominalization illustrated in (5):

(5) *уагъэшIущт*
w-a-ʁe-ŝʷə-š’t
2PL.ABS-3PL.ERG-CAUS-good-FUT
‘they will humour you (lit. make you good)’

къекIокIакIо
q-je-ḳʷe-č’-aḳʷe
DIR-DAT-go-go.out-AG
‘vagrant’

Nevertheless, even in the presence of such morphemes it is not always possible to unambiguously assign one of the POS tags to the token. For instance, when both the causative prefix and the nominalization suffix are present, it is not clear what applies the first and what applies the second. For example, in (6a) the causative clearly applies to the nominalization, but in (6b) the nominalization applies to the causative, as shown by brackets:

(6) a. *Зыжъугъэбэнак1у!*
zə-ẑʷ-ʁe-[ben-akʷ]
RFL.ABS-2PL.ERG-CAUS-[fight-AG]
‘Make yourselves fighters!’

b. *А гъэрэхьэтакIор сэры!*
a [ʁe-šx]-aḳʷe-r se-rə
that [CAUS-console]-AG-ABS I-PRED
‘That consoler is me!’

In order to enable searching for tokens for which it is possible to define a single POS tag, we suggest tagging affixes which clearly indicate the part of speech with additional labels such as NOMINAL or VERBAL. Such tagging will allow searching for e. g. all tokens which can be safely analyzed as nominal, by automatically transforming the query into “find all tokens which have a stem or a derivational affix marked as nominal and no derivational affixes marked as verbal”. At the same time, the decision will make it possible to look for any roots with any derivational suffixes, without specifying the final, resulting POS attribution.

3.4. Glossing and search capabilities

In nearly all large automatically tagged corpora each token is annotated with what is called ‘a bag of tags’, without specifying number of occurrences of each tag or their relative order. This approach is fully justified for Standard Average European languages, however it is hardly appropriate for West Circassian. One of the obstacles is recursion, whereby one affix or group of affixes may be used more than once during derivation (cf. Lander and Letuchiy 2010), as in example (7) below, which contains two benefactive applicative prefixes. The first of them introduces the argument corresponding to the phrase 'for them' in the English translation, and the second introduces the recipient argument translated as 'to him/her':

(7) *сафыфэтхэ*
s-a-fə-Ø-f-e-txe
1SG.ABS-3PL.IO-BEN-3SG.IO-BEN-DYN-write
'I write to him/her for them'

Another obstacle is variable morpheme order. While in some cases order may be irrelevant, in some others it affects the meaning because of the morpheme scope hierarchies. Finally, it is often important whether two morphemes border each other: e. g. when searching for a combination of an indirect object personal affix with an applicative, like in (7).

In order to address these problems, full glossing rather than a mere set of tags is stored for each token in the database of the corpus engine. We use abstract glosses, which are commonly used by typologists (see Lehmann 1982; Haspelmath 2002: 34–36) and present in the examples above. The query interface, still allowing for the 'bag-of-tag'-style queries, has been enhanced with a "glossing" search field which works with a glossing query language. When designing such a language, we considered the tradeoff between expressive power of the query language and the speed, as overly complex queries are usually hard to implement efficiently.

The query language we propose allows for any number of elementary queries joined by Boolean operators. Each elementary query can include grammatical tags and wildcard characters ? and *, the former standing for exactly one morpheme and the latter standing for any number of morphemes. Left and right word boundaries are marked by #. Morpheme adjacency is indicated by hyphens and their order is taken into account. For example, the query "#DIR-*-PST-?-ADV" will find all words starting with the directive prefix, following any number of morphemes, then a past tense marker, then another morpheme, and then and adverbial case marker. The language also allows using umbrella tags that unite several grammatical tags, e. g. the tag "APP" matches any applicative derivation affix such as BEN in (3), (4) and (7). Such elementary queries are currently transformed to SQL-queries containing regular expressions.

4. Conclusion

Corpus linguists dealing with polysynthetic language data face new kinds of challenges which are characteristic and often unique for these languages. It turns out that for such languages, many traditional techniques and concepts are not directly applicable to the data, and novel ways of text processing and corpus design should be developed.

We identified some of the problems which arise in the course of development of West Circassian language corpus, and offered possible solutions for them. The most important challenges, from our point of view, include somewhat vague boundary between morphology and syntax (hence tokenization problems), not well defined concepts of a lemma and a part of speech, and annotating the texts in such a way to enable search queries that could take into account phenomena like recursion and relative order of morphemes.

