Неологизмы в социальной сети Фейсбук

Муравьев Н.А., Панченко А.И., Объедков С.А.

nikita.muraviev@gmail.com

Лаборатория цифрового общества, Москва

Лаборатория цифрового общества

- Маркетинговые, социальные, лингвистические исследования социальных сетей
- Сбор, обработка и анализ больших массивов данных
- Профилирование рекламы в соцсетях
- Разработка интернет-приложений нового поколения

Актуальность исследований заимствований и неологизмов

- Состояние вопроса
- **Широкий круг теоретических и сравнительных исследований по заимствованиям и неологизмам**
- Haugen 1950, «The analysis of linguistic borrowing»
- Capuz 1997, Winter-Froemel 2008, Peperkamp&Dupoux 2003, LaCharité&Paradis 2005 и многие другие
- **Конкретно языковые исследования, русский язык**
- Крысин 1968 "Иноязычные слова в современном русском языке"
- Брейтер 1997, Дьяков 2003, Маринова 2013

Заимствования и неологизмы в соцсетях

- Специфичность языка интернета
- Интернет-дискурс как особый промежуточный модус между устным и письменным
- Особый пласт интернет-лексики (*трекер, паблик, репост*)
- Высокая интенсивность изменений в лексике и грамматике языка
- Значительное влияние английского языка
- Специфика коммуникации в социальной сети – посты, комментарии, лайки и т.д.

Заимствования и неологизмы в соцсетях

- Задачи нашего исследования
- Создание исследовательской базы – корпуса и словаря на основе постов в соц. сетях
- Фиксирование текущего состояния языка соц. сетей для отслеживания изменений
- Получение количественных данных
- Вынесение **предварительных** суждений о текущем состоянии языка соц. сетей и языка вообще

Сбор и обработка и анализ материала

- **Источник:** русскоязычный публичный сегмент социальной сети, сообщения пользователей с «открытым» профилем https://developers.facebook.com/tools/explorer
- **Метод извлечения данных:** использование программного интерфейса (API) Фейсбука
- Первый этап: сбор корпуса анонимизированных постов и комментариев Фейсбука
- Второй этап: посторение словаря на основе корпуса полуавтоматическим способом
- Третий этап: лингвистический анализ данных

Первый этап: сбор корпуса

- Всего 573 миллионов анонимизированных постов и сообщений 3,2 млн. пользователей социальной сети
- Только тексты на русском языке
- Язык каждого из входных текстов был определён автоматически при помощи модуля *langid.py* [Lui & Baldwin 2012].
- Посты за период с 2006 по 2013 год.
- Первый пост датируется 5 августа 2006 года, последний – 13 ноября 2013 года.
- Подавляющее большинство постов приходится на период с 2011 по 2013 год.

Статистика корпуса

Параметр	Значение
Количество анонимизированных пользователей	3,190,813
Язык	Русский
Количество постов	426,089,762
Количество комментариев	147,140,265
Количество текстов (посты и комментарии)	573,230,027
Количество словоформ в постах	20,775,837,467
Количество словоформ в комментариях	2,759,777,659
Количество словоформ (посты и комментарии)	23,535,615,126
Средняя длина поста, словоформ	49
Средняя длина комментария, словоформ	19

Второй этап: построение словаря

- Морфологическая обработка данных
- Токенизация и лемматизиация при помощи морфологического анализатора, основанного на словаре *AOT* на базе словаря А.А.Зализняка [Зализняк 1977].
- Мы использовали собственную *MapReduce* реализацию модуля построения частотного словаря, основанную на модуле морфологического анализа *RussianMorphology*
- Аннотация при помощи морфологического анализатора *PyMorphy*
- Каждой лемме была присвоена часть речи

Второй этап: построение словаря

- Автоматическая фильтрация
- Для каждой леммы было указано, входит ли она в словарь *OpenCorpora* (основан на словаре *AOT*, содержит информацию о 388,790 леммах и 5,094,925 словоформах).
- Отсев имеющихся в этом словаре единиц
- НО: Произошел отсев новых слов, омонимичных существующим словам русского языка (пример: слово «пост»)

