References

1 *On implementation of the Strategy of Industrial and Innovation Development of Kazakhstan for 2003–2015 by region.* [ER]. Access mode: http://www.mit.kz

2 Sklyarenko R.P. *Information economics: from theory to practice*. [ER]. Access mode: http://pipa.ru.

3 Muhanov D. *Kazakhstan: a breakthrough in the innovation economy*, Almaty, 2007.

4 Anfilatov V.S., Emelyanov A.A., Kukushkin A.A. *System analysis in management: Manual* / Ed. A.A.Emelyanov, Moscow: Finance and Statistics, 2002.

5 Kleandrov D.I., Frenkel A. *Statistical analysis of economic time series and forecasting: Proceedings of Statistics*, XXII–XXIII, Moscow: Nauka, 1973, p. 148–164.

6 Taubaev A.A. *Formation and development of the high technology sector in Kazakhstan*, Karaganda: Sanat-Printing, 2007, 170 p.

УДК 517.956.3

Л.В.Устинова, Л.С.Фазылова

Карагандинский государственный университет им. Е.А.Букетова
(E-mail:ustinovakrg@mail.ru)

Автоматизация оценки сложности учебных текстов на основе статистических параметров

В статье на основе статистических параметров текста исследуются вопросы количественной оценки сложности текста. На языке VBA создан макропакет, позволяющий определить стиль научной работы, уровень удобочитаемости текста курсовых и дипломных работ. В алгоритме программы были использованы индексы Флеша, Флеша-Кинкейда, индекс туманности Ганнинга. В работе приведены результаты тестирования разработанного макропакета.

Ключевые слова: автоматизация, оценка, сложность текста, научный стиль, статистические параметры, индекс Флеша, макропакет, алгоритм, программа, тестирование.

В XX веке появился ряд дисциплин прикладного характера на стыке лингвистики, математики и информатики. В частности, статистическая лингвистика — это дисциплина, изучающая количественные закономерности естественного языка, проявляющиеся в текстах. В ее основе лежит предположение, что некоторые численные характеристики и функциональные зависимости между ними, полученные для ограниченной совокупности текстов, характеризуют язык в целом или его функциональные стили (публицистический, художественный, научный и т.п.). Накопленные данные используются для решения задач теории связи, стенографии, информатики, а также выявления особенностей стиля отдельных авторов. На данный момент существует ряд исследований, в которых представлены математические модели оценки сложности текста. Однако эти модели получены в основном для английских текстов и не подкреплены соответствующими системами автоматизированного анализа. Между тем, необходимость подобных систем и соответствующих методик анализа текстов возникает у экспертов-методистов, создателей учебников, а также учителей, разрабатывающих различные методические материалы. С развитием системы экспертизы и сертификации учебной и методической литературы появилась необходимость в объективных и быстро реализуемых оценках ряда параметров сложности учебных текстов [1].

Задачей нашего исследования является изучение количественной оценки сложности текста. В качестве основных критериев используются статистические параметры текста, такие как длина слова, средняя длина предложения, процент многосложных слов и др. Названные параметры требуют достаточно сложных методов и технологий определения. Полученные на основе этих параметров различные формулы оценивают так называемую удобочитаемость или сложность текста. Эти параметры легко поддаются количественному выражению и могут быть использованы для автоматизации

оценки. Следует отметить, что формулы удобочитаемости не являются единственным критерием качества восприятия текста, они не оценивают тонкостей авторского стиля, но чётко отличают ясный простой текст от сложного.

Целью работы является разработка макропакета для автоматизации оценки сложности учебных текстов. Программа может применяться в учебном процессе для верификации курсовых и дипломных работ и определения соответствия публикаций стилю научной статьи. Автоматический классификатор функционального стиля текста создан на базе текстов, относящихся к четырём различным функциональным стилям. Критерием классификации является спектр длин слов.

