С. А. Бурлак
Институт востоковедения РАН; svetlana.burlak@bk.ru

Устойчивость и частотность: есть *ли* корреляция?

Значение слова может быть названо базисным, если для носителей легче вспомнить уже существующее в языке слово с этим значением, чем заменять его заимствованием или производным. Устройство человеческого мозга таково, что проще вспоминается то, что встречается чаще (поскольку задействованные в этом нейронные связи усиливаются в процессе использования, согласно правилу Хебба). Эта гипотеза была проверена на 15 частотных словарях языков разных семей и ареалов. Двадцать одно сводешевское значение выражается словом из первой тысячи по частотности во всех рассмотренных языках; больше 60 слов почти в любом из рассмотренных языков входят в первые две тысячи, более 80 – в первые три.

Если слово оказывается высокочастотным в одном из языков, велика вероятность, что оно окажется высокочастотным и в другом. Исключения связаны с полисемией: многозначные слова частотнее однозначных. Поскольку полисемия может быть унаследована из языка-предка, представительный универсальный список значений, частотность которых не зависела бы от полисемии, составить невозможно. Частотность до некоторой степени зависит от образа жизни, тем не менее, слова из сводешевского списка сохраняют высокую частотность при любом жизненном укладе. При этом изменения частотности отдельных слов не связаны ни с общим происхождением, ни с ареальным влиянием.

Наблюдения над частотностью позволяют предложить следующие улучшения методики сбора стословников: для предельных глаголов выбирать форму перфектива, для непредельных – имперфектива, для глаголов движения – форму однократного направленного действия, для глаголов восприятия – форму ненамеренного восприятия. Для слова ‘рука’ лучше брать не анатомические, а функциональные контексты.

Степень устойчивости слова с определённым значением различается в разных языках, тем не менее, разные выборки базисной лексики демонстрируют примерно одинаковую связь с частотностью. Хотя прямой зависимости между частотностью лексики и её устойчивостью нет, всё же соотношение между базисностью лексики и её частотностью является неслучайным. Существует связь между частотностью (но не стабильностью) и ранним освоением слов. Вероятно, это следствие того, что частотность в большей степени связана с нынешним образом жизни, тогда как стабильность в большей степени отражает прошлые эпохи.

Частотность слов с одним и тем же значением – разная в разных языках и даже в одном языке в разное время. И это, видимо, универсальный закон: если бы было иначе, слова из списка не могли бы выпадать (потому что их бы часто употребляли, не забывали и не заменяли). Сам же список базисных значений – каков бы он ни был – имеет вероятностную природу: какие именно значения войдут в первые тысячи по частотности в данном конкретном языке в заданный момент времени – непредсказуемо, но в каждый момент в любом языке бóльшая часть стословника в них входит.

Ключевые слова: базисная лексика; 100-словный список Сводеша; историческая устойчивость; частотность; усвоение языка.

Кажется логичным, что вхождение того или иного значения в список базисной лексики – т. е. лексики особо устойчивой, мало подверженной заменам и заимствованиям – основывается на том, что носителям языка оказывается проще вспомнить уже существующее слово, чем заменять его заимствованием или новообразованием. В силу устрой-

ства нашего мозга, то, что чаще встречается, вспоминается лучше, поскольку соответствующие нейронные контуры оказываются усилены (Hebb 1949).

Идея того, что более частотные слова должны лучше наследоваться, проводится в работе (Арапов, Херц 1974): М. В. Арапов и М. М. Херц предлагают не составлять списки базисной лексики, а использовать весь словарь языка, ранжировав слова на основе их частотности (Арапов, Херц 1974: 30–33). Список Сводеша авторы критикуют за то, что он, по их мнению, является не лингвистическим, а этнографическим (Арапов, Херц 1974: 26). Однако, как показал М. Н. Саенко (Саенко 2014: 50–53), слова сводешевского списка являются в большинстве своём достаточно частотными: в чешском языке (по словарю Jelínek et al. 1961) «64 слова вошли в первую тысячу, 78 в первые две тысячи самых частотных слов» (Саенко 2014: 51), в польском (по словарю Kurcz et al. 1990) «51 слово из стословника входит в первую тысячу, 76 в первые две тысячи наиболее частотных слов» (там же: 52), в русском (по словарю (Brown 2003), составленному на основе словаря (Частотный... 1977) объёмом 40 тысяч слов) «68 слов из стословника вошли в первую тысячу, 84 в первые две тысячи самых частотных слов русского языка» (Саенко 2014: 53).

Для проверки этой гипотезы за пределами славянской группы были проведены подсчёты по частотным словарям семи языков разных семей и ареалов: русского (Ляшевская, Шаров 2009), английского (Davies, Gardner 2010), турецкого (Aksan Y. et al. 2017), японского (Tono et al. 2013), арабского (Buckwalter, Parkinson 2011), грузинского (Шмальцель, Нозадзе 2012) и игбо (Anon. 2015–2017). Впоследствии в подсчёт были добавлены также данные частотных словарей китайского (Xiao et al. 2009), испанского (Davies, Hayward Dawies 2018), португальского (Davies, Preto-Bay 2008), французского (Lonsdale, Le Bras 2009), немецкого (Jones, Tschirner 2006), нидерландского (Tiberius, Schoonheim 2014), чешского (Čermák, Křen 2011) и русского (Sharoff et al. 2013) языков, вышедшие в той же серии и составленные по той же методике, что и словари английского, турецкого, японского и арабского языков, а также данные по польскому языку (Kurcz et al. 1990, цит. по: Саенко 2014: 51–52).

Словари английского, нидерландского, немецкого, французского, испанского, португальского, чешского, польского, турецкого, японского, китайского и арабского охватывают первые 5 000 наиболее частотных слов соответствующего языка, словарь грузинского – первые 3525 слов, словарь игбо – первую тысячу слов.

В нидерландском словаре Tiberius, Schoonheim 2014 частоты приведены отдельно по разным жанрам и группам жанров, общего списка, упорядоченного по частотности, нет, поэтому подсчёты по нему отличаются некоторой долей условности.

