Анализ методов оценки сложности текста

Мизернов И.Ю., Гращенко Л.А.
Академия ФСО России
Mizerov2015@mail.ru, graschenko@mail.ru

Аннотация. В статье приводятся предварительные результаты ретроспективного анализа основных методов оценки сложности текстов на естественных языках. Приводится описание экспериментального программного стенда для исследования применимости существующих моделей и методов к оценке сложности текстов информационно-аналитических документов. Показана низкая интерпретируемость существующих индикаторов сложности, представленных индексами удобочитаемости. В качестве перспективного направления исследований предлагается разработка моделей сложности в рамках алгоритмического подхода.

Ключевые слова: сложность текста, индекс удобочитаемости, модель сложности, автоматическая обработка текста.

1 Введение

Совершенствование систем документооборота (в том числе электронного) в различных организациях подразумевает, с одной стороны, оперативное прохождение служебных документов по инстанциям, а с другой – унификацию документов для повышения качества управленческих решений. Это, в свою очередь, предполагает эффективный менеджмент по синхронизации документарных потоков среди исполнителей и руководителей, а значит – автоматизированное прогнозирование временных и трудозатрат на подготовку и отработку документов. Особенно это касается отчетных и информационно-аналитических документов, объемы которых в крупных организациях могут быть значительными, вследствие чего качество их подготовки оказывает существенное влияние на время ознакомления с ними и принятия решений руководителями.

Известно, что неоправданно длинные предложения или сложные лексические конструкции затрудняют восприятие текста документа, поэтому требования делового и научного стиля содержат положения о краткости и лаконичности текста. Однако для объективной количественной оценки сложности текста документа, позволяющей прогнозировать эффект от подготавливаемой в организации документации, необходим обоснованный выбор и программная реализация соответствующего метода, адекватного языку и стилистике документооборота.

Поэтому в настоящей статье предпринята попытка в первом приближении охарактеризовать существующие подходы и способы к

оценке сложности текста с позиций автоматизации, а также исходя из практических потребностей различных организаций. Также представлен предварительный анализ результатов оценки сложности различных информационно-аналитических документов с помощью разработанного программного испытательного стенда.

2 Подходы к определению сложности текста

Под сложностью, в общем случае, понимают составленность объекта из нескольких частей; многообразность по составу входящих элементов и связей между ними. Синонимами данного понятия являются трудность, запутанность. Примечательно, что само слово «текст» (от лат. - *textus*) означает ткань, сплетение, соединение. Соответственно, состоящий из множества элементов, объединенных различного рода связями, текст описывается такой характеристикой, как сложность. Сложность произвольного объекта может быть описана в рамках одного из пяти подходов к определению сложности. Применимы они и к тексту, рис. 1.

Рис. 1. Подходы к оценке сложности текста, как объекта

В рамках каждого подхода рассматриваются теоретические аспекты сложности и математический аппарат для получения численных оценок сложности. В рамках *системного подхода* сложность системы тем выше, чем больше число и разнообразие ее элементов и связей между ними, вследствие чего сложная система приобретает множество новых свойств, не присущих отдельным подсистемам. *Теория хаоса* оперирует понятием нелинейной динамической системы, обладающей нестационарностью, т.е. случайными скачкообразными изменениями характеристик во времени. С позиций теории алгоритмов, *вычислительная сложность* есть функция объёма операций, выполняемых некоторым алгоритмом, от размера входных данных. Идея *колмогоровской сложности* заключается в том, что чем сложнее объект, тем длиннее его математическое (формальное) описание. Здесь мерой сложности является длина минимально возможного описания объекта [Hale, 2002]. *Организационный* (кибернетический) подход оперирует понятием управляющей подсистемы, способностью системы к активной адаптации. Чем труднее идентифицировать поведение системы и процессы управления в ней, тем сложнее система.

В XX веке при рассмотрении понятия сложности текста обнаруживаются такие синонимичные понятия, как удобочитаемость, читабельность (readability) [Микк, 1974], трудность текста и благозвучие [Иванов, 2013]. Они отражают, насколько удобным для зрительного либо слухового восприятия является текст, а факторами выступают размер букв, цвет шрифта и фона, наличие жаргонизмов и неологизмов и т.п. Однако такой подход к сложности текста не позволяет оценить содержание текста, его структуру и характеристики. Кроме того, исследования сложности текста во многом велись психологами, которые учитывали личностные характеристики понимающего субъекта при оценке природы текста и характера его понимания.