Второй этап: построение словаря

- Экспертное удаление «шума»
- Ручная фильтрация 10,000 наиболее частотных слов.
- Отсев разнообразного «шума»: нерусские слова, неверно лемматизированные слова и другие артефакты автоматической обработки (*ть, нибыть, гый, санкт, що, ул, пр, нью, грн, ца, рожение, т.д, від, україни, вебинара*).
- Отбор из полученного списка 624 наиболее частотных несловарных слов

Второй этап: построение словаря

- Экспертное удаление нерелевантных единиц
- В списке из 624 выявлено большое количество несловарных слов, которые не относятся к неологизмам:
  - Имена собственные
  - «Сниженная» лексика
  - Слова вроде «авиаперелет», «переориентироваться» или «однодневный», которые не содержатся в *OpenCorpora*, но есть в других словарях.
- Дополнительная ручная фильтрация списка по данным *Яндекс.Словари* и НКРЯ с 2000 года (поиск по всему корпусу)

Второй этап: построение словаря

- Итог – список из 168 популярных неологизмов, извлеченных из корпуса текстов социальной сети.
- 10 наиболее частотных лексем: *вконтакт, твиттер, перепост, скайп, фейсбук, плэйкаст, демотиватор, плейлист, личка, перепостить*

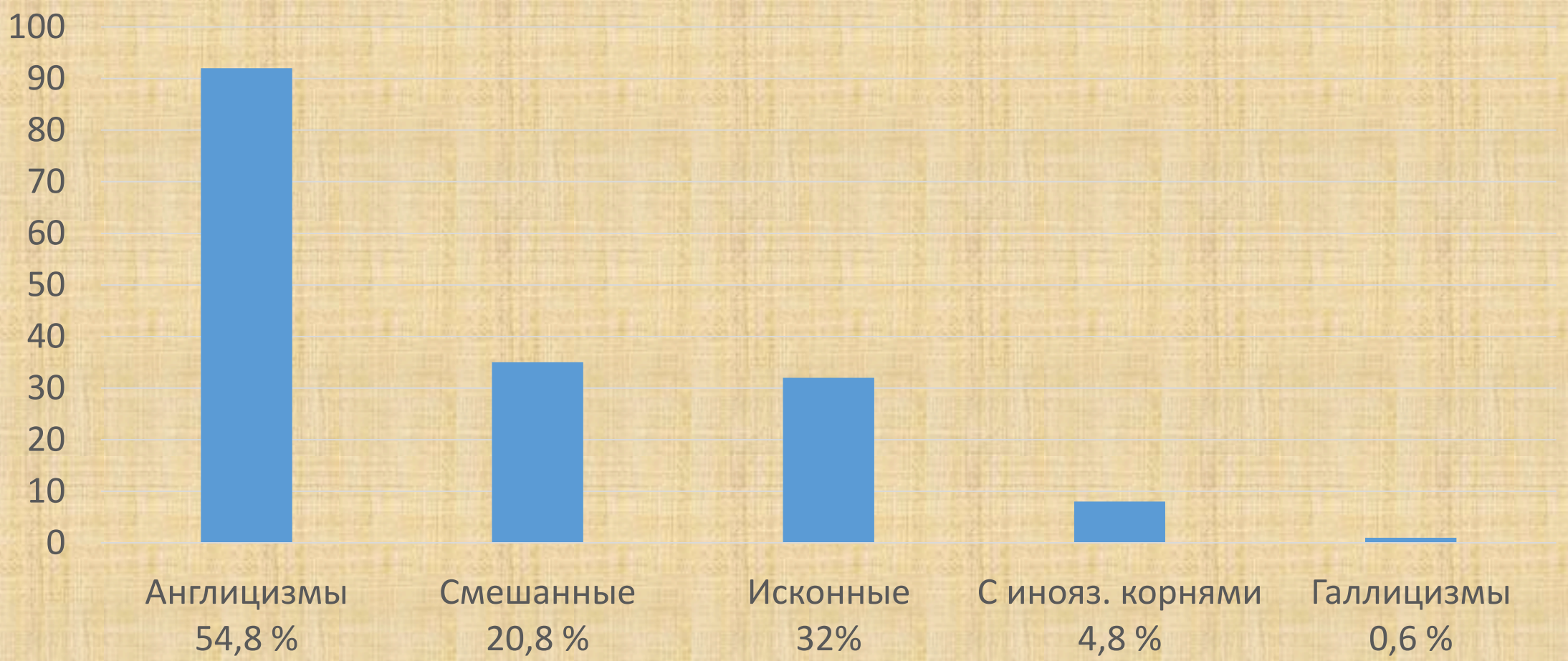
Третий этап: лингвистический анализ материала