В ходе создания макропакета были проанализированы существующие программы поиска и анализа текстовой информации: продукт Кирсанова компании «Гарант-Парк-Интернет», инструмент удобочитаемости, «Худломер», «Орфограммка».

Данные программы используются для решения следующих задач:

- анализ и классификация текстов, автоматическое реферирование;
- различные варианты поиска текста;
- морфологический, синтаксический и семантический анализ текста;
- средства навигации по большим массивам текстов.

Принцип работы макропакета основывается на следующих оценках:

- формула Флеша (Flesch readability formula);
- формула Флеша-Кинкейда (Flesch-Kincaid Grade Level);
- индекс туманности Ганнинга (Gunning Fog Index);
- график читабельности текста по Фраю (Fry Readability graph);
- оценка читабельности Рэйгора (Raygor Readability Estimate).

Тесты Флеша и Фога были разработаны для англоязычной аудитории и являются способом определения того, будет ли материал воспринят целевой аудиторией. Результаты любой проверки на удобочитаемость основаны на величине среднего числа слогов в слове и слов в предложении. Так как среднее число слогов в английском языке меньше, чем в русском, то необходимо модифицировать алгоритм с учетом специфики русскоязычных текстов. Данная оценка может быть получена в MSWord. Кроме проверки правописания и грамматики, на экран выводится информация о следующих показателях удобочитаемости документа:

- удобочитаемость по Флешу;
- уровень образования по Флешу-Кинкейду;
- число сложных фраз;
- благозвучие.

Недостатком данной проверки является влияние языков на определение уровня удобочитаемости. Если документ является полиязычным, то статистика удобочитаемости в Word выводится для последнего фрагмента с учетом языка, на котором он написан.

Для вывода статистики удобочитаемости с учетом нескольких языков нами разработан макрос на языке Visual Basic for Application (VBA). Такой выбор связан с популярностью текстового редактора MSWord. Макрос анализирует текст или любой его фрагмент на русском или английском языках. Отчет представлен в виде таблицы значений статистических характеристик текста и оценки его сложности, в соответствии с формулой Флеша (табл. 1).

Таблица 1

Результаты оценки сложности текста

Всего в тексте	Количество		
Слов	4099		
Символов	30865		
Абзацев	297		
Предложений в абзаце	453		
Среднее количество			
Предложений в абзаце	1.5		
Слов в предложении	8.0		

Продолжение таблицы 1

Символов в тексте	7.0		
Средняя длина слова	6.81		
Показатели легкости чтения		Уровень	Шкала оценки
Уровень образования (FKincadeE)	11.2	Уровень студентов	1–20
Легкость чтения (FlashRE)	42.55	Уровень студентов	0–100
Fog Index	13.23	Уровень студента 1 курса	0–20
Число сложных фраз	3.3		%
Благозвучие	88.8		0–100
Дисперсия	14.58		
Теоретическая дисперсия	15.84		
Стиль: КУРСОВАЯ РАБОТА — уровень студента 1 курса — **20,87 %**			

Результатом выполнения макроса является соответствие или несоответствие стиля уровню курсовой работы, с выводом дополнительных характеристик проверяемого текста (рис. 1).