Сопоставление данных русского языка по словарям Ляшевская, Шаров 2009 и Sharoff et al. 2013 показывает, что, хотя корпуса и разные (в Ляшевская, Шаров 2009 использован Национальный корпус русского языка (НКРЯ), у Sharoff et al. 2013 – тексты с более, чем 75 тысяч веб-страниц, охватывающие 150 миллионов словоупотреблений), корреляция достаточно высока: почти все слова попадают в ту же полутысячу по частотности. Однако по сравнению со словарём Brown 2003 картина в Ляшевская, Шаров 2009 выглядит несколько более пессимистичной: в первую тысячу по частотности попали не 68, а лишь 63 слова из стословника, в первые две – не 84, а 81, что говорит о некоторой зависимости рангов от выборки.

Сходная ситуация с чешским языком: согласно словарю Jelínek, Bečka, Těšitelová 1961, в первую тысячу попадают 64 слова, а согласно Čermák, Křen 2011 – лишь 57, но при этом в первые две тысячи в более позднем словаре вошли не 78, а 80 слов. Для большинства слов ранг различается не очень значительно. Самое большое расхождение – у слова *tamten* ‘тот’: в словаре Jelínek, Bečka, Těšitelová 1961 он имеет ранг 8785 (крайне нетипич-

ный для слова с таким значением – в большинстве языков оно входит в первую тысячу), а в словаре Čermák, Křen 2011 – 1730.

В одной из работ С. А. Старостина показано, что частотность корневой морфемы в языке в некоторый момент времени не зависит от характера выбираемого текста (Старостин 1989: 18). Со словами это очевидным образом не так. Частотные словари основываются на корпусах письменных текстов, тогда как передача языка в череде поколений осуществляется прежде всего за счёт устной речи. Таким образом, наиболее информативен был бы частотный словарь, составленный на корпусе бытовых разговоров, но таких словарей в настоящее время не существует. Тем не менее, даже имеющиеся материалы позволяют, как кажется, сделать некоторые интересные наблюдения.

Выяснилось, что между рангами слов с одинаковым значением в разных языках обнаруживается достаточно высокая корреляция: если слово входит в первую тысячу наиболее частотных слов в одном языке, велика вероятность, что слово с этим значением войдёт в первую тысячу и в другом языке (см. Табл. 1). Если же слово в одном из языков является достаточно редким (не входит в первые три тысячи по частотности), то велика вероятность, что так же будет и в других языках. Из этого есть ряд объяснимых исключений: если слово совмещает в себе несколько значений, его частотность будет выше, чем у слова, имеющего лишь одно из соответствующих значений. Так, в чешском языке слово *měsíc* совмещает значения ‘луна’ и ‘(календарный) месяц’, в японском слово 羽 *hane* – ‘перо’ и ‘крыло’, в грузинском პირი *p’iri* – это и ‘рот’, и ‘человек’ (ср. совр. рус. разг. *в одно рыло* ‘в одиночку’), в арабском صغير ṣaġīr – это и ‘маленький’, и ‘молодой’, в русском слово *язык* – это и ‘tongue’, и ‘language’, а *рука* – и ‘hand’, и ‘arm’.

Из-за полисемии будет различаться частотность в разных семьях, унаследовавших от своих праязыков разные совмещения значений. Универсальный список – такой, чтобы в нём было представительное количество базисной лексики, но при этом ни в каком языке не было бы влияющей на частотность полисемии, составить, видимо, нельзя.

Интересно, что разные списки базисной лексики демонстрируют примерно одинаковую соотнесённость с частотностью. В рассмотренных языках среди 38 слов, входящих в список «Лейпциг – Джакарта» (Haspelmath, Tadmor 2009: 67), но не входящих в список Сводеша, от 11 до 22 входят в первую тысячу по частотности, среди слов, входящих только в список Сводеша, но отсутствующих в списке «Лейпциг – Джакарта», таких от 19 до 29 (грузинский язык из этого подсчёта исключён, поскольку данный словарь составлен по выборке текстов, из которых большинство составляют статьи законов, и выборка тем самым оказывается смещённой; так что лишь 35 слов из стословника входят в грузинском частотном словаре в первую тысячу по частотности). При этом во всех без исключения языках (даже в грузинском!) число слов, входящих в первую тысячу по частотности, в списке «Лейпциг – Джакарта» несколько меньше, чем в списке Сводеша. Напротив, слов низкочастотных (т. е. таких, которые не менее, чем в двух из рассмотренных языков не входят в первые три тысячи) в списке «Лейпциг – Джакарта» несколько больше, чем в списке Сводеша.

Если рассматривать не отдельные слова, а списки в целом, можно отметить, что есть немало значений (21, что составляет чуть больше одной пятой списка), которые во всех рассмотренных языках выражаются словами, входящими в первую тысячу по частотности. В предварительном варианте подсчёта, проведённом на меньшем количестве языков (были задействованы только русский, английский, турецкий, японский, арабский, грузинский и игбо), таких слов было 22. Это слова: ‘go’, ‘good’, ‘head’, ‘I’, ‘know’, ‘man’, ‘many’, ‘name’, ‘new’, ‘person’, ‘road’, ‘say’, ‘small’, ‘this’, ‘thou’, ‘two’, ‘water’, ‘we’, ‘what’, ‘who’, ‘woman’. Возможно, в этот же список можно включить слово ‘not’: из рассмотрен-

