NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

Timofey Arkhangelskiy, Yury Lander

SOME CHALLENGES OF THE WEST CIRCASSIAN POLYSYNTHETIC CORPUS

BASIC RESEARCH PROGRAM

WORKING PAPERS

SERIES: LINGUISTICS
WP BRP 37/LNG/2015

This Working Paper is an output of a research project implemented at the National Research University Higher School of Economics (HSE). Any opinions or claims contained in this Working Paper do not necessarily reflect the views of HSE.

Timofey Arkhangelskiy[1]*, Yury Lander*[2]

SOME CHALLENGES OF THE WEST CIRCASSIAN POLYSYNTHETIC CORPUS[3]

Although there exist comprehensive morphologically annotated corpora for many morphologically rich languages, there have been no such corpora for any polysynthetic language so far. Polysynthetic languages raise a variety of theoretical and practical challenges for corpus linguistics. Some of these challenges have been partly addressed when developing corpora for e. g. Turkic or Uralic languages, while others are unique for this kind of languages. Our paper identifies the most prominent challenges that we are facing in the course of development of West Circassian (Adyghe) corpus, and offer possible 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 problems.

Keywords: language corpora, polysynthesis, West Circassian

JEL classification code: Z

[1] National Research University Higher School of Economics, School of Linguistics; email: tarkhangelskiy@hse.ru

[2] National Research University Higher School of Economics, School of Linguistics; email: yulander@hse.ru

[3] This paper is based on our research supported by Russian Foundation for Basic Research (grant No. 15-06-07434a) and made within the framework of the Academic Fund Program at the National Research University Higher School of Economics (HSE) in 2015 (grant No. 15-09-0263).

1. Introduction

While corpus linguistics is now in its period of growth, to the best of our knowledge, no big tagged corpora of polysynthetic languages seem to exist up to date. In this paper, we outline some issues which arise during the development of such a corpus, using the data from West Circassian (also known as Adyghe), a language belonging to the Circassian branch of the Northwest Caucasian (Abkhaz-Adyghe) family. Polysynthetic languages can be informally described as languages that may convey morphologically much information that in standard synthetic languages like English or Russian is conveyed by syntax. Consider the following West Circassian example:[4]

(1) ...*тыгъужъыми ыцэхэр къыфыIуигъэпсыгъэх*

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 noninflectional 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, because for such languages it is important to annotate not only words but also (at least some) morphemes.

[4] West Circassian examples are given both in the Cyrillic orthography (as they will look in our corpus) and in the simplified phonological transcription with morpheme-by-morpheme glosses.

The structure of this paper is as follows. First, in Section 2, we outline the general framework within which we are developing our West Circassian corpus. The questions and challenges that arise in the course of development are discussed in detail in Section 3. The conclusions are presented in Section 4.

2. Overview of the proposed corpus

Generally, any collection of texts in a given language whose primary purpose is to serve as a source of language examples for linguistic research can be called a corpus of that language. Currently, there are thousands of digital corpora of various languages. Although they all fall under this broad definition, their properties, which are influenced by both linguistic traits of the idiom and target research, may vary significantly. Since different kinds of corpora pose different kinds of challenges and call for different kinds of solutions, it is important to describe goals and features of our proposed corpus.

Linguistic corpora differ in what kinds of texts are included in them, what levels of annotation accompany the texts, and what search capabilities are available to the researcher. In this paper, we will mostly discuss the two latter questions and the challenges they raise. Although the text composition of the corpus also poses a lot of challenges, specifically those connected to representativeness and balancedness, these challenges are quite different in nature and do not depend on the grammatical profile of the language. West Circassian is a written language which has standard orthography based on the Cyrillic alphabet. The corpus will include literary texts and spoken texts, both kinds being written in standard orthography. That said, we turn to the question of annotation, leaving texts-related problems beyond the scope of this paper.