- Краткая этимологическая, морфологическая, синтаксическая и семантическая характеристика слов
- Классификация по 4 основаниям
- Тип (источник) заимствования
- Часть речи
- Тип и модель деривации
- Тематика

Тип заимствования

- Исконные: печенька, улыбизм, приколист и др.
- Слова с иноязычными корнями: мульт, видеорепортаж, евроинтеграция
- Англицизмы: фолловер, битрейт, флешмоб и др.
- Галлицизмы: декупаж
- Смешанные слова: ***пере***пост, имхо***нуть***, реферал***ьный***

Тип заимствования (кол-во слов из 168, %)



Тематика

- Интернет: оффлайн, браузерный, мем
- Оценка: суперский, треш/трэш, жжот
- Маркетинг: реферал, продакшн, инфопродукт
- Мультимедиа: фотопроект, плэйкаст, гифка
- Техника: айпад, ноут, флешка
- Культура: флешмоб/флэшмоб, демотиватор

Тематика (кол-во слов из 168, %)

Тематика и этимология

- Поле **«интернет»** представлено практически исключительно англицизмами и образованными от них словами
- Большинство слов полей **«маркетинг»** и **«техника»** – тоже англицизмы
- 16 из 25 слов поля **«оценка»** состоят из исконно русских корней (= 0,5 от общего числа исконных)
- Поле **«мультимедиа»** состоит в основном из слов с иноязычнми корнями и смешанных слов (обычно композиты от *фото-*, *видео-*, *аудио-*, *теле-*)