входящих в первую тысячу в языке: \ не входят в первую тысячу слов среди слов, в языке:	рус.	англ.	тур.	яп	араб.	груз.	кит.	игбо
рус. (63 слова)		11 (все входят во вторую тысячу)	4 (все входят во вторую тысячу)	10 (1 из них не входит во вторую тысячу)	10 (3 из них не входят во вторую тысячу)	27 (12 из них не входят во вторую тысячу)	20 (9 из них не входят во вторую тысячу)	12
англ. (57 слов)	5 (все входят во вторую тысячу)		1 (входит во вторую тысячу)	9 (1 из них не входит во вторую тысячу)	10 (3 из них не входят во вторую тысячу)	21 (6 из них не входят во вторую тысячу)	14 (5 из них не входят во вторую тысячу)	10
тур. (71 слово)	12 (4 из них не входят во вторую тысячу)	13 (1 из них не входит во вторую тысячу)		17 (4 из них не входят во вторую тысячу)	22 (8 из них не входят во вторую тысячу)	36 (18 из них не входят во вторую тысячу)	23 (8 из них не входят во вторую тысячу)	15
яп. (57 слов)	10 (2 из них не входят во вторую тысячу)	10 (1 из них не входит во вторую тысячу)	4 (1 из них не входит во вторую тысячу)		16 (5 из них не входят во вторую тысячу)	27 (13 из них не входят во вторую тысячу)	14 (6 из них не входят во вторую тысячу)	9
араб. (52 слова)	2 (1 из них не входит во вторую тысячу)	5 (1 из них не входит во вторую тысячу)	3 (1 из них не входит во вторую тысячу)	10 (1 из них не входит во вторую тысячу)		21 (7 из них не входят во вторую тысячу)	13 (7 из них не входят во вторую тысячу)	11
груз. (36 слов)	0	2 (1 из них не входит во вторую тысячу)	0	7 (2 из них не входят во вторую тысячу)	4 (2 из них не входят во вторую тысячу)		3 (2 из них не входят во вторую тысячу)	5
кит. (47 слов)	5 (1 из них не входит во вторую тысячу)	5 (все входят во вторую тысячу)	0	4 (1 из них не входит во вторую тысячу)	9 (4 из них не входят во вторую тысячу)	13 (7 из них не входят во вторую тысячу)		5
игбо (61 слово)	12 (3 из них не входят во вторую тысячу)	14 (3 из них не входят во вторую тысячу)	7 (3 из них не входят во вторую тысячу)	14 (3 из них не входят во вторую тысячу)	20 (6 из них не входят во вторую тысячу)	32 (15 из них не входят во вторую тысячу)	21 (9 из них не входят во вторую тысячу)	

Таблица 1. Сравнение частотностей слов с одинаковым значением в разных языках

ных языков оно отсутствует в турецком, поскольку в этом языке отрицание выражается аффиксом. В предварительном подсчёте было ещё слово ‘one’, которое оказалось за пределами первой тысячи в немецком языке (возможно, из-за того, что в составе числительных больше 20 оно имеет не такую форму, как в изолированном употреблении). Ещё для десяти слов (‘all’, ‘come’, ‘eye’, ‘full’, ‘give’, ‘hand’, ‘heart’, ‘long’, ‘one’, ‘see’) ранг больше тысячи показывает лишь один язык из рассмотренных.

В большинстве языков не менее 60 слов входит в первые две тысячи (среди рассмотренных языков отклоняются лишь нидерландский – 59 слов и грузинский – 56 слов), в первые три – не менее 70 слов (а чаще – более 80).

Для сопоставления частотности и устойчивости были взяты данные из работ, где списки базисной лексики ранжировались по степени стабильности: Старостин 2007 (с поправкой на то, что в настоящей работе не учитываются слова, добавленные С. Е. Яхонтовым; ни одно из них не входит по устойчивости в верхнюю полусотню), Holman et al. 2008, Поздняков 2014, Коровина 2019 для слов сводешевского списка, а также данные из работы Tadmor 2009: 67, подсчитанные для списка «Лейпциг – Джакарта», на 38 слов отличающегося от списка Сводеша.

Степень устойчивости слова в разных языках, как показывают подсчёты, проведённые на различных выборках и по различным методикам (см. Старостин 2007, Holman et al. 2008, Коровина 2019, Поздняков 2014), оказывается неодинаковой: так, слова ‘cloud’ или ‘tail’ оказываются чрезвычайно устойчивы в тюркских языках, но крайне неустойчивы в индоевропейских (Старостин 1989: 15), слово ‘name’ стабильно в языках Евразии, но не Африки (Поздняков 2014). Однако средняя частотность по списку является достаточно стабильной величиной: бóльшая половина списка обычно входит в первую тысячу по частотности, а за пределами третьей тысячи оказываются лишь 10–15 слов.

Есть слова, которые очень сохранны, но низкочастотны, например, ‘louse’ (ни в одном из рассмотренных языков слово с этим значением не входит в первые 5000 по частотности). Вероятно, это связано с изменением хозяйственного уклада: в современном мире педикулёз не особенно распространён, тогда как у охотников-собирателей дело обстояло иначе, некоторые народы вшей даже ели (Mowat 1958: 18).

Частотность может различаться в зависимости от ареала: так, в языках Севера слово ‘seed’ отсутствует или низкочастотно, в полинезийских языках отсутствует или низкочастотно слово ‘horn’.

Таким образом, М. В. Арапов и М. М. Херц в некотором смысле правы, характеризуя список Сводеша как скорее этнографический, чем лингвистический. Однако чисто лингвистический список (в том смысле, который предлагается в работе Арапов, Херц 1974) составить нереально: верхние позиции по частотности занимают предлоги, союзы, частицы, артикли и вспомогательные глаголы, которые бывает довольно непросто перевести с языка на язык (Саенко 2014: 49), кроме того, то, что в каком-то одном языке выражается предлогом или т. п., в другом может выражаться аффиксом (например, в турецком частотном словаре отсутствует перевод слова «не», поскольку в турецком языке отрицание выражается внутри глагола). К тому же, предлоги, союзы, частицы, артикли и вспомогательные глаголы, во-первых, возникают в результате грамматикализации (часто с утратой части фонемного состава и морфемной структуры), а во-вторых, подвергаются дальнейшей грамматикализации (с дальнейшей редукцией). Поэтому подход, основанный на списках устойчивой базисной лексики, представляется более продуктивным.

Редкими, но устойчивыми являются все названия нечеловеческих частей тела (‘feather’, ‘horn’, ‘tail’; ‘wing’) и насекомых (‘louse’; ‘ant’), а также ‘fingernail’, ‘dry’ и ‘liver’. Частотны и устойчивы местоимения и первые числительные, а из знаменательных слов –

‘name’ и ‘water’. Частотны и неустойчивы слова, содержащие оценку: ‘good’, ‘small’ и ‘many’. Редки и неустойчивы слова ‘bark’, ‘big’ и ‘belly’.