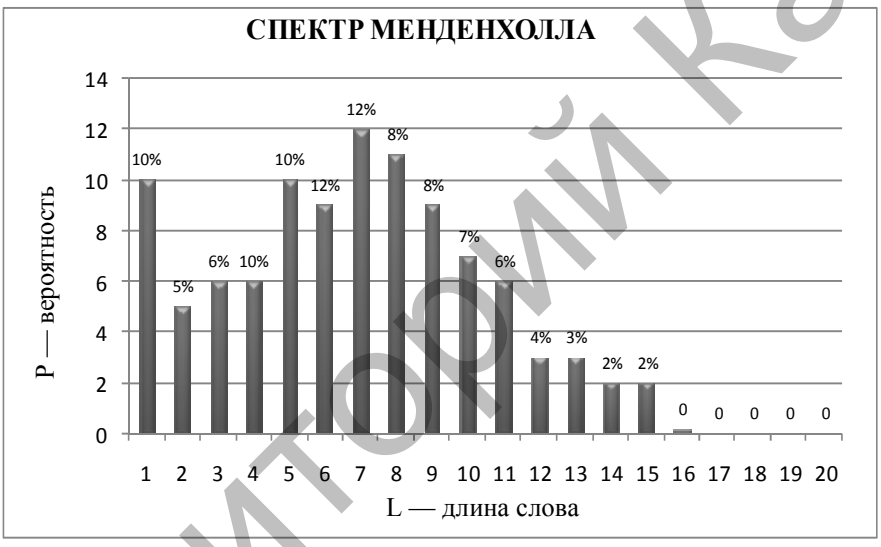

Рисунок 1. Дополнительные характеристики

Результаты статистики удобочитаемости сохраняются в отдельный файл с использованием коллекции readabilitystatistics объекта Document:

```
K_h = 0 :  'Количество характеристик текста
fn_othet = "отчет" + ActiveDocument.Name + ".doc"  'Файл отчета
For Each rs In ActiveDocument.ReadabilityStatistics
K_h = K_h + 1
StatText1(K_h) = rs.Name & " : " & rs.Value & vbCr :        'Характеристики текста
Next rs
Documents.Add Template:="Normal": Application.Keyboard (1087)
ActiveDocument.SaveAs FileName := fn_othet
Documents(fn).Activate
For ind = 1 To 12
Selection.TypeText Text:=StatText1(ind)    'Запись в файл отчета
Next
'Форматирование данных
Call format_res: '….
```

В основе всех указанных выше оценок лежит формула читаемости Флеша, которая позволяет оценить удобочитаемость текстовых материалов.

Проверка удобочитаемости по Флешу оценивается по 100-балльной шкале. Чем больше значение, тем понятнее текст. Для большинства текстов рекомендуемым значением является диапазон 60–70 баллов.

Формула Флеша является наиболее совершенной и распространённой для определения параметров удобочитаемости текста. Флеш определил основные характеристики текста, оказывающие влияние на его восприятие. Это среднее число слогов в слове и средняя длина предложения.

Формула расчета показателя удобочитаемости по Флешу:

$$Индекс\ \ Флеша_{(англ)} = 206.835 - 1.015 * ASL - 84.6 * ASW,$$

где ASL — среднее число слов в предложении (число слов, деленное на число предложений); ASW — среднее число слогов в слове (число слогов, деленное на число слов).

Свои выводы Флеш сделал на основе исследования текста «Экзаменационные уроки для чтения», которые традиционно используются в американской школе при переводе учеников из одного класса в другой. Данная методика получила название «формулы читабельности Флеша».

Формула Флеша, скорректированная для русского языка, прогнозирует лёгкость чтения текста [2]:

$$Индекс\ \ Флеша_{(рус)} = 206.835 - 1.3 * ASL - 60.1 * ASW.$$

Данная формула тесно связана с уровнем понимания текста учеником (табл. 2).

Таблица 2

Проверка удобочитаемости по Флешу

Уровень образования	Показатель
5 класс	91–100
6 класс	81–90
7 класс	71–80
8–9 классы	61–70
10 класс	51–60
Студент	31–50
Выпускник вуза	0–30

Аналогом индекса Флеша является индекс Флеша-Кинкейда. Уровень образования основан на индексе Флеша-Кинкейда. Он показывает, каким уровнем образования должен обладать читатель проверяемого документа. Школьный тест по Флешу-Кинкейду также используется для оценки текстов на экзаменах в школах США. Расчет показателя выполняется на основе вычисления среднего числа слогов в слове и слов в предложении. Значение показателя изменяется от 0 до 20. Значения от 0 до 10 соответствуют номеру класса школы, значения от 11 до 15 соответствуют курсам высшего учебного заведения. Диапазон значений от 16 до 20 относят к сложным научным текстам.

