Квантитативный анализ переключения кодов в четырех языках России

В.В. Дьячков, П.С. Плешак, Н.М. Стойнова, И.А. Хомченкова

“Проблемы языка - 2020”, 19-20 марта 2020 г., ИЯз РАН, Москва

Исследование поддержано грантом РФФИ №18-312-00155мол.а «Переключение кодов в речи русскоговорящих носителей малых языков России: комплексное исследование четырех контактных ситуаций»

Введение: переключение кодов

- Количественное корпусное исследование переключения кодов в языках России:
  - между автохтонным языком (IL) и русским (RUS).
- На материале четырех коллекций спонтанных устных текстов:
  - двух – на уральских языках (мокшанском и горномарийском);
  - двух – на тунгусо-маньчжурских (нанайском и ульчском).
- В этом докладе: часть результатов исследования.

Введение: переключение кодов

Переключение кодов (code-switcing, CS): в широком понимании, с акцентом на структурные аспекты

→ см. Poplack 1980; Myers-Scotton 1993; 2002; Muysken 2000, ...

- межклаузальное (1) и внутриклаузальное (смешение кодов);
- однословное (2) и неоднословное (3);
- внутрисловное (4);
- в т.ч. случаи, промежуточные между переключением кодов и заимствованием.

Sugdatawa=də gəsə-gəsə burməči. (1) *А то кто кормить будет. Кормильца одна я.* (2) *Господи, tara* (3) *в сорок шестом году ǯiǯihən min amin= ambi, bəgdin* (4) *рана-ču=də.*
‘*Рыбу-то между собой делили. А то кто кормить будет. Кормилец одна я. Господи, потом в сорок шестом году вернулся мой отец с ран-ой на ноге.* (aid, ульчский)

Введение: разметка переключения кодов

- Готовые текстовые коллекции (транскрипция + перевод + поморфемное глоссирование: не для всех).
- Для исследования переключения кодов мы добавили к ним:

1) **(полу)автоматическую пословную разметку языка:**
    - для каждого слова - RUS vs. IL.

NB внутрисловные переключения (русская основа с аффиксами IL) размечены как RUS.

2) **ручную разметку структурных типов переключения кодов:**
    - **по объему**: внутрисловное / однословное / неоднословное / клауза
    - **по синтаксическому типу**: NP, ADJ, ADV, NUMP, CONJ, DISC и др.
    - особо помечены клаузы с русской вершиной;
    - особо помечены случаи, когда переключенный фрагмент не образует составляющую;
    - дополнительные пометы: имя собственное, дублирование, flagging, дефолтная форма и проч.

Введение: задачи

Пословная разметка языка:

- очень простая и грубая характеристика CS;
- сделана для большего количества текстов;
- в этом докладе будет обсуждаться в основном она;
- к ней мы применили несколько количественных мер, позволяющих охарактеризовать тип переключения кодов.

Вопросы:

- ? Дают ли меры, основанные на разметке языка, разные результаты для 4 коллекций?
- ? Коррелируют ли различия с различиями между языками (социолингвистическими)?
- ? Коррелируют ли результаты с результатами ручной разметки типов CS (=насколько информативна такая грубая характеристика CS)?

План доклада

1. Материал исследования: языки, корпуса
2. Разметка типов переключения кодов: основные результаты (кратко)
3. Разметка языка: простые меры переключения кодов (подробно)
4. Результаты и обсуждение

1. Материал исследования: языки

2 уральских языка:

- **горно-марийский** (уральские, марийские, Марий-Эл, 23 062 носителей по переписи-2010);
- **мокшанский** (уральские, мордовские, Мордовия и сопредельные регионы, 2025 носителей по переписи-2010), см. Коряков & Холодилова (2018).

2 тунгусо-маньчжурских языка:

- **нанайский** (тунгусо-маньчжурские, южно-тунгусские, Хабаровский край, 1347 носителей по переписи-2010, 11 % от этнической группы),см. Герасимова (2002); Калинина & Оскольская (2016);
- **ульчский** (тунгусо-маньчжурские, южно-тунгусские, Хабаровский край, 154 носителя по переписи-2010, 6 % от этнической группы), см. Герасимова (2002); Сумбатова & Гусев (2016).

1. Материал исследования: языки

- Для всех языков значительная часть или все носители владеют также русским. Та или иная стадия языкового сдвига.
- Разные ситуации с т.зр. **доминирования** автохтонного языка (IL) vs. русского:
  - от стабильного сбалансированного билингвизма для носителей горномарийского до доминирующего русского практически для всех носителей ульчского.

