Discovering dialectal differences based on oral corpora

Vasilisa Andriyanets, Michael Daniel, Brigitte Pakendorf

Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2018"

Moscow, May 30—June 2, 2018

DISCOVERING DIALECTAL DIFFERENCES BASED ON ORAL CORPORA[1]

Andriyanets V. (blindedbysunshine@gmail.com),
Daniel M. (misha.daniel@gmail.com)
International Linguistic Convergence Laboratory, NRU HSE, Moscow, Russia

Pakendor B. (brigitte.pakendorf@cnrs.fr)
Laboratoire "Dynamique du Langage", UMR5596, CNRS & Université de Lyon, Lyon, France

This paper discusses a method to detect statistically significant linguistic differences between corpora while factoring in possible variability within the very corpora to be compared. Specifically, we compare two small corpora of dialects of Even, Bystraja and Lamunkhin Even, in an attempt to identify morphemes that are more frequent in either of the corpora. To investigate whether this difference might be due to an over-representation of a speaker who happens to be an outlier in terms of using a particular morpheme, we use DP, a measurement of evenness of the distribution of a specific linguistic feature across subcorpora of the same corpus.

Keywords: linguistic corpora, Even, dialect variation, outliers

[1] This article was prepared within the framework of the Basic Research Program at the National Research University Higher School of Economics (HSE) and supported within the framework of a subsidy by the Russian Academic Excellence Project '5-100'. BP is grateful to the LABEX ASLAN (ANR-10-LABX-0081) of Université de Lyon for its financial support within the program "Investissements d'Avenir" (ANR-11-IDEX-0007) of the French government operated by the National Research Agency (ANR).

ИЗВЛЕЧЕНИЕ ДИАЛЕКТНЫХ РАЗЛИЧИЙ НА МАТЕРИАЛЕ УСТНЫХ КОРПУСОВ

Андриянец В. (blindedbysunshine@gmail.com),
Даниэль М. (misha.daniel@gmail.com)
Международная лаборатория языковой конвергенции, НИУ ВШЭ, Москва, Россия

Пакендорф Б. (brigitte.pakendorf@cnrs.fr)
Лаборатория «Динамика языка», UMR5596, Национальный центр научных исследований и Университет Лион, Лион, Франция

1. Introduction

Even is a North Tungusic language spoken in Siberia and the Russian Far East. The overall number of speakers is probably not more than two to three thousand, and these are settled over a vast territory in small individual speech communities. It is a dialectally diverse language, with 13 dialects (диалекты) and up to 24 subdialects (говоры) recognized [Burykin 2004: 85]. This paper compares data from two Even dialects, Lamunkhin Even and Bystraja Even. Lamunkhin is the westernmost still viable dialect of Even spoken in the village Sebjan-Küöl in Yakutia. It has been in close contact with the Turkic language Sakha (Yakut) for decades, and Sakha influence is registered at all levels of the language [Pakendorf 2009]. Bystraja, one of the easternmost Even dialects, is spoken in the Bystraja district in central Kamchatka. The extent to which it may have undergone contact influence from the neighbouring language Koryak is yet to be elucidated. This dialect is undergoing a shift to Russian, with no confident speakers younger than 40–45 years of age.

The texts in the corpora were collected in several field trips between 2007 and 2015 by Brigitte Pakendorf with contributions by Natalia Aralova. These are spoken texts, mostly narratives that were glossed and translated in Field Linguist's Toolbox[2], with the majority time-aligned in ELAN[3]. The Lamunkhin corpus comprises ~50,000 tokens and is recorded from 37 speakers, and the Bystraja corpus comprises ~34,000 tokens and is recorded from 26 speakers. Importantly, as will be described in Section 2, neither corpus is balanced in terms of contributions by individual speakers.