Формула школьного теста Флеша-Кинкейда:

$$Индекс\ \ Флеша - Кинкейда = 0.39 * ASL + 11.8 * ASW - 15.59.$$

Следует отметить недостаток обеих формул: линейная зависимость оценки от входных параметров. Хотя графики читабельности Рэйгора (рис. 2) [3] и Фрая (рис. 3) однозначно указывают на нелинейность построенной функции.

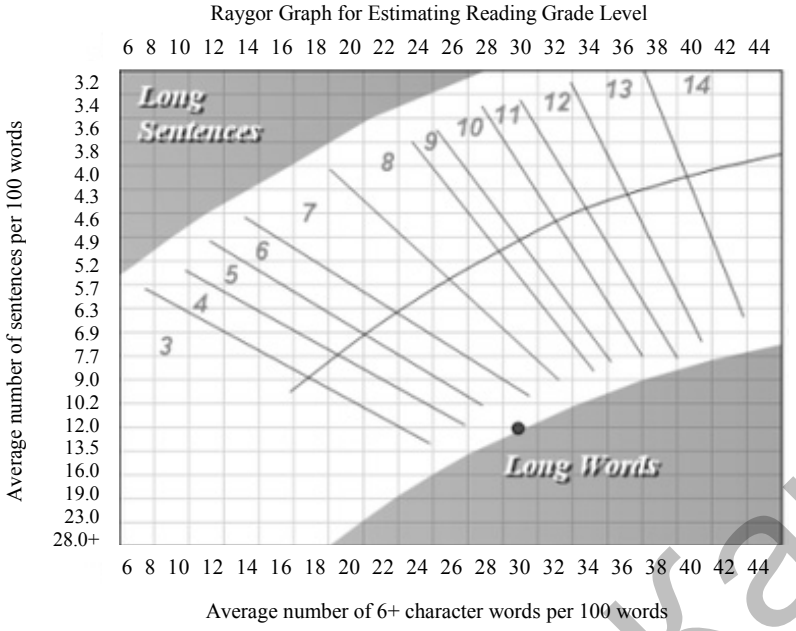

Рисунок 2. График читабельности текста по Рейгору

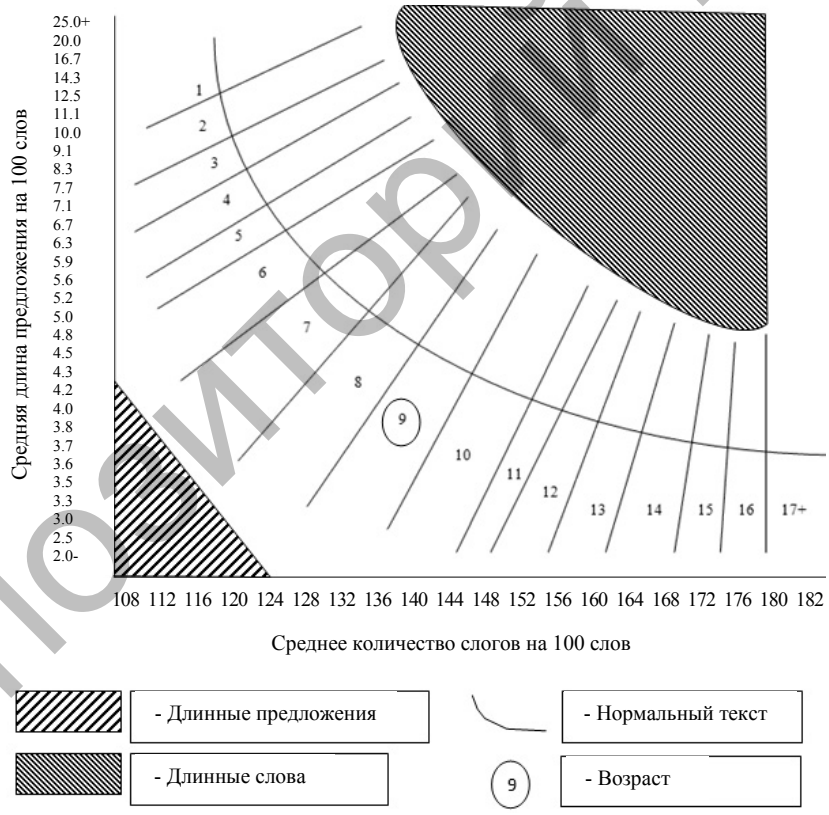

Рисунок 3. График читабельности текста по Фраю

Один из наиболее популярных методов оценки удобочитаемости текста — индекс Фога («индекс туманности»), разработанный в 1952 году американским ученым Р. Ганнингом [4]. Он используется в американской журналистике и позволяет определить минимальный возраст читателя, которому будет доступен данный текст. Индекс туманности определяет сложность чтения на основании средней длины предложения и процента слов, состоящих из трех и более слогов. Чем больше значение индекса

туманности, тем сложнее читать текст. Оценка выполняется для двух и более произвольных фрагментов текста, содержащих по 100 слов. Индекс туманности определяется по формуле

$$Индекс\ \ Ганнинга\ _{(англ)} = 0.4 * (w + l),$$

где w — средняя длина предложения; l — среднее количество «длинных» слов (из трех и более слогов).

Для текстов, написанных на русском языке, индекс Ганнинга вычисляется по формуле

$$Индекс\ \ Ганнинга\ _{(рус)} = 0.4 * \left[0.78 * \left(\frac{слов}{предложений} \right) + 100 * \left(\frac{число\ сложных\ слов}{число\ слов} \right) \right],$$

где *число сложных слов* — количество слов с числом слогов больше четырёх; 0.78 — поправочный коэффициент для русского языка.