стадия языкового сдвига: доминирующий RUS ← доминирующий IL
(1) ульчский < нанайский < мокшанский << горно-марийский

1. Материал исследования: языки

- **По уровню компетенции носителей**: для носителей, вошедших в выборку, различий нет (все свободно владеют русским и владеют IL как минимум достаточно свободно для порождения спонтанного текста).

NB Несмотря на различия в доминирующем языке (см. о том, что эти параметры не всегда коррелируют и могут давать разные эффекты на переключение кодов, Benthalia & Davies 1992).

- **По степени структурной близости к русскому**: принципиально не различаются.

NB Хотя из двух разных языковых семей.

1. Материал исследования: корпуса

- **Горномарийский и мокшанский**: коллекции, собранные и обработанные участниками коллективных экспедиций ОТиПЛ МГУ;
- **Нанайский и ульчский**: коллекции, собранные и обработанные С.А. Оскольской и Н.М. Стойновой.
- **По типам текстов и жанрам**: коллекции сопоставимы:
  - устные спонтанные тексты, в осн. монологи, по заданию лингвиста “расскажите на своем языке”;
  - NB особый искусственный тип текстов, во всех IL заведомо преобладает (даже если те же носители в обычной ситуации используют исключительно русский);
  - биографические рассказы, фольклор, этнографические описания и под.
- **По объему**: коллекции сопоставимы.

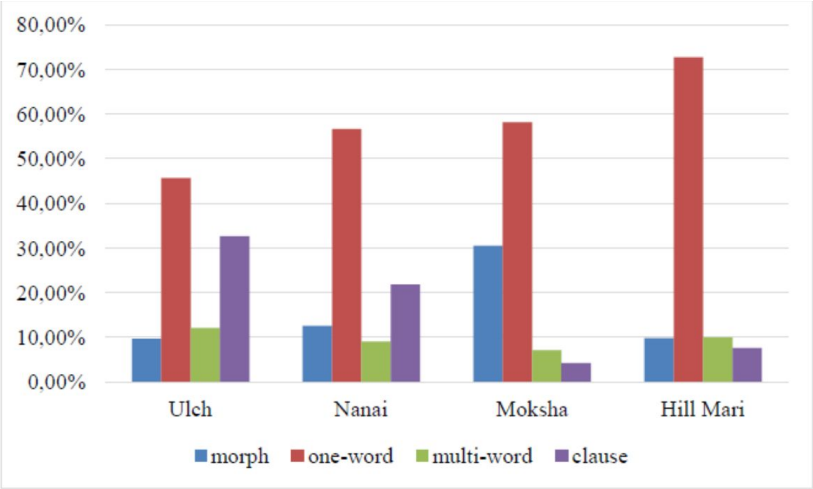
1. Материал исследования: корпуса

	с разметкой LANG, сл.-употр. (текстов)	с разметкой CS_TYPE, сл.-употр. (текстов)	другие характеристики коллекции
ульчский	47 509 (179)	11 334 (50)	синхронизовано со звуком
нанайский	47 411 (167)	16 368 (52)	синхронизовано со звуком
мокшанский	17 578 (53)	17 578 (53)	отглоссировано
горномарийский	15 901 (17)	15 901 (17)	отглоссировано

2. Разметка по типам переключения кодов: основные результаты

Объем переключенного фрагмента:

- % межклаузального переключения: ульчский > нанайский > мокшанский ≈ горномарийский;
- % однословных: ульчский < нанайский ≈ мокшанский < горномарийский.

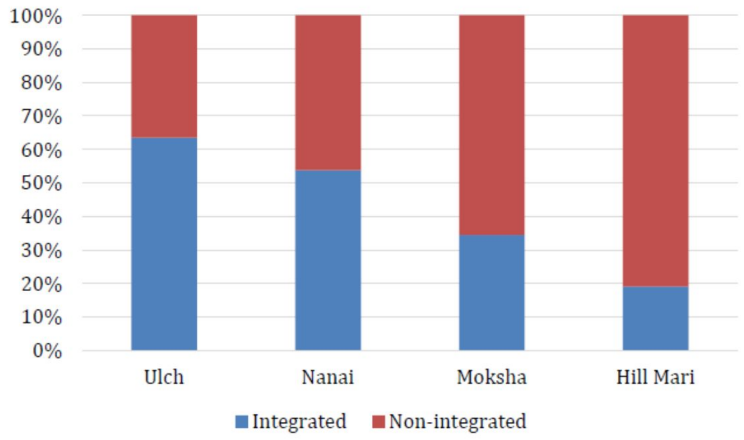


- % внутрисловного переключения: резко повышен в мокшанском.