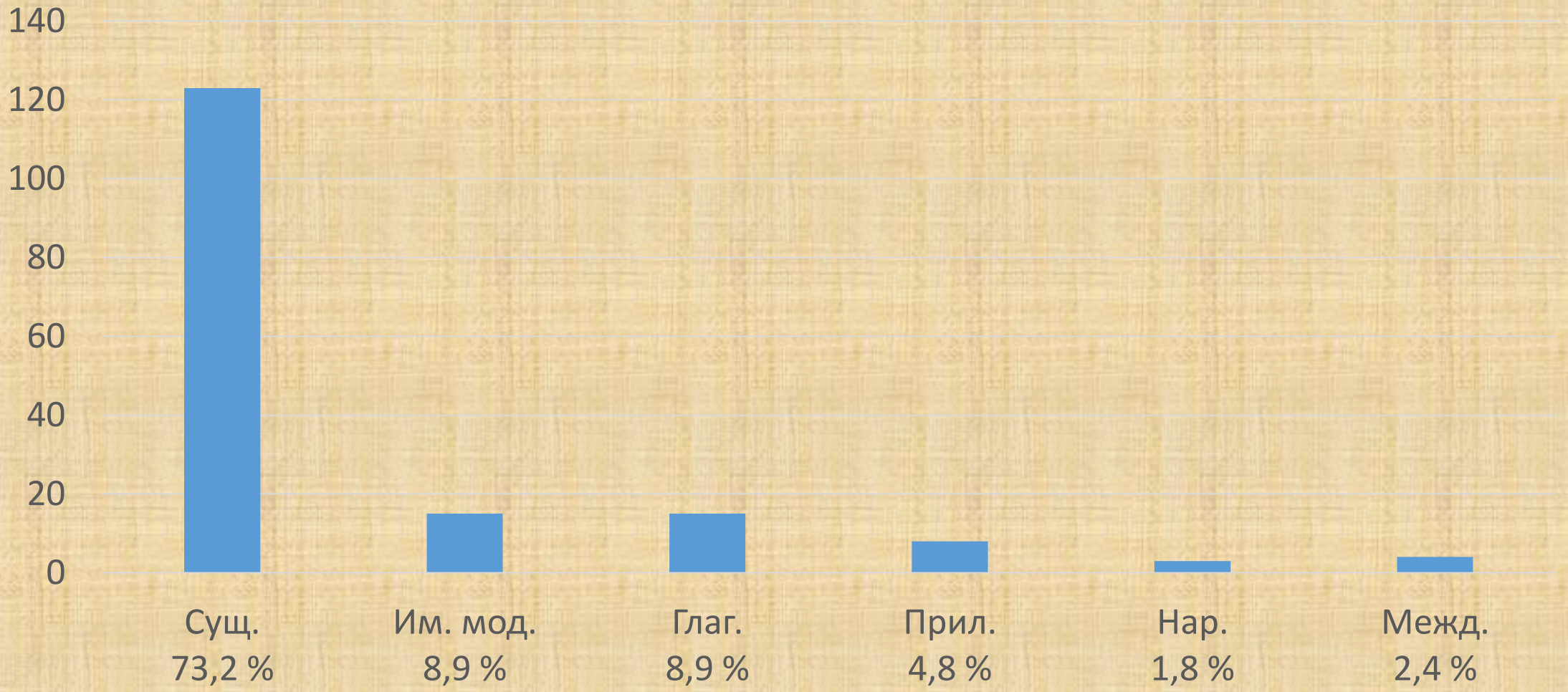
Часть речи

- Традиционные классы (N, V, Adj, Adv, Interj)
- Дополнительный класс (Nmod):
  - ***брэйн****-система,* ***лайф****-коуч,* ***трэш****-комедия,* ***байк****-центр*
  - Довольно распространенный в нашем материале (15 слов из 168)
  - Промежуточный класс между сущ. и прил.
  - Морфологически не изменяемые
  - Модифицируют существительные