Even dialects are known to differ across linguistic domains [cf. Rišes & Cincius 1952], not only in lexicon, but also in phonology, morphology and syntax. The two dialects included in this study are no exception (e.g. [Matić & Pakendorf 2013], [Pakendorf & Krivoshapkina 2014], Pakendorf to appear). For instance, the simultaneous converb

[2] https://software.sil.org/toolbox/

[3] https://tla.mpi.nl/tools/tla-tools/elan/

–nikEn[4] is used more frequently in the Lamunkhin dialect, where in addition the converb of 'say' has taken on extended functions, such as that of complementizer. Furthermore, habitual aspect is expressed with the 'generic' suffix in the Bystraja dialect and with the 'habitual' suffix in the Lamunkhin dialect[5] (1). In the nominal domain, the dative case is extending to addressees of verbs of speech in the Lamunkhin dialect, a function that is fulfilled only by the allative case in the Bystraja dialect (2).

(1a) Bystraja (RME_fox_wolf_053)
nan *tačin* *tar* ***go:-niken*** *ereger* *ŋene-d-**ʤo:t**-te-n*
and dist.qual dist **say-sim.cvb** always go-prog-**gnr**-nfut-3sg
'Saying like this [the fox] was coming all the time…'

(1b) Lamunkhin (stado#10_SEN_poems_084)
Mitja *ihu-riʤi* *ọralči-mŋa* *bi-ʤi-n* ***go:-niken*** *ʤọmkak-**kara**-m*
Mitja grow-ant.cvb herd.reindeer-agnr be-fut-3sg **say-sim.cvb** think-**hab**[nfut]-1sg
'I think that Mitja, having grown up, will be a reindeer herder.'

(2a) Bystraja (NAT_rabotajushaja_010)
ose:l-če-l *eniŋe:-wu* *gia-**tki*** *atikan-**taki***
go:n-ʤid-de-n
become.tired-pf.ptc-pl grandmother-poss.1sg next-**all** old.woman-**all**
say-prog-nfut-3sg
'My tired grandmother said to the other old woman,…'

(2b) Lamunkhin (KKK_Emcheni_056)
asatka-čan *bọllaɣina* *tar* *ọmọlgọ* *kuŋa-**du*** *go:n-če*
girl-dim dp.Y dist boy child-**dat** say-pf.ptc
'And the girl said to the boy:…'

Such differences may result from divergence of dialects of the same language (i.e. independent evolution of mutually isolated linguistic systems) or convergence of dialects or languages (via contact-induced changes in socially and geographically adjacent linguistic systems). However, as yet no comprehensive account of the morphosyntactic differences between the two dialects exists. In this paper we undertake the first step towards filling this gap by assessing the issues involved in detecting significant quantitative differences between the dialects using small corpora of oral narratives. To the best of our knowledge, this is the first attempt at using text corpora to evaluate differences between dialects. We assume that any statistically significant quantitative difference deserves a qualitative interpretation, but such interpretation is outside the scope of this paper.

---

[4] Even retains vestiges of vowel harmony, so that vowels in suffixes can vary between [e] and [a], and consonants undergo various assimilation processes. When suffixes are shown in isolation, capital letters indicate phonemes that undergo changes.

[5] Abbreviations used in the glosses are: agnr: agent nominalizer; all: allative; ant: anterior; cvb: converb; dat: dative; dim: diminutive; dist: distal (demonstrative); dp: discourse particle; fut: future; gnr: generic; hab: habitual; nfut: non-future; pf: perfect; poss: possessive; prog: progressive; ptc: participle; qual: qualitative; sim: simultaneous; Y: Sakha (Yakut) borrowing

2. Data

The texts in the two corpora were glossed in Toolbox over a period of several years, and contained, in addition to typos, some inconsistencies and traces of the evolution of the researcher's conceptions that needed to be corrected in an initial step. Although most of the texts are narratives, in some of them speakers other than the narrator step in for a phrase or two, and a few are conversations. Each annotation unit (roughly, a sentence, to the extent that the notion of sentence can be applied to spoken text) is thus associated with a speaker (represented by a two- or three-letter code).