2. Разметка по типам переключения кодов: основные результаты

Степень синтаксической интеграции переключенного фрагмента:

- высокая: np, pp, nump
- низкая: disc, pred, interj, voc, conj, adv



- % синтаксически интегрированных русских вставок:

ульчский > нанайский > мокшанский > горномарийский

3. Меры CS, основанные на простой разметке языка

- **5 количественных мер переключения кодов**, описанных в Guzmán et al. (2016a; b 2016b; 2017) (см. также Barnett et al. 2000; Gardner-Chloros et al. 2007; Gambäck & Das 2014, 2016; Goh & Barabási 2008):
  - **Multilingual Index** и **Integration Index**: характеризуют распределение слов на IL vs. RUS;
  - **Burstiness**, **Memory** и **Span Entropy**: характеризуют распределение последовательностей слов (фрагментов, spans) на IL vs. RUS.

3. Меры CS, основанные на простой разметке языка

Задачи:

- 1) Проверить, как соотносятся с социолингвистическими параметрами

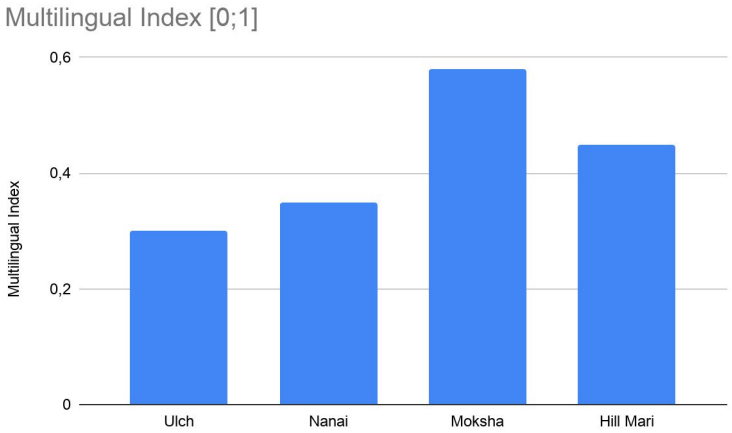
→ Соблюдается ли иерархия языкового сдвига?
ульчский < нанайский < мокшанский << горно-марийский?

- 2) Проверить, как соотносятся с результатами более детальной разметки по структурным типам переключения кодов

→ Можно ли “поймать” разницу в структурных типах CS с помощью этих простых мер (основанных на самой примитивной автоматической разметке)?

3. Меры CS: Multilingual Index $MI = \frac{1 - \sum p_j^2}{\sum p_j^2}$

- Характеризует соотношение слов на IL и RUS: [0 - весь текст на IL; 1 - равное количество слов на IL и RUS].

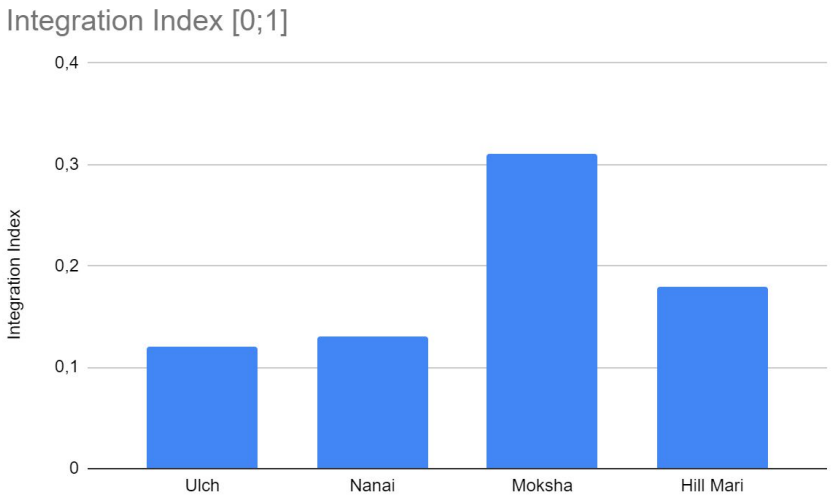


- Для всех языков: заметное преобладание IL (max - 0,58).
- Социолингвистической иерархии не соответствует.
- Выделяется мокшанский (больше RUS): как и по обилию внутрисловных переключений.