При этом среди слов, имеющих высокую частотность (входящих в первую тысячу) во всех рассмотренных языках, 6 входят в первую десятку наиболее сохранных слов (по ранжированию С. А. Старостина), 12 – в первую треть, 3 – в последнюю десятку, 8 – в последнюю треть, а середины почти нет. Зато слова, которые входят в первые две тысячи, представляют собой среднюю часть списка. По подсчётам Е. В. Коровиной, в первую десятку по устойчивости входит слово ‘fingernail’, которое во всех рассмотренных языках не вошло даже в первые 2,5 тысячи, а слово ‘many’, завершающее список устойчивости, во всех рассмотренных языках входит в первые полтысячи наиболее частотных слов. К. И. Поздняков отмечает, что «наиболее универсальны начало и конец списка. Наиболее стабильные значения универсального списка являются наиболее стабильными в большинстве языковых семей. Наименее стабильные значения являются таковыми в большинстве семей. Средняя часть “универсального списка” наименее универсальна» (Поздняков 2014: 206).

При любой ранжировке стословника – и по Старостину, и по Вихману, и по Позднякову, и по Коровиной, и по рангам списка «*Л*ейпциг – *Д*жакарта» – слов верхней половины по устойчивости, которые имеют низкую частотность (более, чем в одном языке выходят за пределы первых трёх тысяч), несколько меньше, чем низкочастотных слов из нижней половины.

Безусловно, частотность по современным словарям показывает срез лишь одного из бесчисленных этапов эволюции. В современных языках высокие позиции в частотном списке занимают слова, связанные с современными реалиями – ‘государство’, ‘система’, ‘проблема’, ‘закон’, ‘производство’, ‘политический’ и т. п. В языках, на которых говорили люди другого социально-экономического уклада, выше могли быть слова типа ‘сеять’, ‘пасти’, ‘доить’; когда люди перемещались в другие природные зоны, менее частотными становились слова типа ‘семя’ или ‘рог’; с победой над педикулёзом ушло в число редких слово ‘вошь’. Но интересно, что даже в такой ситуации значительная часть базисной лексики всё равно сохраняет высокие позиции. Таким образом, можно сделать вывод, что хотя прямой зависимости между частотностью лексики и её устойчивостью нет, всё же соотношение между базисностью лексики и её частотностью является неслучайным.

Наблюдения над частотностью разных лексем заставляют задуматься о том, что, возможно, стоило бы внести некоторые корректировки в методику отбора единиц сводешевского списка.

В соответствии с принципами, изложенными в работе Kassian et al. 2010, следует включать в список наиболее частотный и стилистический нейтральный вариант обозначения соответствующего понятия (Kassian et al. 2010: 48). В соответствии с этим для предельных глаголов, видимо, следует брать форму совершенного вида, а для непредельных глаголов – несовершенного вида: частотности соответствующих словарных единиц различаются очень сильно (см. Табл. 2 и Табл. 3).

Единственно исключение из этих закономерностей – глагол ‘bite’, у которого в русском языке частотности совершенного и несовершенного вида практически одинаковы: *кусать* – 8,5 ipm, *укусить* – 8,3 ipm. Возможной причиной этого является то, что характер соотношения между совершенным и несовершенным видом в данном случае – промежуточный между тем, что можно наблюдать в глаголах *есть*/*съесть* и *пить*/*выпить*, с одной стороны, и *жечь*/*сжечь* и *убивать*/*убить*, с другой.

Для глаголов перемещения наиболее частотной является форма однократного направленного перемещения (см. Табл. 4), для глаголов восприятия – форма ненамеренно-

го восприятия (см. Табл. 5 и Табл. 6). Интересно было бы сравнить частотность разных форм в других языках, имеющих систему аспектуальных противопоставлений.

Частично рекомендации такого типа отмечены в Kassian et al. 2010: так, для 'say' указано, что имеется в виду однократный речевой акт, для глаголов местоположения 'sit', 'stand' и 'lie', а также для глагола 'sleep' – что следует собирать формы, обозначающие именно состояние, а не переход в него. Для 'see' и 'hear' указано, что следует собирать именно формы ненамеренного восприятия – но лишь для 'see' сказано, что нужны формы именно несовершенного вида. Вероятно, сходную рекомендацию имеет смысл распространить и на глагол 'hear'.

Для глагола 'go' сказано, что нужна форма длительного действия (Kassian et al. 2010: 62), что не очень хорошо отличимо от 'ходить' (ср. *Он каждый день ходит на озеро*), для глагола 'swim' приводятся контексты как для 'плыть', так и для 'плавать', при глаголе 'fly' (Kassian et al. 2010: 60) отмечено, что если в языке различаются корни для 'лететь' и 'летать' (а также для 'идти' и 'ходить'), то допустимо включать в список оба корня на правах синонимов. Между тем, чем меньше в списке синонимов, тем точнее результат. Так что, возможно, стоит и на глаголы перемещения распространить принцип максимизации частотности и требовать для них форм направленного перемещения.

Предельные глаголы				
Единица стословника	Несов. вид	Частотность (ipm)	Сов. вид	Частотность (ipm)
burn (tr.)	жечь	14,1	сжечь	21,6
come	приходить	218,2	прийти	523,3
die	умирать	62,0	умереть	179,7
give	давать	370,7	дать	573,1
kill	убивать	49,8	убить	144,7
say	говорить	1755,0	сказать	2396,6
Непредельные глаголы				
Единица стословника	Несов. вид	Частотность (ipm)	Сов. вид	Частотность (ipm)
drink	пить	200,9	выпить	129,9
eat	есть	94,7	съесть	43,2
see	видеть	818,2	увидеть	452,4
hear	слышать	256,1	услышать	160,4
know	знать	1713,8	узнать	238,1
sleep	спать	221,9	заснуть уснуть	29,7 26,5

Таблица 2. Сопоставление частотностей совершенного и несовершенного вида (по Ляшевская, Шаров 2009).