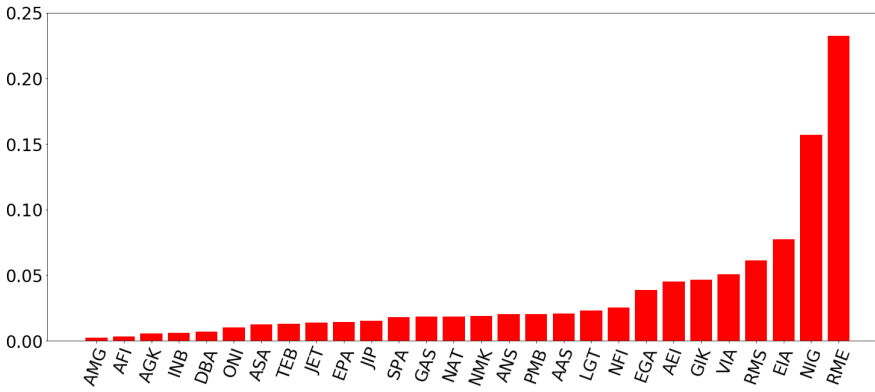

Fig. 1. Relative contributions of individual speakers to the Bystraja corpus

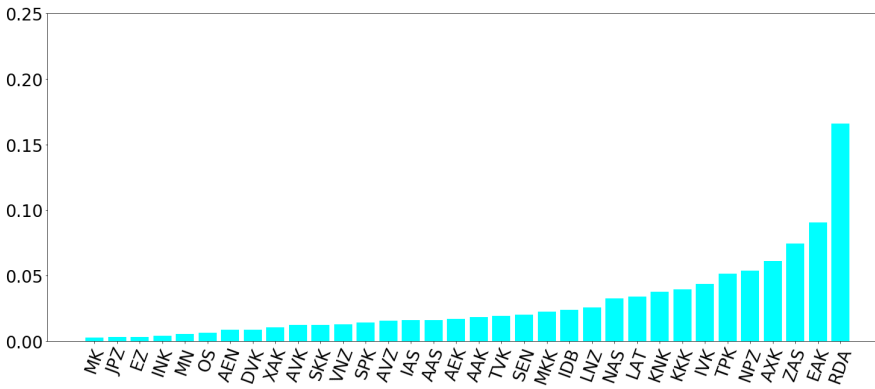

Fig. 2. Relative contributions of individual speakers to the Lamunkhin corpus

As discussed above, the Bystraja corpus contains 34,000 tokens and the Lamunkhin corpus contains 50,000 tokens. However, both corpora are very unbalanced in terms of contributions of individual speakers (Fig. 1, 2). This is particularly striking in the Bystraja corpus, where the two speakers with the largest contributions (RME

and NIG) account for 39% of the corpus (with just RME's contribution amounting to 23%)—practically equal to the contribution of the ten speakers with the next largest contributions (41%). Although the Lamunkhin corpus is somewhat more balanced, even here one third was contributed by only three speakers (RDA, EAK, and ZAS), with RDA alone accounting for 13%. Estimates of dialectal frequency differences based on such unbalanced corpora might well be distorted by the linguistic idiosyncracies of the speakers who contributed the most, rather than reflecting actual differences between the dialects.

3. Method

Notwithstanding the heterogenous nature of the corpora, our first approach to a quantitative analysis of the data was the conventional log-likelihood measure (as used in [Rayson & Garside, 2000]) to compare the relative frequencies of each suffix in the corpora. The log-likelihood comparison is based on the overall size of the compared corpora and the frequency of a given suffix in each corpus[6]. The comparison indicates noticeable differences between the corpora in the use of several morphemes; here, we focus on the spatial case markers (locative, allative and dative, which is also used as a directional marker) and converbs[7]. Table 1 shows the observed and the expected (under the assumption of an even distribution in the two corpora) frequencies of the markers in the two dialects.

The difference in the frequencies is overwhelming in the case of the simultaneity converb and the allative, but it is also deemed significant for the other suffixes (log-likelihood values above 10 are considered to be statistically significant).