The kinds of annotation available in contemporary corpora include morphological tagging, syntactic annotation, various kinds of semantic annotation, anaphora resolution, recognition of named entities, and many others. As we are developing a general purpose corpus, we have to provide at least the annotation that is necessary for virtually any research. For a polysynthetic language like West Circassian, it is crucial to have morphological annotation in texts. Unlike in languages with simpler morphology, the commonly used solution which only includes lemmatization and part-of-speech tagging is inapplicable for a polysynthetic language, since that kind of annotation would ignore lots of important morphological information. However, there is another, less evident reason why this simplistic approach is not sufficient for such languages: the very concepts of part of speech and lexeme are much more vague in polysynthetic languages than in

Standard Average European languages that most corpus linguistic efforts have been dedicated to so far. These problems will become our main topic of discussion in sections 3.2 and 3.3. Although other levels of annotation would undoubtedly be beneficial for the corpus, their introduction would be extremely time-consuming. Given that they are not as necessary for a general-purpose corpus as the morphological annotation, we are leaving them for future work.

Our West Circassian corpus will be 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.[5] 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 analysis 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 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. Particularly, 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, has to be replaced with a mechanism that would allow specifying relative position of morphemes in a search query. This enhancement is discussed in detail in section 3.4.

3. Challenges and solutions

In this section, we will examine the challenges that the polysynthetic language data raises in corpus construction, and offer possible solutions for them. The order of subsections will roughly correspond to the order in which the problems described appear in the pipeline. Thus, the discussion will start with tokenization issues, followed by questions raised in the process of compiling the grammatical dictionary and challenges of formal description of morphology and search queries. For

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

any West Circassian token, we assume that ideally the following types of morphological and lexical information should be included in our corpus:

(i) lexical attribution linking the token to the lexicon (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.[6] The discussion will cover these topics in that order.

3.1. Tokenization: the subtle boundary between syntax and morphology

The first task in the text processing pipeline, splitting the text into word units, or tokenization (see e. g. Grefenstette 1999), already poses a problem specific for West Circassian. There admittedly exist difficulties for tokenization even in non-polysynthetic languages, e.g. annotation of named 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 шхъуэнтIэ дахэхэр*

jə-ǯene-šχ^{w}enṭe-daxe-xe-r

[6] Distinguishing between PNC and non-regular derivation is a separate issue, which remains beyond the scope of this paper.

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 indicate that this is in fact a single wordform (see Lander 2015 for details). One of the main arguments for such an analysis comes from morphophonology. In West Circassian words, the root together with certain affixes undergoes regular alternation process, whereby under certain conditions the underlying segment /CeCe/ at the right edge of a combination is replaced by surface-level /CaCe/[7] (Arkadiev, Testelets 2009). According to this rule, the word for ‘her dress’ in isolation should look like *jəʒane*. The fact that the alternation did not occur in this word indicates that its graphical right edge does not correspond to the word boundary. It follows that the whole complex (2) is in fact a single word with a single manifestation of the alternation in the last part, where the stem *dexe* ‘beautiful’ appears as *daxe*.

The reason why such nominal complexes and similar structures pose a problem for corpus construction is the way morphology treats them. The West Circassian nominal and verbal morphology includes both prefixes and suffixes. When attached to a nominal complex, they 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 presumably 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, there seem to be two approaches for detecting such complexes. One of them is based on the aforementioned alternation pattern: if the token does not have the alternation it was supposed to have, include the word to its right in the complex. The other takes morphology into account: continue to the right if the word does not have suffixes it was supposed to have (and vice versa with the prefixes and leftward search). Unfortunately, both methods will not provide accurate results, as most words do not have alternations, and in most cases not having any prefixes or suffixes is

[7] Plural marker -*xe*-, which is found in the last word of the example, is not included in the set of alternating affixes.

perfectly normal for a West Circassian word. Even if we can identify the complexes accurately enough, we face the problem of annotation. 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. It would, therefore, be necessary to have several layers of annotation, so that the information about both the graphical words and the linguistic tokens is accessible to the search engine. At the current stage, we are not including nominal complexes recognition in our tokenization module. However, in the process of morphological analysis we will tag tokens with no expected alternation, which can help in recognizing complexes in the future.