  - Некоторые не имеют самостоятельного употребления в позиции существительного (этот трэш/*этот брейн/*этот лайф)

Часть речи

- Всего из 168 слов (а точно ли не 165?):
- Существительные (N): 123 (73,2 %)
- Именные модификаторы (Nmod): 15 (8,9 %)
- Глаголы (V): 15 (8,9 %)
- Прилагательные (Adj): 8 (4,8 %)
- Наречия (Adv): 3 (1,8 %)
- Междометия (Interj): 4 (2,4 %)

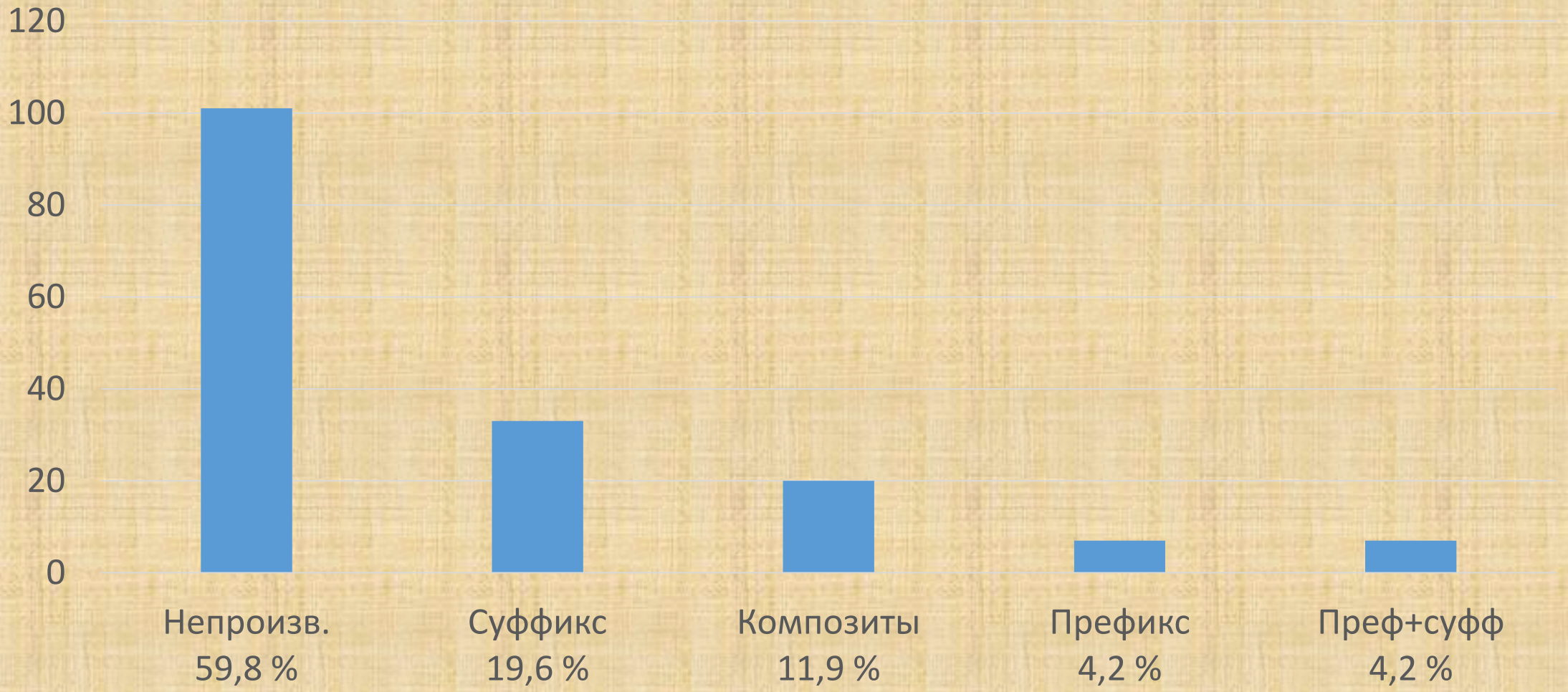
Часть речи (кол-во слов из 168, %)