A potential problem with this analysis is that, as explained in Section 2 above, the two corpora are unbalanced in terms of speaker representation. Given the small size of the corpora (~50,000 tokens for Lamunkhin and ~34,000 tokens for Bystraja) and the large variation in speaker contributions illustrated in Fig. 1 and Fig. 2, the differences between the corpora shown in Table 1 might reflect not actual differences between the dialects, but rather be the result of a greater contribution to either of the corpora by a speaker or speakers who are (in)frequent users of a particular grammatical category.

---

[6] The expected value of a morpheme for each corpus is calculated on the basis of the overall frequency of the morpheme in both corpora in relation to the relative size of each corpus. See http://ucrel.lancs.ac.uk/llwizard.html for a technical description of the log-likelihood calculation.

[7] Converbs are subordinate verb forms typically introducing adverbial clauses and specifying a temporal, logical or other relation of the event described by such subordinate clauses to the main clause. See [Haspelmath, König 1995] for a typological overview of the category. They are not obligatory, since they can be replaced by alternative subordinate constructions using case-marked participles or by chains of finite verb clauses.

Table 1. Log-likelihood comparison for spatial case markers and converb suffixes, ordered by decreasing log-likelihood (LL) value

Suffix	Observed, Lamunkhin	Observed, Bystraja	Expected, Lamunkhin	Expected, Bystraja	LL
Simultaneity converb -nikEn	1,083	59	688.72	453.28	739.84
Allative case -t(E)ki	78	378	275.00	181.00	360.17
Multiplicative converb ntEkEn	101	0	60.91	40.09	102.15
Anteriority converb -Ridʒi	721	237	577.75	380.24	95.30
Purposive converb -DE	307	405	429.40	282.60	85.46
Dative case -Du	699	299	601.88	396.12	40.93
DS[8] conditional converb -REk	488	197	413.11	271.88	35.66
Locative case -(du) LE	966	503	885.93	583.07	18.56
SS[9] conditional converb -mi	442	223	401.05	263.95	10.75

To assess the unevenness of the distribution of different morphemes across speakers in each corpus we used the DP metric [Gries 2008]. DP is a measurement that shows how evenly a feature is distributed across corpus parts. Importantly, DP is based on the cumulative difference between the expected and observed numbers of uses in each subcorpus against the total number of its uses in the whole corpus, rather than on the frequency of a category in the subcorpus, as log-likelihood is. Absolute values of pairwise differences are added and divided by two, and the resulting DP value lies between 0 (absolutely even distribution) and 1 (infinitely uneven distribution)[10].

For the Even data, we calculated the evenness of the distribution of a morpheme between the speakers of each dialect separately for each dialect using an in-house Python script. For each of the two corpora, the calculation showed that all frequently occurring suffixes and, importantly, most of the suffixes we were primarily interested in (i.e. the locative and dative case marker as well as converbs) had DP values lower than 0.2. For instance, the DP value for the conditional converb *-REk* is ~0.19 in the Bystraja corpus and ~0.17 in Lamunkhin; that for the locative *-(du)LE* is ~0.09 (Table 2). We interpret this as an indication that the feature is more or less evenly distributed. More precisely, it means that less than 20 percent of the distribution of a suffix is not where it is expected to be under the assumption of its even distribution between the speakers.

[8] DS = different-subject, i.e.subordinate and main clause subjects are non-coreferential

[9] SS = same-subject, i.e. subordinate and main clause subjects are coreferential

[10] DP = sum(abs(exp-obs))/2.

The following table shows the DP values for the suffixes of interest in this study, computed across all speakers in each corpus and compared to the corresponding log-likelihood values. Note that we excluded the multiplicative converb shown in Table 1 because it does not occur in the Bystraja dialect at all, and therefore the DP metric cannot be calculated for this dialect.