С началом XXI века при описании сложности текста стал превалировать подход, оперирующий общей идеей понимания сложности как количества затрачиваемых ресурсов для описания какого-либо объекта. Сложность языковой системы (complexity) противопоставляется стоимости и трудности (cost and difficulty) [Dahl, 2008]. В терминах работы [Miestamo, 2008] сложности текста соответствует абсолютная (объективная) сложность, зависящая от выбранного подхода к ее оценке, а стоимости и трудности – относительная (субъективная) сложность. Впрочем, выделение субъективных и объективных факторов сложности текста довольно условно. Например, информативность и абстрактность текста могут быть оценены как относительно, так и безотносительно к субъекту, рис. 2.

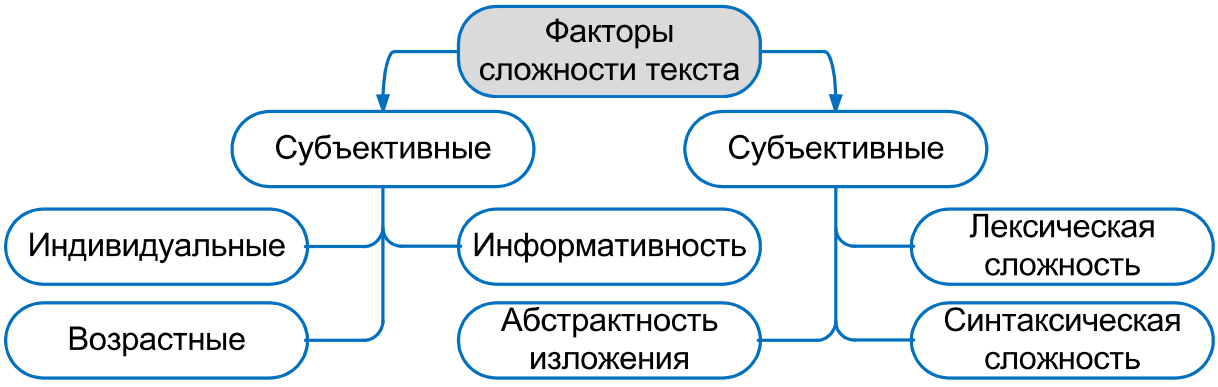

Рис. 2. Факторы сложности текста

Тем не менее, объективно текст можно рассматривать, по меньшей мере, на трех уровнях, для каждого из которых задаются различные числовые показатели, являющиеся основой практической оценки сложности, табл. 1.

Таблица 1. Статистические параметры текста

Уровень рассмотрения текста	**Параметр**	**Сокращение**
Макроуровень (уровень текста, абзаца)	Длина текста в абзацах	ДТА
	Длина текста в словах	ДТС
	Длина текста в буквах	ДТБ
	Средняя длина абзаца (в словах, буквах)	СДАС
Синтаксический	Средняя длина предложения во фразах	СДПФ
	Средняя длина предложения в словах (слогах)	СДПС
	Средняя длина предложения в буквах	СДПБ
Лексический	Средняя длина слова в буквах	СДСБ
	Процент односложных слов	ПОС
	Процент сложных слов	ПСС
	Процент неповторяющихся слов	ПНС
	Средняя частота повторения слова	СЧПС
	Процент частей речи (существительных, глаголов, прилагательных)	ПЧР

3 Математические модели сложности текста

На начальном этапе изучением сложности занимались ученые-лингвисты, реализуя свои методы для оценки сложности учебной литературы школьных учебных пособий и учебников. С 20-х годов XX века предпринимались многочисленные попытки численной оценки сложности текста индексами удобочитаемости, сначала для английского языка, а затем для европейских языков. Первый индекс удобочитаемости был предложен в 1923 году американскими учеными-лингвистами Б. Лайвли (B.Lively) и С. Пресси (S. Pressey) [Оборнева, 2006].Дальнейшие исследования в этом направлении были предприняты в 1947-1958 гг., 1967-1976 гг., а также с начала 2000-х годов наблюдается всплеск интереса к рассматриваемому вопросу, рис. 3.