3.2. Lemmatization

Lemmatization, i.e. providing each token with its dictionary form, is one of the basic steps in text processing pipelines in corpus linguistics and in natural language processing applications. The idea that a lemma can be unequivocally attributed to every or almost every word is usually taken for granted in contemporary corpus linguistics. While this statement undoubtedly holds for Standard Average European languages and other 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 regular from the point of view of semantics, in some cases they add to the meaning of the word in a non-compositional way. Consider the following Udmurt (Uralic > Permic) example:

(3) puk-i̯ni̯ vs. puk-t-i̯ni̯
sit-INF sit-CAUS-INF

The first verb translates as ‘to sit’. While one of the possible translations for the second verb is ‘to make someone sit’, which is compositionally deducible from the meanings of the root and the causative suffix, another, and quite frequent, translation is ‘to build (a house)’. Since the meaning of this combination is non-compositional, the word *pukti̯ni̯* should appear as a separate dictionary entry in a bilingual dictionary. Correspondingly, it makes sense to lemmatize this word in a corpus as *pukti̯ni̯*. However, annotating it as a form of *puki̯ni̯* also makes sense, since otherwise the query "*pukti̯ni̯* + CAUS" will not find anything in the corpus.

There are two solutions to this problem that have been offered in corpora and dictionaries. The first one is annotating the tokens with their root instead of their lemma, thus eliminating the concept of the lemma altogether. This approach does not seem to be widespread, presumably because most morphological analyzers for such languages existing today use the data from bilingual dictionaries that had already existed in digital form. Since the dictionaries always list non-compositional derived forms (and, unfortunately for corpus linguistics, often list compositinal ones as well), usually another approach is chosen for such corpora. Under this approach, such items obtain both analyses, i.e. get both the derived lemma and the non-derived one (such approach was used e. g. in the Udmurt corpus and in the Tatar corpus). Such a solution allows users to search for both 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; Lander and 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 (4):

(4) *афэтшIыщт*

[a-fe]-t-ṣ̂ə-š’t

3PL.IO-BEN-1PL.ERG-do-FUT

‘We will do this for them.’

In (4) we find the applicative complex *a-fe-* ‘for them’, which can be easily omitted (resulting in the form *tṣ̂əš’t* ‘we will do this’). Since the applicative *fe-* is highly productive and its semantic contribution is purely compositional here, it makes no sense to lemmatize the form with this prefix.

The situation is different in (5), though.

(5) *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. The level of idiosyncrasy of this combination is much higher than in most Turkic or Uralic examples, including (3). In fact, in languages like West Circassian, this kind of idiomatic lexicalisation of morpheme combinations is quite widespread. It seems 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. While in the case of Turkic and Uralic languages both alternative stems can be represented as contiguous segments of the word (usually starting the word, as prefixes are virtually nonexistent in these languages), 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.

In this case, we propose to use two different levels of annotation which are filled one after the other. During the main stage of morphological tagging, the tokens should be split into morphemes and glossed. Thus, every successfully analyzed token will be assigned a root, the description of which will be stored at the first level. Then, the annotated token should be passed to the second-level annotation module which will do the lemmatization. The lemmatizer should use a separate dictionary containing rules that would look like “if a token has root X together with affixes X, Y and Z, it should be assigned lemma L”. After applying the rules to the first level of annotation, all possible lemmata will be written to the second level. For the word in the example (5), the first level will contain only information about the root *ṣ̂ə* ‘do’. At the second level, it will be (ambiguously) associated with two dictionary entries, *ṣ̂ə* ‘do’ and *feṣ̂əʁe* ‘prosperous’. The search interface, correspondingly, has to provide the possibility of searching by the stem as well as searching by the lemma.

Singling out lemmatization in a separate procedure that takes place after the main part of the tagging is performed allows us to avoid adding complex stems in the grammatical dictionary and also moves the ambiguity to only one of the analysis levels.

3.3. Parts-of-speech (POS) tagging.

Another step commonly used in text processing is POS-tagging. It is a trivial observation that for languages with rich morphology simple POS-tagging alone is not enough and detailed morphological tagging is required for the corpus to be a useful resource. Nevertheless, together with lemmatization, POS-tagging is traditionally an obligatory step in the pipeline of text processing. The same kind of problems we face in lemmatization lead 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.[8] In West

[8] 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.

Circassian, for example, tense affixes may attach to clearly nominal stems, as with the borrowed stem *oficerə* ‘officer’ in the following example :

(6) *Офицэрыщтыгъ, шыудзэм хэтыгъ.*

oficerə-š’tə-ʁ, šəw-ʒe-m xe-tə-ʁ

officer-AUX-PST rider-army-OBL LOC-stand-PST

‘He was an officer, served in cavalry.’