3. Меры CS: Integration Index

$$II = \frac{1}{n-1} \sum_{1 \le i < j \le n} S(l_i, l_j)$$

- Вероятность слова на IL или RUS: характеризует “равномерность” распределения слов на IL и RUS [0 - неравномерно; 1 - равномерно].

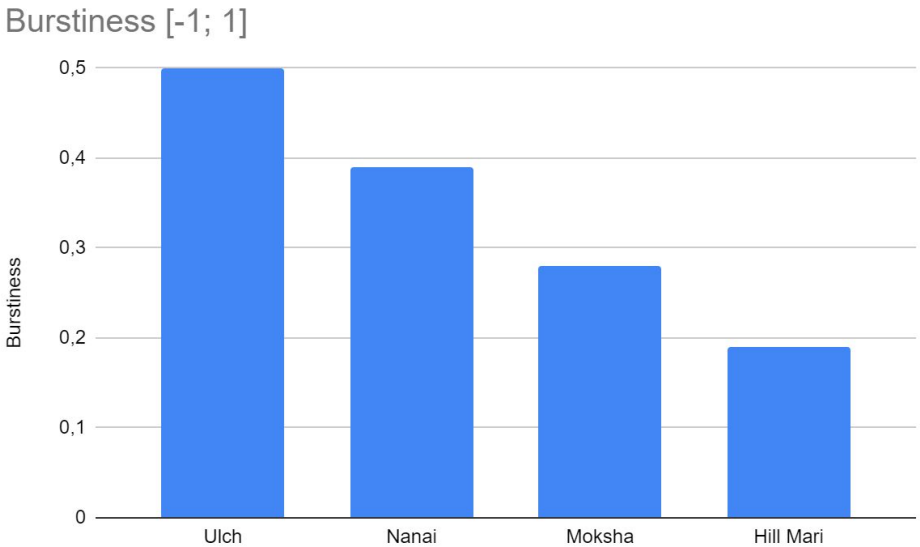


- Для всех выборок скорее неравномерно (max - 0,31).
- Социолингвистической иерархии не соответствует.
- Выделяется мокшанский (наиболее равномерно): как и по обилию внутрисловных переключений.
- Результаты такие же, как для Multilingual Index.

3. Меры CS: Burstiness

$$Burstiness = \frac{(\sigma_\tau - m_\tau)}{(\sigma_\tau + m_\tau)}$$

- Характеризует периодичность переключения: [-1 - фрагменты на IL vs. RUS периодичны, “как кардиограмма”; 1 - хаотичны].

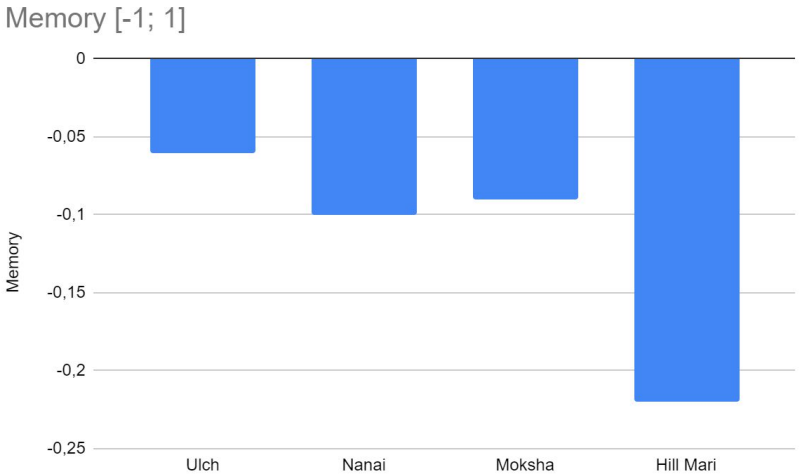


- Во всех выборках не очень периодичны: [0;0,5].
- Соответствует социолингвистической иерархии (ульч.: наиболее периодично).
- Соответствует % межклаузального переключения.
- *И соответствует % синтаксически интегрированных вставок.*