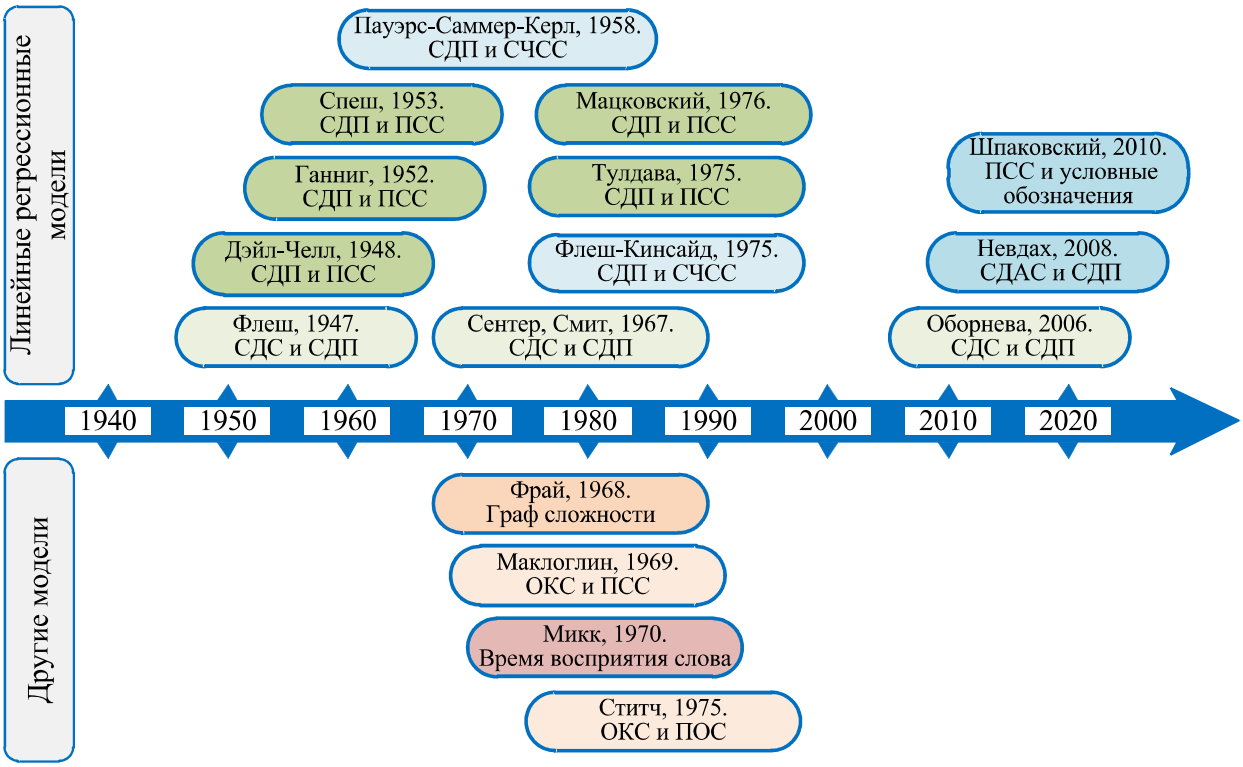

Рис. 3. Динамика исследований сложности текстов

В основу большинства предложенных формул легла линейная регрессионная модель (ЛРМ), переменными которой выступают статистические параметры текста:

$$f(x,b) = b_0 + \sum_{i=1}^{k} b_i x_i,$$

где: b_i - параметры (коэффициенты) регрессии; x_i - регрессоры (факторы сложности модели); k -количество факторов модели.

К достоинствам данной модели относят простоту, гибкость, а также единообразие анализа и проектирования, проводимого при помощи данных моделей. При использовании ЛРМ результат прогнозирования может быть получен быстрее, чем при использовании остальных моделей. Кроме того, достоинством является прозрачность моделирования, т.е. доступность для анализа всех промежуточных вычислений. Основным недостатком ЛРМ является сложность определения вида функциональной зависимости, а также трудоемкость определения параметров модели. К недостаткам также относят низкую адаптивность и отсутствие способности моделирования нелинейных процессов, а также высокую корреляцию результатов для разных методов оценки.

Как правило, индексы удобочитаемости на основе ЛРМ учитывали 2-3 параметра - по 1-2 с лексического и синтаксического уровней и применялись для оценки учебных текстов (для учащихся школ и колледжей). Коэффициенты регрессии подбирали таким образом, чтобы полученное значение показывало, каким уровнем образования должен

обладать либо в каком возрасте должен находиться читатель для успешного понимания предложенного текста.