Table 2: DP values within corpora for the spatial case markers and converb suffixes compared to log-likelihood values[11]

Suffix	DP Lamunkhin	DP Bystraja	Log-likelihood
locative (nouns) -(du)LE	0.085	0.098	18.56
DS conditional converb -REk	0.165	0.187	35.66
dative (nouns) -Du	0.194	0.168	40.93
purposive converb -DE	0.208	0.155	85.46
SS conditional converb -mi	0.248	0.168	10.75
simultaneity converb -nikEn	0.139	0.294	739.84
anteriority converb -Riʤi	0.149	0.318	95.30
allative (nouns) -t(E)ki	0.360	0.170	360.17

From the data in Table 2 it appears as if the simultaneity converb –nikEn and anteriority converb—Riʤi, as well as the allative case suffix –t(E)ki are quite unevenly distributed in the Bystraja and Lamunkhin corpus, respectively, since the DP values for these suffixes are three times higher than the DP value calculated for the locative case marker (which is similar to the lowest DP value calculated for a sample of common English function words in the BNC Sampler corpus; [Gries 2008: 421]). This therefore indicates that the putative dialectal differentiation emerging in the log-likelihood scores (Table 1) needs to be evaluated with caution, since the observed differences might be due to speaker idiosyncracies rather than to true dialectal differences.

4. Discussion and Conclusion[12]

Dialectal differences between languages may be due to both linguistic divergence (independent innovations such as the loss of inherited categories) and linguistic convergence (innovations due to the influence of other dialects or unrelated languages). Often, descriptions of dialectal differentiation are based on categorical differences between varieties (presence or absence of a category). Categorial changes can, however, be preceded by changes in usage [Johanson 1999: 52]; [Aikhenvald 2002: 238]; [Heine & Kuteva 2005: 50]; frequency of use therefore needs to be taken

[11] The DP values are arranged by increasing order of the higher of the two DP values for each morpheme.

[12] We sincerely thank an anonymous reviewer for her/his thoughtful feedback on our manuscript, which has greatly influenced our interpretation of our results.

into account in evaluations of dialect divergence. However, for any qualitative interpretation of the frequency differences between two linguistic varieties, it is important to ensure that the observed differences reflect true dialectal divergence rather than biases in the data under comparison. This is especially important when working with minority languages where the data are scarce and one cannot from the outset exclude utterances by under- or overrepresented speakers in order to homogenize potentially heterogenous corpora.

It is therefore important when evaluating dialectal differences based on oral corpora of minority languages to first apply methods that take into account potential heterogeneity in speaker contributions. In this paper, we found skewed distributions in the frequency of use of spatial case markers (especially the allative) and of converb suffixes (especially the simultaneous converb) between two small dialectal corpora, as reflected in their very high log-likelihood values (Table 1). While this may indicate dialectal differentiation, the DP metric [Gries 2008] indicates that the difference in their use varies greatly between speakers. Further tests are therefore necessary before one can safely conclude that the apparent frequency differences between the two corpora are indeed due to dialectal differentiation.

A first qualitative assessment of the impact of the heterogeneity in speaker contributions is provided by the breakdown of frequency of use by individual speakers of the three morphemes with the highest DP values: the allative case marker for the Lamunkhin dialect and the anterior and simultaneous converb for the Bystraja dialect (Fig. 3–5). All three suffixes are used less frequently than expected in the respective dialect (Table 1). Were this underuse driven by individual speakers' idiosyncracies, we would expect to find that some of the speakers with large contributions to the corpora use these suffixes less than expected.

As can be seen from the figures, of the ten speakers with the highest contribution to the Lamunkhin corpus, NPZ, and especially TPK and KNK use the allative suffix less often than would be expected merely by the number of tokens they contributed to the corpus. Of the ten speakers with the highest contribution to the Bystraja corpus, LGT, NFI, and AEI use both the simultaneity and the anterior converb less than expected judging from the size of their contributions. Although it is not the speakers with the absolutely largest contributions who appear to be underusing these morphemes, the impact of this potential speaker bias should be evaluated before a qualitative investigation of the putative dialectal differences can be approached. One way to do this would be to exclude these speakers from the corpus of the respective dialect and to redo the log-likelihood calculations to assess the degree of impact of their idiosyncratic behaviour on the differences between the corpora.