Единица стословника	Состояние	Частотность (ipm)	Переход в состояние	Частотность (ipm)	Вторичный имперфектив	Частотность (ipm)
lie	лежать	318,1	лечь	63,5	ложиться	39,5
sit	сидеть	538,1	сесть	175,9	садиться	77,9
stand	стоять	419,3	встать	172,1	вставать	69,1

Таблица 3. Частотности глаголов местоположения (по Ляшевская, Шаров 2009).

Единица стословника	Направленное движение	Частотность (ipm)	Ненаправленное движение	Частотность (ipm)
fly (v.)	лететь	82,9	летать	46,6
go	идти	957,1	ходить	296,6
swim	плыть	42,5	плавать	33,1

Таблица 4. Сопоставление частотностей глаголов направленного и ненаправленного перемещения (по Ляшевская, Шаров 2009).

Единица стословника	Невольное восприятие	Частотность (ipm)	Целенаправленное восприятие	Частотность (ipm)
hear	слышать	256,1	слушать	239,5
see	видеть	818,2	смотреть	667,2

Таблица 5. Частотности глаголов восприятия в русском языке (по Ляшевская, Шаров 2009).

Единица стословника	Ранг	Глагол целенаправленного восприятия	Ранг
hear	198	listen	619
see	58	look	87

Таблица 6. Частотности глаголов восприятия в английском языке (ранги по Davies, Gardner 2010).

Интересно сопоставить данные о частотности и устойчивости слов с результатами, полученными в рамках исследования детской речи. Для оценки речевого развития ребёнка используется так называемый Макартуровский опросник (версия для русского языка – Елисеева и др. 2017). Там перечислены слова, которые нормально развивающийся ребёнок, скорее всего, будет знать к полутора и к трём годам. Разумеется, рамки нормального развития достаточно широки и для того, чтобы ребёнок не получил диагноза «задержка развития», надо знать не все слова, а лишь большую их часть (так, 50% девочек в полтора года понимают 263 слова, 50% мальчиков – 248 слов), но сами списки примечательны. В список, большую часть которого ребёнок должен освоить до трёх лет, входит 716 слов (включая звукоподражания и имена близких). В это число попадает более 2/3 стословного списка Сводеша – 71 слово; из них более 50 надо освоить до полутора лет. Бóльшая часть этих слов входит в первую тысячу по частотности (см. Табл. 7). Из тех слов, которые надо знать до полутора лет, в первую тысячу входят 40 (71%), во вторую – 11 (20%), в третью 4 (7%), и лишь одно (2%) не входит даже в третью. Из слов, которые надо выучить до трёх лет, 9 (60%) входят в первую тысячу, 3 (20%) во вторую, 2 (13%) в третью, одно (7%) является ещё менее частотным. Среди слов, которые надлежит выучить после трёх лет, процент низкочастотных выше: 9 из 29 этих слов (31%) не входят даже в третью тысячу по частотности.

Слово из списка Сводеша	Русский эквивалент (по Елисеева и др. 2017)	Возраст освоения	Ранг частотности в русском языке (по Ляшевская, Шаров 2009)	Ранг устойчивости по Е. В. Коровиной
not	не*	1,5	3	37
I	я	1,5	5	1
what	что	1,5	9	18

Таблица 7. Рано осваиваемые единицы сводешевского списка (в русском языке).

Таблица 7. Рано осваиваемые единицы сводешевского списка (в русском языке) (*продолжение*)

Слово из списка Сводеша	Русский эквивалент (по: Елисеева и др. 2017)	Возраст освоения	Ранг частотности в русском языке (по: Ляшевская, Шаров 2009)	Ранг устойчивости по Е. В. Коровиной
this	этот	1,5	14	42
we	мы	1,5	18	5
thou	тебе**	1,5	33	2
all	все	1,5	35	102
person	человек	1,5	39	79
who	кто	1,5	67	22
hand	рука	1,5	74	24
walk (go)	идти	1,5	95	95
big	большой	1,5	96	109
eye	глазки**	1,5	110	4
head	голова	1,5	132	70
give	дать	1,5	157	29
sit	сидеть	1,5	169	76
water	вода	1,5	191	23
good	хороший	1,5	199	99
foot	нога	1,5	205	49
stand	стоять	1,5	224	35
small	маленький	1,5	229	107
white	белый	1,5	290	81
black	чёрный	1,5	293	60
tongue	язык	1,5	306	7
lie	лежать	1,5	312	97
red	красный	1,5	442	87
sleep	спать	1,5	486	75
drink	пить	1,5	551	43
tree	дерево	1,5	659	46
sun	солнце	1,5	690	36
hair	волосы**	1,5	842	38
ear	ушки**	1,5	860	28
nose	нос	1,5	863	33
mouth	рот	1,5	901	88
dog	собака	1,5	909	11
stone	камень	1,5	910	30
green	зелёный	1,5	958	98
star	звезда	1,5	975	26
cold	холодный	1,5	1008	104
tooth	зубы**	1,5	1032	10
mountain	горка**	1,5	1040	100
warm	тёплый	1,5	1164	52

Таблица 7. Рано осваиваемые единицы сводешевского списка (в русском языке) (*продолжение*)

Слово из списка Сводеша	Русский эквивалент (по: Елисеева и др. 2017)	Возраст освоения	Ранг частотности в русском языке (по: Ляшевская, Шаров 2009)	Ранг устойчивости по Е. В. Коровиной
eat	есть	1,5	1290	17
bird	птичка**	1,5	1294	62
fish	рыба	1,5	1436	39
neck	шея	1,5	1449	91
rain	дождь	1,5	1490	44
yellow	жёлтый	1,5	1582	101
meat	мясо	1,5	1691	73
belly	живот	1,5	1831	106
sand	песок	1,5	2150	108
moon	луна	1,5	2525	34
egg	яйцо	1,5	2802	67
bite	кусать	1,5	9101	89
that	тот	3	36	74
say	сказать***	3	42	84
one	один	3	48	8
two	два	3	70	3
new	новый	3	73	31
see	видеть	3	113	90
night	ночь****	3	236	71
hear	слышать / слушать*****	1,5	411	51
full	полный	3	250	15
road	дорога	3	300	61
long	длинный	3	667	48
knee	колено	3	1178	69
dry	сухой	3	1493	56
fly (v.)	лететь	3	1495	68
cloud	облако	3	2643	94
swim	плыть / плавать *****	3	2668	54
claw(nail)	ногти*	3	4112	9

* Только в составе выражений *не хочу*, *не буду*.