В начале 70-х годов свои методы оценки сложности текстов стали предлагать психологи и социологи, которые рассматривали сложность с точки зрения возраста человека и его психологических особенностей и т.д. Предлагались альтернативные варианты формул расчета сложности, а также графы и таблицы. Так, в 1969 году Маклоглин предложил индекс «SMOG» читабельности текста, которым рассчитывал возраст читателя прозаического произведения через квадратный корень от доли многосложных слов в тексте [McLaughlin, 1969]. Также в 70-х года были предложены индексы удобочитаемости для русскоязычных текстов и обоснована методика их построения [Микк, 1974].

С начала 2000-х годов отмечается возвращение интереса к оценке удобочитаемости текстов, прежде всего, в России. Помимо адаптации известного индекса Флеша к русскому языку [Оборнева, 2006] предприняты перспективные попытки предложить индексы удобочитаемости для текстов «взрослой аудитории». Так, представляют интерес работы М.М. Невдаха и Ю.Ф. Шпаковского, в которых учитываемые факторы сложности переходят на макроуровень, так как у взрослого человека не возникает трудностей восприятия группы слов или целого предложения [Невдах, 2012].

На современном этапе изучения сложности текста предпринимаются попытки автоматизации процесса ее расчета, исследования проводятся на аудитории студентов вузов. В глобальной сети Internet появились онлайн сервисы, позволяющие оценить сложность вводимого пользователем текста, однако в них реализованы лишь отдельные формулы, зачастую не соответствующие уровню анализируемого текста.

Требуется автоматизировать расчет вышеуказанных методов и проверить пригодность их использования к текстам типовых информационно-аналитических документов на русском языке.

4 Текущие разработки

С целью проведения сравнительного анализа и применимости различных показателей сложности текстов с использованием среды Embarcadero® C++ Builder® начата разработка исследовательского программного стенда, рис. 4. После загрузки текста программа рассчитывает исходные статистические параметры (табл. 1) и рассчитывает 12 индексов удобочитаемости как для всего текста, так и по абзацам, что позволяет оценивать динамику изменения удобочитаемости внутри текста.