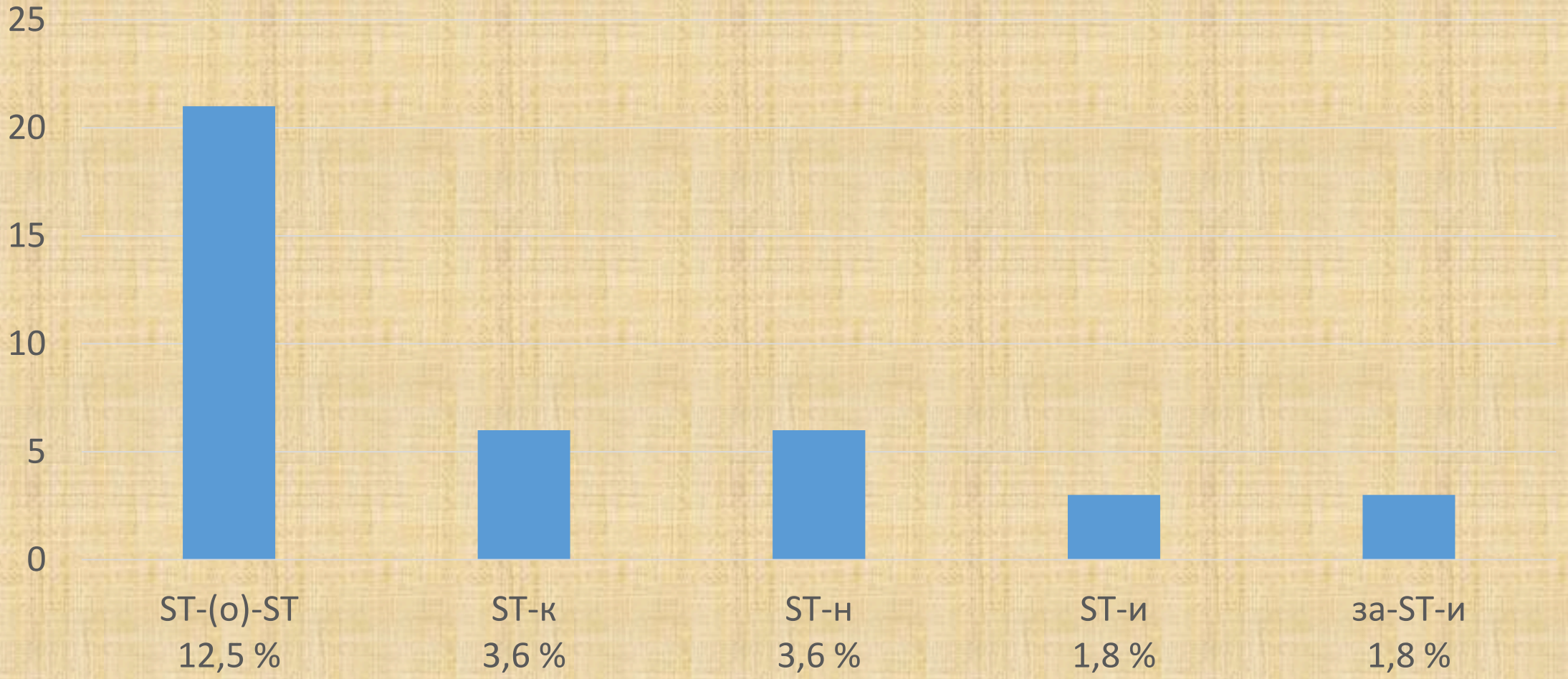
Тип и модель деривации

- Традиционное деление на префиксальный, суффиксальный, префиксально-суффиксальный и композитный способ словообразования
- Модели словообразования, разнообразие аффиксальных средств
- Суффикс, ST-к: флешка, гифка, личка
- Префикс, пере-ST: перепост
- Префикс+суффикс, за-ST-и: забанить, запостить
- Словосложение, ST-ST: фотоподборка, госуслуга

Типы деривации (кол-во слов из 168, %)



Модели деривации (кол-во слов из 168, %)



Тип и модель дериваци

- Консерватизм в деривационных моделях
- Единственная инновация – модель ST-ч
- Предположительно, выглядит как V (inf) + j
- В нашем материале только одно слово *ржач*
- Известно также о существовании некоторых других: *срач*, *махач*

Некоторые итоги и обобщения

- Опыт построения корпуса и словаря
- База для дальнейших исследований
- Теоретические наблюдения касательно современного состояния русского языка в области лексики
  - Значительное влияние английского языка
  - Семантические поля с англоязычным vs исконным наполнением
  - Консерватизм в грамматических свойствах и деривационных моделях

Дальнейшие направления исследований

- Корпусные исследования языка соцсетей
- Временная динамика языковых изменений
- Социология и социолингвистика пользователей
- Взаимосвязь языковых изменений и событий реального мира
- Автоматическая обработка новых, некодифицированных лексических единиц
- Сравнительный анализ данных Фейсбука и В контакте

Спасибо за внимание!

Литература

- *Брейтер М. А.* (1997), Англицизмы в русском языке: история и перспективы: Пособие для иностранных студентов-русистов. // Владивосток, Диалог-МГУ.
- *Дьяков А. И.* (2003), Причины интенсивного заимствования англицизмов в современном русском языке. // Язык и культура. - Новосибирск. - С. 35-43
- *Зализняк, А. А.* (1977). Грамматический словарь русского языка: словоизменение: около 100 000 слов. // Изд-во" Русский язык".
- *Крысин Л.П.* (1968) *Иноязычные слова в современном русском языке - М.*: Наука
- *Маринова. Е.В.* (2013) Иноязычная лексика современного русского языка: учеб. пособие, 2-е изд., М. : ФЛИНТА
- *Capuz Juan Gómez.* (1997), Towards a Typological Classification of Linguistic Borrowing (Illustrated with Anglicisms in Romance Languages) // Revista Alicantina de Estudios Ingleses 10: 81-94

Литература

- *Haugen E*. (1950), The analysis of linguistic borrowing. // Language: 210-231
- *LaCharité, D. & Paradis, C.* (2005). Category Preservation and Proximity versus Phonetic Approximation in Loanword Adaptation. // Linguistic Inquiry 36/2, 223-258.
- *Lui M. and Baldwin T.* (2012). langid. py: An off-the-shelf language identification tool. // In Proceedings of the ACL 2012 System Demonstrations, pages 25–30. Association for Computational Linguistics.
- *Peperkamp, S. & Dupoux, E.* (2003). Reinterpreting loanword adaptations: The role of Perception // Proceedings of the 15th International Congress of Phonetic Sciences 2003, 367-370.
- *Winter-Froemel E.* (2008), *Studying loanwords and loanword integration*: *Two criteria* of *conformity //* Newcastle Working Papers in Linguistics 14: 156–176.