Скорректированный индекс Ганнинга определяет образовательный уровень, необходимый для усвоения данного материала. Чем меньше значение индекса, тем большей аудитории текст будет понятен. Диапазон значений 16–20 соответствует уровню высшего образования, диапазон 9–10 — газетный уровень, 7–8 — школьный уровень.

Приведенные выше примеры демонстрируют разнообразие подходов к оценке уровня удобочитаемости текстов. В созданном макропакете для определения средней длины слова используются два критерия: количество слогов или количество символов в слове. Количество символов в слове автоматически определяется средствами VBA. Для определения количества слогов в слове был использован алгоритм Ляна-Кнута [5], который по словарю с расставленными переносами строит компактный набор правил, позволяющий в точности эти места переносов восстановить. Преимуществом использования первого критерия является высокая скорость выполнения. Хотя использование второго критерия перегружает память, он позволяет более точно определить оценку сложности текста. Для автоматизации проверки большого количества работ файлы (в формате *doc*) загружаются в одну папку, и запускается процесс проверки файлов:

```
if Right(whatdir, 1) <> Application.PathSeparator Then
whatdir = whatdir & Application.PathSeparator
End If
f1 = Dir(whatdir & "*.doc")
L = 0
Do While f1 <> ""
If f1 <> "отчет.doc" Then
Documents.Open FileName:=whatdir & f1
' Характеристики L текста
StatText1(L, 0) = whatdir & f1: 'Полное имя файла
' кол-во орфографических ошибок
StatText1(L, 15) = ActiveDocument.SpellingErrors.Count
' кол-во синтаксических ошибок
StatText1(L, 16) = ActiveDocument.GrammaticalErrors.Count
Document(whatdir & f1).Activate : Selection.WholeStory
 Application.Run MacroName:="ManStatist" :  'Анализ характеристик
L = L + 1 : 'L — счетчик анализируемых файлов
End If
f1 = Dir : '…..
Loop
  Call FormaStatALL: ’Форматирование и сохранение результатов
```

Результат проверки представлен в следующей таблице.