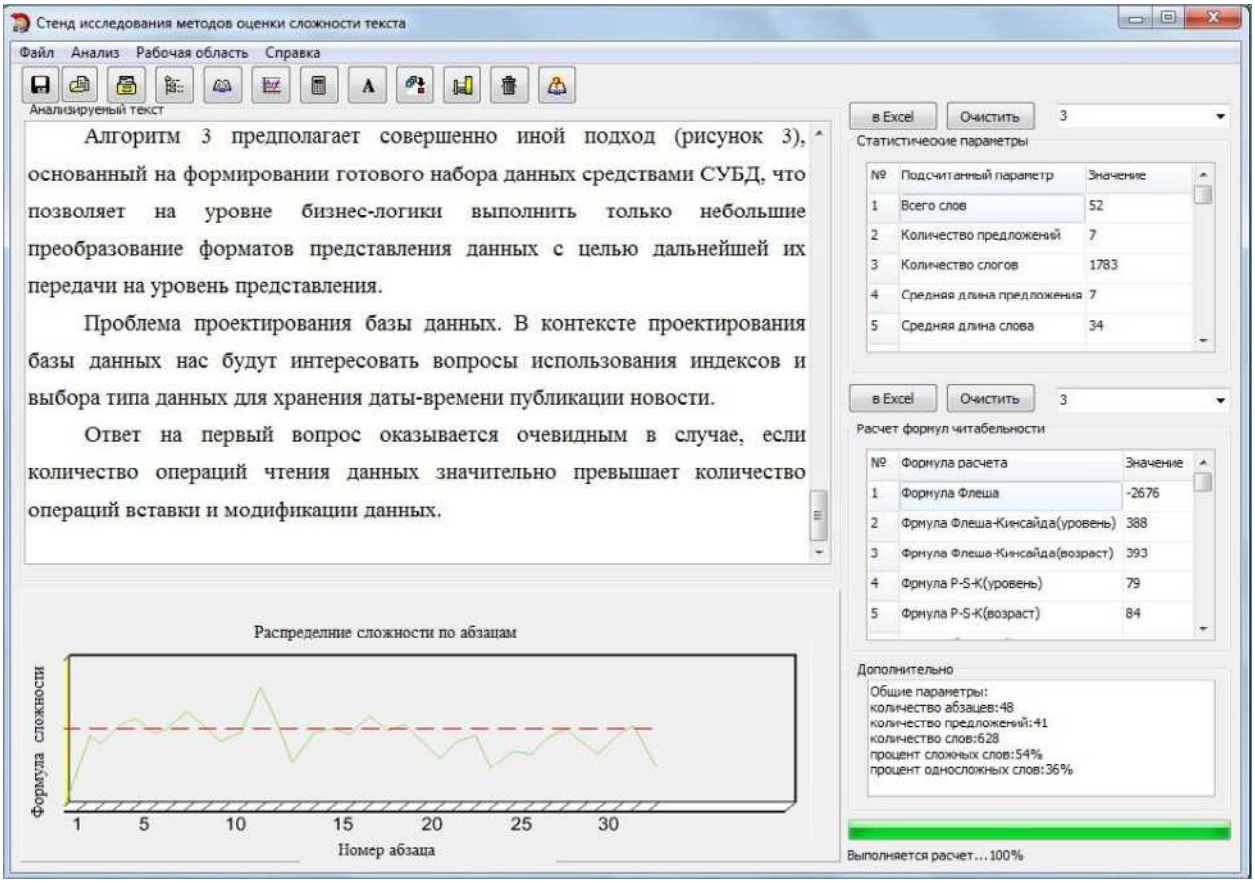

Рис.4. Главная форма экспериментального программного стенда

Индексы удобочитаемости были разбиты на пять групп по однородности статистических параметров, используемых для их расчета, рис. 3:

– Первая группа: тесты Флеша (1.1), Тулдавы (1.2), Колемана (1.3), АРИ (1.4) и Оборневой (1.5);

– Вторая группа: тест Флеша-Кинсайда (2.1 - возраст и 2.2 - уровень), тест Мацковского (2.3);

– Третья группа: тест Дейла-Челла (3.1) и тест Ганнига (3.2);

– Четвертая группа: тест Пауэрса (4.1);

– Пятая группа: SMOG-индекс (5.1).

5 Полученные результаты

Для проведения исследования методом информационного поиска на веб-сайтах государственных и коммерческих организаций была сформирована выборка информационно-аналитических документов по трем отраслям:

– маркетинговые и финансовые отчеты компаний;

– управленческие - отчеты о деятельности государственных учреждений;

– технические - итоги проведения научных и исследовательских работ.

Было проанализировано по 15 документов каждого направления, средний размер текста которых составил около 45000 печатных знаков.

С помощью разработанного программного стенда выполнен расчет различных индексов сложности, усредненные значения индексов по отраслям сведены в таблицу 2. Также приведен диапазон допустимых значений индекса и референсные значения (показывают, какой должна быть сложность текста документа, что его смысл мог успешно освоить взрослый человек либо студент, оканчивающий высшее образовательное учреждение).

Таблица 2 Усредненные значения индексов удобочитаемости для информационно-аналитических документов

Отрасль документа	Группы индикаторов											
	1					2			3		4	5
	1	2	3	4	5	1	2	3	1	2	1	1
Финансовые	-46	7	149	22	13	24	36	93	29	10	157	21
Технические	-22	13	162	32	23	41	55	120	36	17	174	29
Управленческие	-50	6	130	21	12	21	32	85	27	8	145	19
Диапазон значений	0-100	0-30	0-30	0-30	0-100	0-100	0-100	0-100	0-30	0-10	0-100	0-240
Референсные значения	0-30	20-30	20-30	20-30	0-30	23-50	23-50	0-30	20-30	9-10	23-50	70-240

6 Выводы и предложения

Исходя из полученных значений, видно, что большинство методов при оценке русскоязычных информационно-аналитических документов дают оценки, выходящие как за интерпретируемый диапазон значений, так и за референсные значения. Отмечается также, что тексты маркетинговых документов в среднем имеют наибольшую сложность, а организационные - наименьшую.

На основании проведенного анализа предоставляется возможным сделать следующие выводы:

– Полученные результаты характеризуются высокой степенью корреляции, в силу использования разработчиками одной математической модели (ЛРМ), а также однообразия применяемых в них параметров текста (средняя длина слова, средняя длина предложения).