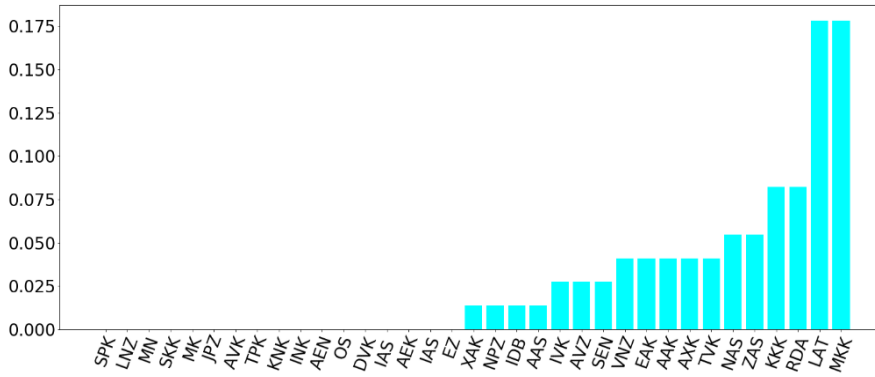

Fig. 3. Frequency of use of allative suffix by individual speakers in the Lamunkhin dialect

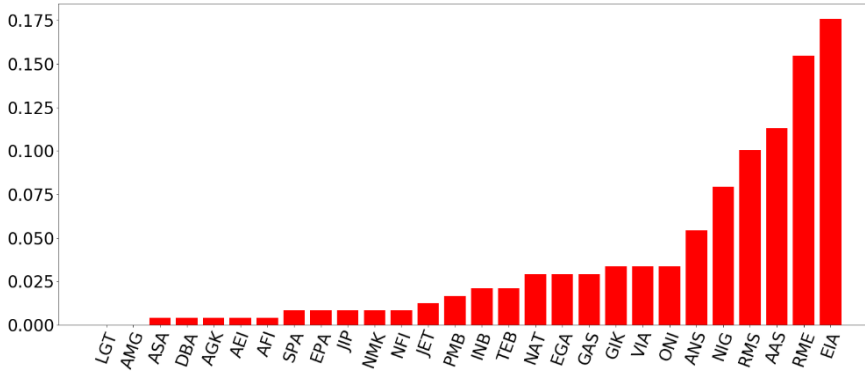

Fig. 4. Frequency of use of anteriority converb by individual speakers in the Bystraja dialect

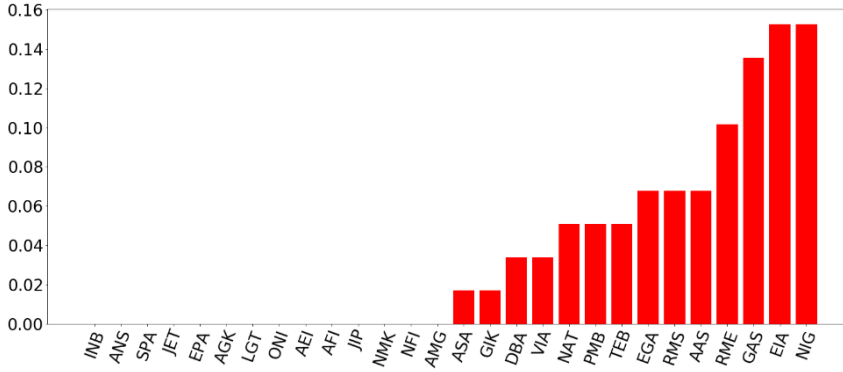

Fig. 5. Frequency of use of simultaneity converb by individual speakers in the Bystraja dialect