3. Меры CS: Memory

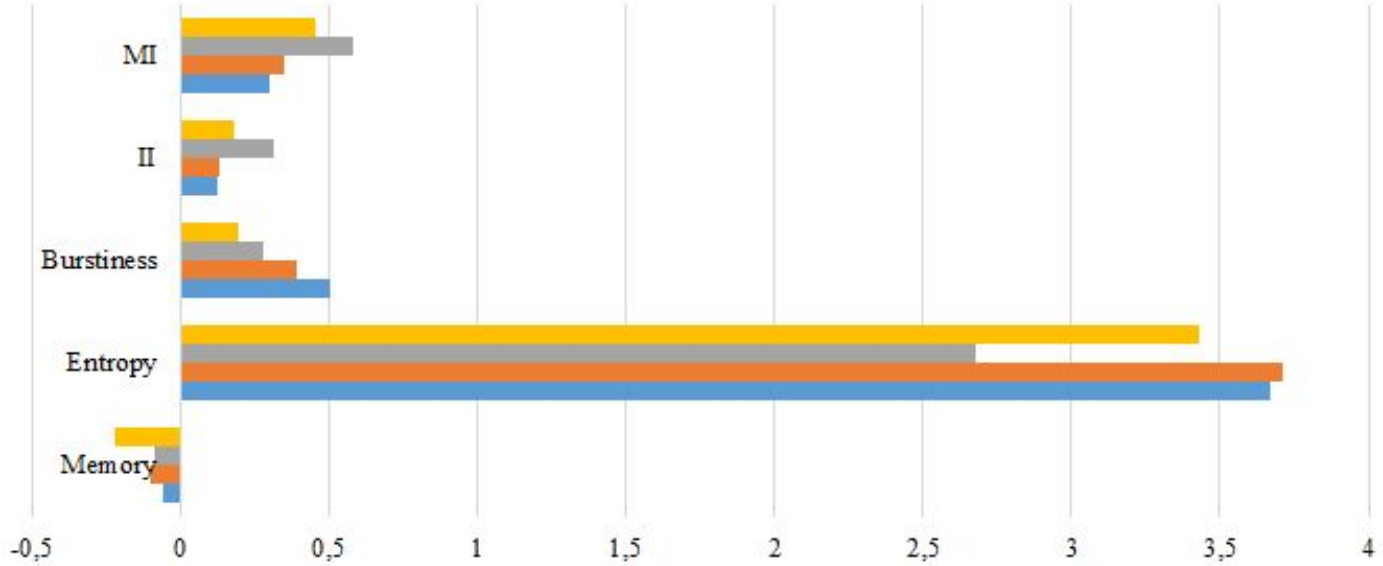
$$Memory = \\ = \frac{1}{n_r - 1} \sum_{i=1}^{n_r - 1} \frac{(\tau_i - m_1)(\tau_{i+1} - m_2)}{\sigma_1 \sigma_2}$$

- Характеризует, насколько длина переключенного фрагмента на RUS коррелирует с длиной предыдущего на IL: [-1 - длинные после коротких; 1 - длинные после длинных].



- Результаты, симметричные результатам Burstiness.
- Для всех выборок: скорее не коррелирует: [-0,2;0].

3. Меры CS: результаты



	Memory	Entropy	Burstiness	II	MI
Hill Mari	-0,22	3,43	0,19	0,18	0,45
Moksha	-0,09	2,68	0,28	0,31	0,58
Nanai	-0,1	3,71	0,39	0,13	0,35
Ulch	-0,06	3,67	0,5	0,12	0,3

4. Результаты и обсуждение

Общее для всех корпусов:

→ корпуса текстов, где носитель осознанно контролирует / старается контролировать выбор языка (задача “рассказать на IL”)

- заметное преобладание IL над русским (Matrix Index)
- русские слова распределены по текстам неравномерно (Integration Index);
- русские фрагменты распределены по текстам неравномерно и непредсказуемой длины (Burstiness, Memory).

4. Результаты и обсуждение

Различия между корпусами:

- есть, хотя и не очень резкие;
- единообразные для мер, ориентированных на отдельные слова (Matrix Index, Integration Index);
- единообразные для мер, ориентированных на последовательности слов (Burstiness, Memory).

4. Результаты и обсуждение

С социолингвистической иерархией (стадией языкового сдвига):

(1) ульчский < нанайский < мокшанский << горно-марийский

- не коррелируют меры, ориентированные на отдельные слова;
- (опосредованно?) коррелируют меры, ориентированные на последовательности слов:
  - доминирующий русский ⇒ русские фрагменты распределены более хаотично (Burstiness);
  - доминирующий русский ⇒ длина русского фрагмента менее предсказуема (Memory).