– Реализация автоматической оценки сложности текста является достаточно сложной и языкозависимой задачей на этапе подсчета статистики по тексту (например, вычисление длин фрагментов, выраженных в слогах).

– Ни один из рассмотренных методов не направлен на оценку восприятия текста взрослым человеком, работающим со служебными документами. У профессионала не должно возникать затруднений с пониманием многосложных слов. В конечном итоге фактором сложности выступает семантика текста и абстрактность его изложения.

– Проверенные индикаторы удобочитаемости текста демонстрируют низкую интерпретируемость, поскольку не могут напрямую быть использованы для прогнозирования времени обработки текста тем или иным человеком.

– Целесообразно продолжить исследования в рамках сформулированных задач, в том числе реализовать расчет и проверку показателей сложности текстов на макроуровне представления текстов. Представляется необходимой дальнейшая теоретическая проработка понятия сложности текста.

Список литературы

[Иванов, 2013] Иванов, К.В. Автоматизация оценки благозвучия текстов / К.В. Иванов // Материалы шестнадцатого научно-практического семинара «Новые информационные технологии в автоматизированных системах» / под ред. С.Р. Тумковского. – М.: МИЭМ НИУ ВШЭ, 2013. – С. 253-254.

[Микк, 1974] Микк, Я.А. Методика разработки формул читабельности / Я.А. Микк. – М.: Советская педагогика и школа IX. – Изд-во Тарту, 1974. – 273 с.

[Оборнева, 2006] Оборнева, И.В. Автоматизированная оценка сложности учебных текстов на основе статистических параметров: дис... канд. пед. наук / И.В. Оборнева. - М., 2006. - 120 с.

[Невдах, 2012] Невдах, М.М. Повышение качества учебной литературы / М.А. Зильберглейт, Ю.Ф. Шпаковский, М.М. Невдах // Труды Белорусского государственного технологического университета. - 2012. - №9. - С. 89-92.

[Селезнев, 2007] Селезнев, Г.Д. Природа экспоненциального распределения слов по числу значений / Г.Д. Селезнев // Вестник Воронежского государственного университета. Серия: Лингвистика и межкультурная коммуникация. - 2007. - №2. - С. 42-45.

[Филиппова, 2010] Филиппова, А.В. Управление качеством учебных материалов на основе анализа трудности понимания учебных текстов: автореф. дис... канд. тех. наук: 05.13.10 / А.В. Филлипова; Уфа: ГОУ ВПО «Уфимский гос. авиационный техн. ун-т», 2010. – 24 с.

[Coleman, 1966] Coleman, E.B. Learning of prose written in four grammatical transformations. - 1966. - pp. 332-341.

[Dahl, 2008] Dahl, O. Grammatical resources and linguistic complexity. Siriono as a language without NP coordination / O. Dahl // Language complexity: typology, contact, change. – Amsterdam. - 2008.

[Dale, 1948] Dale, E.A formula for Predicting Readability / E. Dale, J. Chall // Educational Research Bulletin. – 1948. – 28 p.

[Flesh, 1946] Flesh, R. The Art of Plain Talk. - 1946. - 210 p.

[Hale, 2002] Hale, S. The Interaction between Text Difficulty and Translation Accuracy / S. Hale, S. Campbell // Babel. - 2002. - Vol. 48, №1.- pp. 14-33.

[McLaughlin, 1969] McLaughlin, H. SMOG grading – a new readability formula / H. McLaughlin // Journal of Reading. - 1969. - № 22. - pp. 639-646.

[Miestamo, 2008] Miestamo, M. Grammatical complexity in a cross-linguistic perspective / M. Miestamo, K. Sinnemaki, F. Karlsson // Language complexity: Typology, Contact, Change. - Amsterdam: John Benjamins, 2008. - pp. 23-42.

[Powers, 1993] Powers, R.D. A recalculation of 4 readability formulae / R.D. Powers, W.A. Summer, B.E. Kearl // Educational Psychology. - 1993. - № 49. - pp. 99-105.

[Stitch, 1973] Stitch, T.G. Research towards the design, development and evaluation of a job–functional literacy training program for the US Army / T.G. Stitch // Literacy Discussion. - 1973. - № 4. - pp. 339-369.

[Williamson, 2009] Williamson, G. Lexical Density.2009. [Электронный ресурс]. - Режим доступа: http://www.speech-therapy-information-and-resources.com/lexical-density.html