Таблица 3

Проверка удобочитаемости по Флешу

Имя документа	Индексы Флеша	Количество ошибок: орфограф., синтакс.	Уровень студента	Стиль
Kurs_3Dmod(Масюк).doc	34	17; 31	4 курс	Курсовая работа
Kurs_3Dmod(Шурыгина).doc	32	2; 5	4 курс	Курсовая работа
Dipl_(Гудз.В).doc	30	6; 10	Выпускник	Дипломная
Stat_Ust_Fazilova.doc	30	0; 0	Выпускник	Научный стиль
Kurs_алг(Nemo).doc	40.97	5; 4	2 курс	Курсовая работа
……	……	……	……	……
K_kompmod(Зимина).doc	32	3; 4	4 курс (89.05)	Курсовая работа
Дата проверки				16/11/2013

Разработанный макропакет требует дополнительного тестирования для обработки больших объемов текстовых материалов. В перспективе предлагаемый макропакет может применяться в учебном процессе для определения соответствия уровня курсовых, дипломных работ; для ускорения проверки на наличие и определение количества ошибок в тексте; оценки сложности школьных учебных текстов с учетом возрастных особенностей учащихся. Автоматизация анализа сложности учебных текстов с применением информационных технологий на основе методов их количественной оценки позволит увеличить эффективность обработки документов.

Список литературы

1 *Попова Я.И., Шишкевич Е.В.* Стандартизация учебной литературы средней школы по критерию удобочитаемости // Севастопольский национальный университет ядерной энергии и промышленности // Научные ведомости БелГУ. Сер. Гуманитарные науки. — 2010. — № 12. — Вып. 6. — С. 142–147.

2 *Оборнева И.В.* Автоматизация оценки качества восприятия текста // Вестн. Москов. городс. пед. ун-та. Сер. Информатика и информатизация образования. — 2005. — № 2 (5). — С. 86–92.

3 Raygor Readability Estimate. — [ER]. Access mode: http://en.wikipedia.org/wiki/Raygor_Estimate_Graph

4 *Рогушина Ю.В.* Использование критериев оценки удобочитаемости текста для поиска информации, соответствующей реальным потребностям пользователя // Проблеми програмування. — 2007. — № 3. — С. 76–87.

5 Алгоритм Ляна-Кнута. — [ER]. Access mode: http://quittance.ru/blog/index.php?category=21

Л.В.Устинова, Л.С.Фазылова

Статистикалық өлшемдер негізінде мәтін күрделілігінің бағасын автоматтандыру

Мақала мәтіннің статистикалық параметрлері негізінде мәтін күрделілігінің сандық бағасы сұрақтары зерттелді. VBA тілінде ғылыми жұмыстың стилін, курстық және дипломдық жұмыстары мәтіндерінің оқылу деңгейін анықтауға мүмкіндік беретін макропакет құрылған. Бағдарлама алгоритмінде Флеш, Флеш-Кинкейд, Ганнинг индекстері қолданылған. Сондай-ақ құрылған макропакетті тестілеу нәтижелері келтірілген.

L.V.Ustinova, L.S.Fazylova

Automation of estimation of complexity of educational texts on the basis of statistical parameters

In this work issues of quantitative estimation of the complexity of the text on the basis of statistical parameters were researched. It is created macropackage in the language VBA that allows to define the style of scientific work, the level of readability of projects and dissertations. Flesch index, Flesch-Kincaid index, Gunning index nebula were used in the algorithm of the program. This work contains the testing results of the developed macropackage as well.

References

1 Popova Y.I., Shishkevich E.V. *Scientific statements of BSU*, Humanities Ser., 2010, 12, 6, p. 142–147.

2 Oborneva I.V. *Bull. of the Moscow City Pedagogical University*, A series of «Science and the computerization of education», 2005, 2 (5), p. 86–92.