** Слово либо не в начальной грамматической форме, либо с уменьшительным суффиксом.

*** Входит в пассивный словарный запас: уже от ребёнка до полутора лет требуется понимать слово *скажи*, хотя и не требуется уметь его произносить. В число слов, осваиваемых до трёх лет, входит *говорить*; *сказать* появляется ещё позже.

**** До полутора лет осваивается *ночью*; *ночь* как существительное осваивается в возрасте от полутора до трёх лет.

***** Авторы русской версии опросника считают данные глаголы за одну лексическую единицу: знание их (либо одного, либо другого) отмечается в одной общей клетке.

В этот список не включены позиции ‘man’ и ‘woman’, поскольку, хотя ребёнок уже в полтора года умеет называть отдельными словами мужчин и женщин, в русском языке для этого используются слова «языка нянь» (так называется «языковая подсистема, которую речевой коллектив считает пригодной в основном для общения с маленькими детьми» Елисеева и др. 2017: 7) – *дядя* и *тётя*. Возможно, при сборе стословника имеет смысл специально оговорить на всякий случай, что слова из «языка нянь» собирать не следует (хотя на практике вероятность этого, скорее всего, невелика).

Если сравнить те значения, которые должен выучить русскоязычный ребёнок, с частотностью соответствующих слов в других языках, выясняется, что в любом из них около 60% слов, входящих в первую тысячу по частотности, обозначают те реалии, названия которых мы осваиваем в раннем возрасте. Естественно, в разных языках в число рано осваиваемых будут входить разные слова (например, в английском языке слова *come* и *name* осваиваются гораздо раньше, чем их русские эквиваленты *прийти* и *имя*), так что подсчёт является лишь приблизительным, но, как кажется, результаты не должны расходиться слишком сильно.

При этом корреляции между ранним освоением и устойчивостью нет: если взять ранжирование по Е. В. Коровиной, то из рано осваиваемых слов половина относится к верхней половине списка, половина – к нижней; другие варианты ранжирования по устойчивости дают сходные результаты. Вероятно, это обусловлено тем, что раннее освоение, как и частотность (по современным словарям) связаны с нынешним уровнем развития культуры и типа хозяйства, тогда как устойчивость в большей степени ориентирована на прошлое.

Макартуровский опросник охватывает раннюю часть чувствительного периода – т. е. того периода, когда ребёнок овладевает родным языком (целиком он продолжается в среднем примерно до шести лет). И вероятность слова перейти к следующим поколениям зависит от того, в каком возрасте человек с ним встречается: чем раньше человек выучил слово, тем вероятнее, что он его не забудет, не станет образовывать обозначение для соответствующего понятия от других корней, а при контактах – заменять на заимствование. Передача лексики происходит через контексты: слыша незнакомое слово в контексте понятной ситуации, ребёнок соотносит это слово с каким-то её элементом. В этой связи кажется оправданным несколько изменить рекомендуемые контексты для сбора лексемы ‘hand’ – с чисто анатомических на функциональные, типа «возьми в руку», «держит ложку/палку в руке», «дай руку» или т. п. Интересно, что для ноги приведены функциональные контексты (Kassian et al. 2010: 61), а для руки, хоть и дана отсылка к лексеме ‘foot’, контексты исключительно анатомические (Kassian et al. 2010: 63).

Изменения частотности слов в разных языках происходят независимо. Если бы это было не так, в парах близко родственных языков или языков, вступавших в интенсивные контакты, слова с одинаковым значением имели бы более сходную частотность. Но имеющиеся данные показывают, вероятность слова войти в первую тысячу в языке 1 при условии, что слово с этим же значением входит в первую тысячу в языке 2, зависит лишь от того, сколько слов в том и в другом языке входят в первую тысячу (см. Табл. 8–10).

Испанский и португальский языки близкородственны, тем не менее, для каждого из них найдётся по 7 слов, которые в одном языке входят в первую тысячу по частотности, а в другом нет. Но во французском языке, отстоящем от них несколько дальше, нет слов, входящих в первую тысячу по частотности, которые бы в испанском или португальском оказались менее частотными. Причина этого – в том, что во французском языке стословных слов, входящих в первую тысячу по частотности, меньше, чем в испанском или португальском.

не входят в первую тысячу слов среди слов, в языке: / входящих в первую тысячу в языке:	польский	чешский по: Jelínek et al. 1961	чешский по: Čermák, Křen 2011	немецкий
польский (51 слово)		1 (оно не входит во вторую тысячу)	3 (все входят во вторую тысячу)	6 (1 из них не входит во вторую тысячу)
чешский по: Jelínek et al. 1961 (64 слова)	14 (2 из них не входят во вторую тысячу)			15 (3 из них не входят во вторую тысячу)
чешский по: Čermák, Křen 2011 (57 слов)	9 (1 из них не входит во вторую тысячу)			7 (2 из них не входят во вторую тысячу)
немецкий (54 слова)	9 (2 из них не входят во вторую тысячу)	5 (2 из них не входят во вторую тысячу)	5 (1 из них не входит во вторую тысячу)	

Таблица 8.

не входят в первую тысячу слов среди слов, в языке: / входящих в первую тысячу в языке:	немецкий	английский	французский
немецкий (54 слова)		5 (1 из них не входит во вторую тысячу)	14 (6 из них не входят во вторую тысячу)
английский (57 слов)	8 (все входят во вторую тысячу)		15 (6 из них не входят во вторую тысячу)
французский (45 слов)	5 (1 из них не входит во вторую тысячу)	3 (1 из них не входит во вторую тысячу)	

Таблица 9.

не входят в первую тысячу слов среди слов, в языке: / входящих в первую тысячу в языке:	французский	испанский	португальский
французский (45 слов)		0	0
испанский (63 слова)	16 (7 из них не входят во вторую тысячу)		7 (1 из них не входит во вторую тысячу)
португальский (63 слова)	18 (7 из них не входят во вторую тысячу)	7 (2 из них не входят во вторую тысячу)	

Таблица 10.