4. Результаты и обсуждение

Отражают ли меры, основанные на простой пословной разметке языка, различия в структурных типах CS?

- меры, основанные на отдельных словах, выделяют мокшанский:
- больше всего русских слов и они наиболее равномерно распределены по тексту;
- ~ в мокшанском наибольший % внутрисловных переключений: эффект за счет них?

4. Результаты и обсуждение

Отражают ли меры, основанные на простой пословной разметке языка, различия в структурных типах CS?

- результаты мер, основанные на последовательностях слов, коррелируют с % межклаузальных переключений:
  - в коллекциях, где больше межклаузальных переключений, русские фрагменты распределены более хаотично и их длина меньше коррелирует с длиной предшествующего фрагмента на IL
- результаты мер, основанные на последовательностях слов, коррелируют с % синтаксически интегрированных внутриклаузальных переключений:
  - скорее случайный эффект.

Литература

- Barnett R., Codo E., Eppler E., Forcadell M., Gardner-Chloros P., van Hout R., Moyer M., Torras M. C., Turell M. T., Sebba M., Starren M., Wensing S. (2000), The LIDES Coding Manual: A document for preparing and analyzing language interaction data, Version 1.1–July, 1999. International Journal of Bilingualism, 4(2), pp. 131–132.
- Bentahila A., Davies E.D. (1992), Code-switching and language dominance // Harris, R.J. (Ed.) Cognitive processing in bilinguals, Amsterdam: Elsevier, pp. 443–58.
- Gambäck B., Das A. (2014), On measuring the complexity of code-mixing, Proceedings of the 11th International Conference on Natural Language Processing, Goa, India, pp. 1–7.
- Gambäck B., Das A. (2016), Comparing the level of code-switching in corpora, Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), pp. 1850–1855.
- Gardner-Chloros, Penelope, Moyer, Melissa, and Mark Sebba (2007), Coding and Analysing Multilingual Data: The LIDES Project // Creating and Digitizing Language Corpora, pp. 91–120.
- Gerasimova A.N. (2002), Nanai and Ulch in Russia: a comparative characteristics of the sociolinguistic situation [Nanajskij I uljčskij jazyki v Rossii: sravniteljnaja harakteristika sociolingvističeskoj situacii], Jazyki Korennyh narodov Sibiri [Languages of Indigenous Peoples of Siberia], 12.
- Guzmán G.A., Serigos J., Bullock B., Toribio A. J. (2016), Simple tools for exploring variations in code-switching for linguists, EMNLP-2016: Proceedings of the Second Workshop on Computational Approaches to Code Switching, pp. 12–20.
- Guzmán G.A., Serigos J., Bullock B., Toribio A. J. (2017a), Moving code-switching research toward more empirically grounded methods, Proceedings of the Workshop on Corpora in the Digital Humanities (CDH 2017), Bloomington, IN, USA, pp. 1–9.

Литература

- Guzmán G.A., Serigos J., Bullock B., Toribio A. J. (2017b), Metrics for modeling code-switching across corpora, Proc. Interspeech 2017, pp. 67–71.
- Goh K.I., Barabási A. L. (2008), Burstiness and memory in complex systems, EPL (Europhysics Letters), 81(4): 48002.
- Kalinina E. Ju., Oskolskaya S.A. (2016), Nanai [Nanajskij jazyk], Language and Society. An Encyclopaedia [Jazyk i obščestvo. Ènciklopedija], Azbukovnik, Moscow, pp. 293–296.
- Muysken P. (2000), Bilingual speech: A typology of code-mixing, Cambridge University Press, Cambridge/New York.
- Myers-Scotton C. (1993), Duelling languages: Grammatical structure in codeswitching, Oxford University Press, Oxford/New York.
- Myers-Scotton C. (2002), Contact linguistics: Bilingual encounters and grammatical outcomes, Oxford University Press, Oxford/New York.
- Poplack Sh. (1980), ‘Sometimes I ’ll start a sentence in Spanish Y TERMINO EN ESPAÑOL’, Linguistics 18, pp. 581–618.
- Sumbatova N.R., Gusev V.Yu. (2016), Ulch [Ul’čskij jazyk], Language and Society. An Encyclopaedia [Jazyk i obščestvo. Ènciklopedija], Azbukovnik, Moscow, pp. 513–515.