3 *Raygor Readability Estimate*, [ER]. Access mode: http://en.wikipedia.org/wiki/Raygor_Estimate_Graph

4 Rogushina Y.V. *Problems of programming*, 2007, 3, p. 76–87.

5 *Algorithm Liang-Knuth*, [ER]. Access mode: http://quittance.ru/blog/index.php?category=21

УДК 517.518.235

Г.Ш.Искакова

Карагандинский государственный университет им. Е.А.Букетова (E-mail: iskakova.1975@mail.ru)

Об одном многовесовом анизотропном неравенстве вложения

В статье получены теоремы вложения одного многовесового многопараметрического пространства Соболева для весов общего типа на областях с произвольной геометрией. Получены условия на весовые функции $\rho_i\,(i=1,...,n),\upsilon$ и ω, при которых справедливо неравенство вложения $\left(\int\limits_G \left|D^a f\right|^q \omega\right)^{1/q} \le C\left(\sum\limits_{i=1}^{n}\left(\int\left|D^l f\right|^{p_i}\rho_i\right)^{1/p_i}+\left(\int\limits_G |f|^{p_0}\upsilon\right)^{1/p_0}\right)$. Приведены примеры с доказательствами.

Ключевые слова: анизотропное, многовесовое, многопараметрическое, произвольная геометрия, пространство.

Пусть G — область в R^n, $l=(l_1,...,l_n)$; $\alpha=(\alpha_1,..,\alpha_n)$ — векторы с целыми координатами $l_i>0$, $\alpha_i\ge 0$. Нами будут использованы обозначения: для $x=(x_i)=(x_1,\dots,x_n)\in(-\infty,+\infty)^n$, $y=(y_i)\in(0,+\infty]^n$, $\lambda=(\lambda_i)\in(0,+\infty)^n$, $t\in(0,+\infty)$ пусть $x\le y$, $x<y$ — запись покоординатного сравнения, $\lambda x=(\lambda_i x_i),(\lambda,x)=\sum\limits_1^n\lambda_i x_i$, $\frac{x}{y}=x:y=\left(\frac{x_i}{y_i}\right)$, $\frac{1}{y}=\left(\frac{1}{y_i}\right)$, $|\lambda|=\sum\limits_1^n\lambda_i$, $t^\lambda=(t^{\lambda_i})$, $|x|_\lambda=\max\limits_{1\le i\le n}|x_i|^{1/\lambda_i}$, $1=(1)$, $\infty=(+\infty)$. Для $x\in R^n$, множеств E, $F\subset R^n$ и $\lambda\in(0,+\infty)^n$ пусть $x\pm\lambda E=\{y:\ y=x\pm\lambda z,\ z\in E\}$, $E\pm F=\{z:\ z=x\pm y,\ x\in E,\ y\in F\}$. Пусть $Q_0=(-1,1)^n$, область $G\subset R^n$, $G\left(\frac{1}{\lambda},t\right)=\left\{x:\ x=y+\left(\frac{t}{2}\right)^\lambda Q_0\right\}\subset G$, $G_t=\{x:\ x\in G, dist(x,\partial G)>t\}$.

Далее $l\in N^n,\alpha\in Z^n,\alpha\ge 0$.

$$Q=Q_d=Q_d(x)\overset{\text{def}}{=}\left\{y\in R^n:\ |y_i-x_i|<d/2,\ i=1,\dots,n\right\}=Q_{(2d,\lambda)}(x)$$

при $\lambda=(\lambda_1,...,\lambda_n)$, $\lambda_1=...=\lambda_n=1$. Положим

$$\tau(x)\overset{\text{def}}{=}\min\left(1,\ \sup_{d>0}\{d:\ 2Q_d(x)\subset G\}\right),\ Q(x)=\tfrac{1}{2}Q_{\tau(x)}(x)$$