В более новом частотном словаре чешского языка (Čermák, Křen 2011) стословных слов, входящих в первую тысячу по частотности, меньше, чем в прежнем (Jelínek et al. 1961) – соответственно, уменьшаются и расхождения в частотности с польским и немецким.

Итак, частотность слов с одним и тем же значением – разная в разных языках и даже в одном языке в разное время. И это, видимо, универсальный закон: если бы было иначе,

слова из списка не могли бы выпадать (потому что их бы часто употребляли, не забывали и не заменяли). Сам же список базисных значений – каков бы он ни был – имеет вероятностную природу: какие именно значения войдут в первые тысячи по частотности в данном конкретном языке в заданный момент времени – непредсказуемо, но в каждый момент в любом языке бóльшая часть стословника в них входит.

Литература

Арапов, М. В., М. М. Херц. 1974. *Математические методы в исторической лингвистике*. Москва: Наука.

Елисеева, М. Б., Е. А. Вершинина, В. Л. Рыскина. 2017. *Макартуровский опросник: русская версия. Оценка речевого и коммуникативного развития детей раннего возраста. Нормы развития. Образцы анализа. Комментарии. Изд. 2-е, испр. и доп*. Иваново: ЛИСТОС.

Коровина, Е. В. 2019. Ранжирование базисной лексики С. А. Старостина: материалы к улучшению. В: Н. Н. Казанский (ред.). *Индоевропейское языкознание и классическая филология – XXIII. Материалы чтений, посвященных памяти профессора Иосифа Моисеевича Тронского. Первый полутом*: 546–556. Санкт-Петербург: Наука.

Ляшевская, О. Н., С. А. Шаров. 2009. Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). Москва: Азбуковник.

НКРЯ — *Национальный корпус русского языка*. Онлайн-версия: https://ruscorpora.ru.

Поздняков, К. И. 2014. О пороге родства и индексе стабильности в базисной лексике при массовом сравнении: атлантические языки. *Вопросы языкового родства* 11: 187–225.

Саенко, М. Н. 2014. *Общие инновации в базисной лексике как аргумент в дискуссии о балто-славянском единстве. Дисс.* на соискание ученой степени канд. филол. наук. Москва: Институт языкознания РАН.

Старостин, С. А. 1989. Сравнительно-историческое языкознание и лексикостатистика. В: И. Ф. Вардуль (ред.). *Лингвистическая реконструкция и древнейшая история Востока (тезисы докладов). Ч. 1*: 3–39. Москва: Наука.

Старостин, С. А. 2007. Определение устойчивости базисной лексики. В.: С. А. Старостин. *Труды по языкознанию*: 580–590. Москва: Языки славянских культур.

Засорина, Л. Н. (ред.). 1977. *Частотный словарь русского языка*. Москва: Русский язык.

Шмальцель Г., Г. Нозадзе. 2012. *Грузинско-азербайджанско-армянско-русский частотный словарь*. Тбилиси: Чешское землячество в Грузии «Злата Прага».

References

Aksan, Yeşim, Mustafa Aksan, Ümit Mersinli, Umut Ufuk Demirhan. 2017. *A Frequency Dictionary of Turkish. Core Vocabulary for Learners*. London: Routledge.

Anon. *SketchEngine: Igbo corpus (igWaC)*. 2015–2017. Available online at: www.sketchengine.eu/igtenten-igbo-corpus.

Arapov, Mikhail V., Maisa M. Kherts. 1974. *Matematicheskije metody v istoricheskoj lingvistike*. Moskva: Nauka.

Brown, Nicholas J. 2003. *Russian learners' dictionary*. London: Routledge.

Buckwalter, Tim, Dilworth Parkinson. 2011. *A Frequency Dictionary of Arabic. Core vocabulary for learners*. London: Routledge.

Čermák, František, Michal Křen. 2011. *A frequency dictionary of Czech: Core vocabulary for learners*. London: Routledge.

Davies, Mark, Dee Gardner. 2010. *A Frequency Dictionary of Contemporary American English. Word sketches, collocates, and thematic lists*. London: Routledge.

Davies, Mark, Kathy Hayward Davies. 2018. *A frequency dictionary of Spanish: Core vocabulary for learners. 2nd edition*. London: Routledge.

Davies, Mark, Ana M. R. Preto-Bay. 2008. *A frequency dictionary of Portuguese: Core vocabulary for learners*. London: Routledge.

Hebb, Donald O. 1949. *The organization of behavior: a neuropsychological theory*. New York: John Wiley and Sons.

Holman, Eric W., Søren Wichmann, Cecil H. Brown, Viveka Velupillai, André Müller, Dik Bakker. 2008. Explorations in automated language classification. *Folia Linguistica* 42 (3-4): 331–354.

Jelínek, Jaroslav, Josef V. Bečka, Marie Těšitelová. 1961. *Frekvence slov, slovních druhů a tvarů v českém jazyce*. Praha: Státní Pedagogické Nakladatelství.

Jeliseeva, Marina B., Elena A. Vershinina, Viktorija L. Ryskina. 2017. *Makarturovskij oprosnik: russkaja versija*. Ivanovo: LISTOS.

Jones, Randall L., Erwin Tschirner. 2006. *A frequency dictionary of German: Core vocabulary for learners*. London: Routledge.

Kassian, Alexei, George Starostin, Anna Dybo, Vasiliy Chernov. 2010. The Swadesh wordlist. An attempt at semantic specification *Journal of Language Relationship* 4: 46–89.

Korovina, Evgenija V. 2019. Ranzhirovanije bazisnoj leksiki S. A. Starostina: materialy k uluchsheniju. In: Nikolaj N. Kazanskij N. (ed.). *Indojevropejskoje jazykoznanije i klassicheskaja filologija XXIII*: 546–556. Sankt-Petersburg: Nauka.

Kurcz, Ida, Andrzej Lewicki, Jadwiga Sambor, Krzysztof Szafran, Jerzy Woronczak. 1990. *Słownik frekwencyjny polszczyzny współczesnej*. *T. 1–2*. Kraków: Instytut Języka Polskiego.