The DP metric thus provides a useful means to highlight potential heterogeneities in corpus data that need to be investigated before conclusions can be drawn concerning differences between corpora. However, it is not evident how to interpret the DP values and which values are to be considered high. The DP statistic is relatively sensitive to small numbers [Gries 2008: 423], so that DP values for relatively small corpora, such as those resulting from linguistic documentation projects, are expected to be relatively large—but it is not clear how large, and it is therefore hard to judge which values would be within the normal range for a given corpus. A potential solution to this problem would be to follow the heuristic suggested by [Gries 2008: 423], namely to "evaluate distributional statistics (…) in terms of the ranking of words in comparison to other words rather than their absolute values in isolation". In the corpora studied here, none of the DP values estimated for the morphemes of interest is among the highest ranking values for the overall set of morphemes found in the corpora. For instance, in the Bystraja corpus, the DP value for the anteriority converb only ranks 33rd out of 183 morphemes, i.e. 32 morphemes in the corpus have a higher DP value. Similarly, the DP value for the allative case suffix in the Lamunkhin corpus is in rank 26 (out of a total of 203 morphemes). These observations would seem to attenuate the conclusion that the DP values of 0.31 and 0.36 for the morphemes of interest indicate their substantial underdispersion in the corpora—but this is not sufficiently clear.

It is thus clearly necessary to perform further analyses, both to better understand the performance of the DP metric in small corpora as well as to obtain a trustworthy estimate of true dialectal frequency differences.

References

1. *Aikhenvald, A. Y.* (2002), Language Contact in Amazonia. Oxford, New York: Oxford University Press.
2. *Burykin, Aleksej A.* (2004), Jazyk maločislennogo naroda v ego pis'mennoj forme. Sociolingvističeskie i sobstvenno lingvističeskie aspekty [The language of a minority people in its written form. Sociolinguistic and linguistic aspects]. St Petersburg: Peterburgskoe Vostokovedenie.
3. *Gries S. T.* (2008), Dispersions and adjusted frequencies in corpora, International journal of corpus linguistics, Vol. 13, 4, pp. 403–437.
4. *Haspelmath M., König E.* (eds.) (1995), Converbs in cross-linguistic perspective. Structure and meaning of adverbial verb forms. Berlin—New York: Mouton de Gruyter.
5. *Heine, B., Kuteva, T.* (2005), Language Contact and Grammatical Change. Cambridge: Cambridge University Press
6. *Johanson, L.* (1999), The dynamics of code-copying in language encounters. In: Brendemoen, B. et al. (eds), Language encounters across time and space. Studies in language contact. Oslo: Novus forlag: pp. 37–62.
7. *Matić, D., Pakendorf, B.* (2013), Non-canonical SAY in Siberia: Areal and genealogical patterns. Studies in Language Vol. 37, 2, pp. 356–412.

8. *Pakendorf, B.* (2009), Intensive Contact and the Copying of Paradigms: An Ėven Dialect in Contact with Sakha (Yakut). Journal of Language Contact Varia 2, pp. 85–110.
9. *Pakendorf, B.* (to appear), Expressing equality, similarity, and pretense in Even (Northern Tungusic, Siberia). To appear in : Treis, Y., Chamoureau, C. (eds) : special issue of Faits des Langues on "Comparaisons d'égalité et de similitude et expression de la simulation".
10. *Pakendorf, B., Krivoshapkina, I. V.* (2014), Ėven nominal evaluatives and the marking of definiteness, Linguistic Typology Vol. 18, 2, pp. 289–331.
11. *Rayson P., Garside R.* (2000), Comparing corpora using frequency profiling, Proceedings of the workshop on Comparing Corpora, Association for Computational Linguistics, pp. 1–6.
12. *Rišes, L. D., Cincius, V. I.* (1952), Kratkij očerk grammatiki ėvenskogo (lamutskogo) jazyka. [A short grammar sketch of the Even (Lamut) language.]. Russko-Ėvenskij Slovar', pp. 693–777. Moscow: Gosudarstvennoe izdatel'stvo inostrannyx i nacional'nyx slovarej.