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 the root. Therefore the POS information will be available at only one of the two levels described in the previous subsection. However, many wordforms with derivational affixes, even morphologically complex ones, 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 clearly defining the class of the derived item. Examples of such affixes include the causative prefix and the agentive nominalization illustrated in (7) and (8) respectively:

(7) *уагъэшIущт*

w-a-ʁe-ṣ̂ʷə-š’t

2PL.ABS-3PL.ERG-CAUS-good-FUT

‘they will humour you (lit., make you good)’

(8) *къекIокIакIо*

q-je-ḳʷe-č̣’-aḳʷe

DIR-DAT-go-go.out-AG

‘vagrant’

Note that 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 (9a). the causative clearly applies to the nominalization, but in (9b) the nominalization applies to the causative, as shown by brackets:

(9) a. *ЗыжъугъэбэнакIу!*

zə-ẑ^{w}-ʁe-[ben-ak^{w}]

RFL.ABS-2PL.ERG-CAUS-[fight-AG]

‘Make yourselves fighters!’

b. *А гъэрэхьэтакIор сэры!*

a [ʁe-šx]-aḳ^{w}e-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 anayzed 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

Due to the active use of PNC affixes in West Circassian, it is clear that the users of our corpus should have access to the morphological information and internal structure of the tokens when making search queries. A common way to provide this capability is annotating tokens with grammatical tags. Because of the large number of grammatical categories in Western Circassian and because it is typologists that constitute a large portion of the target audience of the corpus, it would be impractical to use BNC-style one-letter or several-letter tags first designed for the Brown corpus (Francis, Kucera 1979). Instead, we are going to use abstract glosses, which are commonly used by typologists (see Lehmann 1982; Haspelmath 2002: 34–36) and are shown in the examples above as well as in (10):

(10) q-je-če-ʁa-ç̣’-ew

DIR-DAT-run-PST-IMMEDIATE-ADV

‘just after running to it’

However, there is a larger problem with the traditional approach to tagging. In nearly all sufficiently large automatically tagged corpora each token is annotated with what is called ‘a bag of tags’. Under this approach, the token in (10) would be annotated with a set of tags “adv,dir,dat,immediate,pst” without specifying number of occurrences of each tag or their relative order. This approach is fully justified for Standard Average European languages, as nearly every morpheme has a fixed place in the word, which makes it possible to recover the information about the relative order of morphemes from the bag of tags, and there is no morphological recursion, i.e. multiple appearance of the same morpheme in the wordform. The situation in polysynthetic languages is strikingly different. In West Circassian, there are numerous examples which render the ‘bag of tags’ approach insufficient. 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 (11) below, which contains two benefactive applicative prefixes:

(11) *сафыфэтхэ*

s-a-fə-Ø-f-e-txe

1SG.ABS-3PL.IO-BEN-3SG.IO-BEN-DYN-write

‘I write to him for them’

The traditional approach would not allow differentiating between words with one occurrence of a specific tag and those with multiple occurrences, which is required in research on recursion. If we turn once again to the existing corpora for morphologically rich languages, we can find a halfway solution that enables such queries without changing the annotation model. For example, in the corpus of Kalmyk (Mongolic), where the causative suffix can attach to a stem twice, tokens with two causatives obtain a special tag, CAUS2, alongside the usual tag for causative. However, this solution is hardly applicable to West Circassian data (even though a similar theoretical solution was assumed for double causatives in Circassian languages by Kumakhov 1965), not only because recursion here is more widespread and diverse, but also because there are other obstacles which are generally absent in e. g. Kalmyk or Turkic languages. Consider (12) and (13):

(12) a. *слъэгъуыгъапэ*

s-λeʁʷə-ʁa-pe

1SG.ERG-see-PST-ASSERT

b. *слъэгъуыпагъ*

s-λeʁʷə-pa-ʁ

1SG.ERG-see-ASSERT-PST

‘I really saw that.’ (Korotkova and Lander 2010: 305)

(13) *дэплъыен*

de-pλə-je-n

LOC-look-UPWARDS-MOD

‘to look up’ (Arkadiev and Letuchiy 2011)

As can be seen from (12), there are combinations of affixes that can appear in the word in different orders. Sometimes the order is variable and does not influence the meaning, while in other cases one of the affixes falls into the scope of the other, which is reflected in the semantics of the two variants. Whichever the case, it is obvious that many kinds of research require access to the information on the relative position of affixes. In (13), we see another example where the appearance of the affix *-je* ‘upwards’ in one part of the word requires the appearance of the locative applicative affix *de-* in the other part of the word (cf. Arkadiev and Letuchiy 2011). The choice of the applicative is not always predictable and hence represents a parameter that may be relevant for the search.