Lonsdale, Deryle, Yvon Le Bras. 2009. *A frequency dictionary of French: Core vocabulary for learners*. London: Routledge.

Lyashevskaja, Ol'ga N., Sergey A. Sharov. 2009. *Chastotnyj slovar' sovremennogo russkogo jazyka (na materialax Nacional'nogo korpusa russkogo jazyka)*. Moskva: Azbukovnik.

Mowat, Farley. 1958. *Coppermine Journey: An Account of a Great Adventure – Selected from the Journals of Samuel Hearne*. Boston: Litte, Brown & Co.

Pozdniakov, Konstantin I. 2014. O poroge rodstva i indekse stabil'nosti v bazisnoj leksike pri massovom sravnenii: atlanticheskije jazyki. *Journal of Language Relationship* 11: 187–225.

Saenko, Mikhail N. 2014. *Obshchije innovacii v bazisnoj leksike kak argument v diskussii o balto-slav'anskom jedinstve*. Dissertation Ms. Moskva: Institut jazykoznaija RAN.

Shmal'cel', Garol'd, Givi Nozadze. 2012. *Gruzinsko-azerbajdzhansko-arm'ansko-russkij chastotnyj slovar'*. Tbilisi: European Centre for Minority Issues.

Sharoff, Serge, Elena Umanskaya, James Wilson. 2013. *A frequency dictionary of Russian: Core vocabulary for learners*. London: Routledge.

Starostin, Sergei A. 1989. Sravnitel'no-istoricheskoje jazykoznanije i leksikostatistika. In: I. F. Vardul' (ed.). *Lingvisticheskaja rekonstrukcija i drevnejshaja istorija Vostoka. Part 1*: 3–39. Moskva: Nauka.

Starostin, Sergei A. 2007. Opredelenije ustojchivosti bazisnoj leksiki. In: Serget A. Starostin. *Trudy po jazykoznaniju*: 580–590. Moskva: Jazyki slav'anskix kul'tur.

Tadmor, Uri. 2009. Chapter III. Loanwords in the World's Languages: Findings and results. In: Haspelmath, Martin, Uri Tadmor (eds.). *Loanwords in the World's Languages. A comparative handbook*: 55–75. Berlin: Mouton de Gruyter.

Tiberius, Carole, Tanneke Schoonheim. 2014. *A frequency dictionary of Dutch: Core vocabulary for learners. 2nd edition*. London: Routledge.

Tono, Yukio, Makoto Yamazaki, Kikuo Maekawa. 2013. *A Frequency Dictionary of Japanese. Core Vocabulary for Learners*. London: Routledge.

Xiao, Richard, Paul Rayson, Tony McEnery. 2009. *A frequency dictionary of Mandarin Chinese: Core vocabulary for learners*. London: Routledge.

Zasorina, Lidija N. (ed.). 1977. *Chastotnyj slovar' russkogo jazyka*. Moskva: Russkij jazyk.

Svetlana Burlak. Stability and frequency: is there a correlation?

Words belonging to the Swadesh list of basic vocabulary are not only stable but also frequent. For each word, frequency varies across different languages and periods of time, but in general, more than half of the Swadesh items belong to the first thousand of the most frequent words, and ca. 80% of Swadesh items belong to the first three thousand, irrespective of language. It seems a priori probable that a meaning would belong to the basic vocabulary if it were easier for the speakers to remember an already existing word than to replace it with

a new derivative or borrowing. Given the construction of the human brain, things that are more frequent tend to be more strongly memorized (since the corresponding neuron circuits are strengthened, according to Hebb's postulate). For a preliminary testing of this hypothesis, we analyzed the Swadesh wordlist items in frequency dictionaries of 15 languages belonging to different families and linguistic areas. In all of these languages, 21 meanings were expressed by words belonging to the first thousand of the most frequent words. In most languages more than 60 words belong to the first two thousand and, typically, more than 80 words belong to the first three thousand of the most frequent words.

There is a rather strong correlation between frequency ranks of words with the same meaning in different languages. There are a few easily explicable counterexamples to this rule: e.g., if a word has several meanings, it will have a higher frequency. Languages of different families inherit different polysemy models from their protolanguages, which is why a universal wordlist containing a representative amount of words but not influenced by polysemy cannot be generated. Although frequency may vary depending on the speakers' way of life, a considerable amount of items on the Swadesh wordlist retain high frequency under any circumstances. Our data show that even if languages are closely related or involved in intensive contact, frequency changes within them are independent.

Observations on frequency ratings of different lexemes make it possible to provide some refinements to the method of compiling Swadesh wordlists. Thus, for terminative verbs, a perfective form should be chosen, while for non-terminative verbs, an imperfective form would be more appropriate. For verbs of movement, forms denoting a single directed movement should be chosen. For verbs of physical perception, forms denoting unintentional perception should be preferred. For the word 'hand', one should rely on functional rather than anatomical contexts.

Swadesh wordlist items are learned very early: thus, in Russian, children are expected to know ca. 70 Swadesh items before they are 3 years old. Words are acquired through their contexts, which is why the method of compiling Swadesh wordlist via diagnostic contexts is the most appropriate. The degree of stability of a word with a certain meaning differs in different languages; however, different samples of basic vocabulary show almost the same correlation with frequency. Though there is no direct dependence between frequency and stability of words, the relation between belonging to the subset of basic vocabulary and the subset of frequently used words is not random.

There is a correlation between early acquisition and frequency (but not stability). This is probably due to the fact that early acquisition and modern day frequency are more deeply connected with modern day lifestyle, while stability reflects previous epochs. Frequency of the words with the same meaning differs in different languages (and even in the same language in different times). This is, apparently, a universal principle: if it were not so, words would never have an opportunity to be eliminated from the Swadesh wordlist, since they would be so frequent that nobody could forget them and replace them with derivatives, borrowings, or other words. However, the list itself has a probabilistic nature: it cannot be predicted which meanings would gain higher frequency in a certain language at a certain period of time, but the majority of items on the Swadesh wordlist still belongs to it in any language at any period of time.

Keywords: basic vocabulary; Swadesh wordlist; lexical stability; lexical frequency; language acquisition.