It should be noted that in this paper we only focused on a limited number of issues raised during the elaboration of the corpora. Some others include:

(i) morphophonological rules, which are by no means numerous but still should be accounted because of their high relevance for the analysis of the West Circassian word,

(ii) classes of morphemes: as we noted in Section 3.4, there is a need to group morphemes into classes, but the criteria of such grouping remain obscure,

(iii) the "translation" of our system into the conceptual system which is traditionally used in the descriptions of Circassian languages and in textbooks and hence should be considered for practical reasons.

References

1. *Arkadiev, P., Testelets, Ya.* (2009), On three alternations in the Adyghe language [O trex čeredovanijax v adygejskom jazyke], in Ya. G. Testelets et al. (eds), Aspekty polisintetizma: očerki po grammatike adygejskogo jazyka, 121–145. Moscow: RGGU.
2. *Arkhangelskiy, T., Belyaev, O., Vydrin, A.* (2012), The creation of large-scaled annotated corpora of minority languages using UniParser and the EANC platform, in Proceedings of COLING 2012: Posters, 83–91. Mumbai: The COLING 2012 Organizing Committee.
3. *Baker, M. C.* (2004), Lexical Categories: Verbs, Nouns, and Adjectives. Oxford: Oxford University Press.
4. *Bocharov, V., Granovsky D., Surikov, A.* (2012), Probabilistic tokenization model in the Open Corpus project [Verojatnostnaja model' tokenizacii v proékte Otkrytyj korpus], in Novye informacionnye texnologii v avtomatizirovannyx sistemax: Materialy 15, 15, 176–183.
5. *Grana, J., Barcala, F. M., Vilares, J.* (2002), Formal methods of tokenization for part-of-speech tagging, in Computational linguistics and intelligent text processing, 240–249. Springer Berlin—Heidelberg.
6. *Haspelmath, M.* (2002), Understanding Morphology. London: Arnold.
7. *Khakimov, B., Gil'mullin, R., Gataullin, R.* (2014), Morphological disambiguation in the Tatar corpus [Razrešenie grammatičeskoj mnogoznačnosti v korpuse tatarskogo jazyka], in Učenye zapiski Kazanskogo gosuniversiteta 156(5): 236–244.
8. *Lander, Yu.* (to appear, a), Nominal complex in West Circassian: between morphology and syntax, in Studies in Language.
9. *Lander, Yu.* (to appear, b), Adyghe, in P. O. Müller et al. (eds), Word Formation, An International Handbook of the Languages of Europe. Berlin: Mouton de Gruyter.

10. *Lander, Yu., Letuchiy, A.* (2010), Kinds of recursion in Adyghe morphology, in: H. van der Hulst (ed.), Recursion and Human Language, 263–284. Berlin: Mouton de Gruyter.
11. *Lander, Yu., Letuchiy, A.* (to appear), Decreasing valency-changing operations in a valency-increasing language? In Í. Navarro and A. Alvarez (eds), On verb valency change: theoretical and typological perspectives (working title).
12. *Lander, Yu., Testelets, Ya.* (2006), Nouniness and specificity: Circassian and Wakashan. Paper presented at the conference on Universality and Particularity in Parts-of-Speech Systems, University of Amsterdam.
13. *Lehmann, Chr.* (1982), Directions for interlinear morphemic translations, in Folia Linguistica 16: 193–224.
14. *de Reuse, W. J.* (2009), Polysynthesis as a typological feature. An attempt at a characterization from Eskimo and Athabascan perspectives, in M.-A. Mahieu and N. Tersis (eds), Variations on Polysynthesis: the Eskaleut Languages, 19–34. Amsterdam: John Benjamins.
15. *Rijkhoff, J., van Lier, E.* (eds). (2013), Flexible Word Classes, in Typological Studies of Underspecified Parts of Speech. Oxford: Oxford University Press.
16. *Sak, H., Güngör, T., Saraçlar, M.* (2008), Turkish language resources: Morphological parser, morphological disambiguator and web corpus, in Advances in natural language processing, 417–427. Springer Berlin—Heidelberg.
17. *Smeets, R.* (1984), Studies in West Circassian Phonology and Morphology. Leiden: The Hakuchi Press.