The arguments above suggest that tokens have to be accompanied by more elaborate annotation, and that search interface has to provide search tools for accessing this annotation. We propose that instead of a set of tags, full glossing is stored for each token in the database of the corpus engine. With such kind of annotation, all morphological information is preserved and remains potentially searchable. However, in order to make use of this information, search interface should be enhanced. Apart from traditional search queries, which in case of the platform we are using for our corpus allows searching for grammatical tags and their combinations, including use of logical functions, the interface should allow queries which can specify relative position of the affixes, indicate whether certain affixes are adjacent in the token, or take recursion into account. When designing an enhanced interface, it should be brought in mind that there is a tradeoff between expressive power of the query language and the speed, as overly complex queries are usually hard to implement efficiently.

Our proposed solution, which we believe is expressive enough and at the same time still allows efficient implementation, is using both a field for ‘bag of tags’ queries and a field for queries related to internal word structure. In the latter, restrictions on the morphological structure can be expressed by means of grammatical tags and a couple of wildcard characters. The wildcard characters, e.g. the question mark and the asterisk, stand for ‘any morpheme’ and ‘any number of morphemes’ (including zero). For example, the word in (10) may be accessed by queries like those in (14) (among others).

(14) *-imm-*

-run--adv

-dir-

-run-pst-

An example of using a question mark for one morpheme is a query in (15), which would also find the word in (10):

(15) dir-?-?-pst-?-adv

This simple convention, on the one hand, makes it possible to take into account various kinds of morphological information, including the points listed in the beginning of the subsection, while, on the other hand, it can be implemented to process such queries sufficiently quickly (implementation details go beyond the scope of this paper and thus will not be covered here).

Yet such a scheme does not provide all information that may be required for a query and has to be further refined. First, it seems important to have a possibility to refer not only to specific morphemes but also to morpheme classes. For example, if the past suffix is considered to be a tense marker, it may be referred not only as pst but also as tense. The word (10) can be found, for example, using queries like:

(16) *-run-tense-*

-root--pst-*

Second, we must have a possibility to refer not to a specific order of morphemes but just to the presence of several morphemes irrespectively of their order. This may be achieved by combining the 'bag of tags' field with the one that refers to the word structure. Both these fields should allow using simple Boolean operators like AND and OR. For (10), for example, this may look as:

(17) *-root-* and dir

-root- and (pst or fut)

Third, not only the presence of morphemes should be thought of, but also the absence thereof. This is important, since given the narrow selective restrictions of affixes, it may still be that the presence of an affix (such as tense) is unmarked for some word classes (like verbs) but is marked for other word classes (like nouns). This can be achieved with a combination of two types of queries and using a Boolean operator NEG in one of them, as in (18):

(18) *-root-* and (pst or fut) and neg(case)

4. Conclusion

Corpus linguists dealing with polysynthetic language data face new kinds of challenges which are characteristic for these languages. Although some of them were studied and addressed in some form when dealing with the data of e.g. Turkic languages, many of which have digital corpora, we showed that some of them are unique for polysynthetic 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 (which poses tokenization problems), not well defined concepts of a lemma and a part of speech, and 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.

We are currently implementing the solutions discussed in the paper. Future prospects of our work include evaluation of the proposed solutions and addressing the issues listed above.

Abbrebiatons

ABS – absolutive, ADD – additive, ADV – adverbial, AG – agentive nominalization, ASSERT – asserive, AUX – auxiliary stem, BEN – benefactive, CAUS – causative, DAT – dative, DIR – directive, DYN – dynamic, ERG – ergative, FUT – future, INF – infinitive, IO – indirect object, LOC – locative preverb, MOD – modal, NEG – negation, OBL – oblique, PL – plural, POSS – possessive, PR – inalienable possessor, PRED – predicative form, PST – past, RFL – reflexive, SG – singular.

References

Arkadiev, P. and A. Letuchiy. 2011. Prefixes and suffixes in the Adyghe polysynthetic wordform: types of interaction. In: V. S. Tomelleri et al. (eds), *Languages and Cultures in the Caucasus*, 495—514 Muenchen: Otto Sagner.

Arkadiev, P. M. and Ya. G. Testelets. 2009. O trex čeredovanijax v adygejskom jazyke. In: Ya. G. Testelets et al. (eds), *Aspekty polisintetizma: očerki po grammatike adygejskogo jazyka*, 121-145. Moscow: RGGU.

Arkhangelskiy T., O. Belyaev, and A. Vydrin. 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.

Baker, M. C. 2004. *Lexical Categories: Verbs, Nouns, and Adjectives*. Oxford: Oxford University Press.

Bocharov, V. V., D. V. Granovsky, and A. V. Surikov. 2012. Verojatnostnaja model' tokenizacii v proékte Otkrytyj korpus. *Novye informacionnye texnologii v avtomatizirovannyx sistemax: Materialy* 15, *15*, 176-183.

Francis, W. N., and H. Kucera. 1979. Brown Corpus manual: Manual of information to accompany a standard corpus of present-day edited American English for use with digital computers. Brown University, Providence, Rhode Island, USA.

Grana, J., F. M. Barcala, and J. Vilares. 2002. Formal methods of tokenization for part-of-speech tagging. In: *Computational linguistics and intelligent text processing*, 240-249. Springer Berlin Heidelberg.

Grefenstette, G. 1999. Tokenization. In: *Syntactic Wordclass Tagging*, 117-133. Dordrecht: Kluwer.

Haspelmath, M. 2002. *Understanding Morphology*. London: Arnold.

Khakimov, B. É., R. A. Gil'mullin, and R. R. Gataullin. 2014. Razrešenie grammatičeskoj mnogoznačnosti v korpuse tatarskogo jazyka. *Učenye zapiski Kazanskogo gosuniversiteta* 156(5): 236-244.

Korotkova, N. and Yu. Lander. 2010. Deriving affix order in polysynthesis: evidence from Adyghe. *Morphology* 20(2): 299—319.

Kumakhov, M. A. 1965. Distributivnuj analiz polisintetičeskogo kompleksa. *Voprosy jazykoznanija*, no. 5: 112-117.

Lander, Yu. To appear. Adyghe. In: P. O. Müller et al. (eds), *Word Formation, An International Handbook of the Languages of Europe*. Berlin: Mouton de Gruyter.

Lander, Yu. and A. Letuchiy. 2010. Kinds of recursion in Adyghe morphology. In: H. van der Hulst (ed.), *Recursion and Human Language*, 263–284. Berlin: Mouton de Gruyter.

Lander, Yu. and A. Letuchiy. 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).

Lander, Yu. and Ya. Testelets. 2006. Nouniness and specificity: Circassian and Wakashan. Paper presented at the conference on Universality and Particularity in Parts-of-Speech Systems, University of Amsterdam.

Lehmann, Chr. 1982. Directions for interlinear morphemic translations. *Folia Linguistica* 16: 193 – 224.

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.

Rijkhoff, J. and E. van Lier (eds). 2013. *Flexible Word Classes. Typological Studies of Underspecified Parts of Speech*. Oxford: Oxford University Press.

Sak, H., T. Güngör, and M. Saraçlar. 2008. Turkish language resources: Morphological parser, morphological disambiguator and web corpus. In: *Advances in natural language processing*, 417-427. Springer Berlin Heidelberg.

Smeets, R. 1984. *Studies in West Circassian Phonology and Morphology*. Leiden: The Hakuchi Press.

Timofey Arkhangelskiy

National Research University Higher School of Economics, School of Linguistics; email: tarkhangelskiy@hse.ru

Yury Lander

National Research University Higher School of Economics, School of Linguistics; email: yulander@hse.ru

Any opinions or claims contained in this Working Paper do not necessarily reflect the views of HSE.