Киев, 1974.

УДК 62.506.2

Е. А. СОЛОВЬЕВА

МОДЕЛИРОВАНИЕ ПРОЦЕССОВ МОРФОЛОГИЧЕСКОЙ КЛАССИФИКАЦИИ С УЧЕТОМ ОМОГРАФИИ

При создании модели классификации словоформ будем исходить из того, чтобы она 1) описывала некоторую функцию, т. е. одинаковым входным сигналам ставила в соответствие одинаковые выходные; 2) была адекватной изучаемому процессу, т. е. поведению грамотного человека при решении рассматриваемой задачи и 3) была действующей, т. е. воспроизводимой с помощью технических средств.

Постановку задач морфологической классификации можно сформулировать следующим образом. Если грамотному испытуемому предъявить упорядоченную пару $\langle x_i, x_j \rangle$ словоформ, где $\langle x_i, x_j \rangle \in X \times X$ (X — некоторое множество словоформ русского языка), то он сможет ответить «да» или «нет» («0» или «1») в зависимости от того, являются ли формы x_i и x_j эквивалентными по заданному признаку. При этом человек реализует некоторую функцию φ, определенную на множестве $X \times X$ и принимающую

значения «0» или «1». Необходимо построить модель функции φ, основываясь на информации в входных и выходных сигналах.

Известно [1, 2 и др.], что если функция удовлетворяет условиям

$$1)\ \varphi(x_i, x_i) = 1;$$

$$2)\ \text{если } \varphi(x_i, x_j) = 1, \text{ то } \varphi(x_j, x_i) = 1;$$

$$3)\ \text{если } \varphi(x_i, x_j) = 1, \varphi(x_j, x_k) = 1, \text{ то } \varphi(x_i, x_k) = 1,$$

то существует некоторое множество Y и функция $f: X \to Y$ такая, что

$$\varphi(x_i, x_j) = L(f(x_i), f(x_j)). \qquad (1)$$

L является характеристической функцией диагонали квадрата $Y \times Y$, т. е.

$$L(y_i, y_j) = \begin{cases} 1, \text{ если } y_i = y_j, \\ 0, \text{ если } y_i \neq y_j, \end{cases} \text{ причем } y_i = f(x_i),\ y_j = f(x_j).$$

Элементы множества Y при решении конкретной задачи интерпретируются как коды полученных классов эквивалентности. В один и тот же класс попадут те и только те элементы $x_i \in X$ и $x_j \in X$, для которых выполняется условие $\varphi(x_i, x_j) = 1$. Классы эквивалентности либо совпадают с грамматическими значениями рассматриваемой категории (значение понимаем в смысле множества [3]), если между ними отсутствует омография, либо определенным образом связаны с этими значениями [3]. Описание функции f является самостоятельной моделью морфологического анализа, построение которой считаем основной задачей наших исследований.

Для представления функции φ в виде (1) необходимо убедиться в выполнении свойств 1—3. Экспериментальная проверка этих свойств не представляет затруднений при изучении процессов морфологической классификации. Свойства 1—3 выполняются при правильной постановке задачи испытуемым даже при наличии омографии.

Словоформы эквивалентны по признаку некоторой категории, если каждой из них соответствует один и тот же набор признаков значений (признак формального значения) этой категории (при необходимости учитывается признак отсутствия значений) [3]. В ряде задач классификации (при наличии омографии между значениями рассматриваемой категории) классы эквивалентности, сформированные на основании ответов испытуемых, могут не совпадать с частными грамматическими значениями.

Рассмотрим задачу морфологической классификации синтетических личных форм глаголов русского языка (X — множество таких форм) по признакам собственно грамматических категорий: наклонения M, времени T, числа N, лица L и рода G, т. е. по признаку $Z = \langle M, T, N, L, G \rangle$. В процессе ее решения грамотный че-

ловек реализует функцию $f_Z^I : X \to Y_Z^I$ (назовем ее общим классификатором личных форм), где $Y_Z^I = \{z_1^I, \ldots, z_{17}^I\}$, причем $z_i^I = y_i\,(i = \overline{1,13})$ [3]: $z_{14}^I = \langle M_3^I, T_4^I, N_1^I, L_2^I, G_4\rangle$, $z_{15}^I = \langle M_3^I, T_4^I, N_1^I, L_5^I, G_4\rangle$, $z_{16}^I = \langle M_3^I, T_4^I, N_2^I, L_2^I, G_4\rangle$, $z_{17}^I = \langle M_3^I, T_5^I, N_3^I, L_6^I, G_4\rangle$.

$z_i (i = \overline{1,17})$ характеризует i-й класс эквивалентности в разбиении множества X по признаку Z. Такое разбиение на основании ответов испытуемых можно получить при классификации по признаку Z, а также в результате классификации по каждому из признаков в отдельности (M, T, N, L или G) и изучения логических возможностей сочетания полученных классов.

Признаки $z_{14}^I - z_{17}^I$ соответствуют глаголам, обладающим свойством омографии, z_{14}^I характеризует омографичные словоформы, оканчивающиеся на *ешь*, z_{15}^I — на *усь*, z_{16}^I — на *ите*, z_{17}^I — на *ли*.

Для построения точной модели функции f_Z^I (с учетом омографии) можно использовать результаты проведенных исследований процессов морфологической классификации омографичных форм [4] и полученные ранее частные модели [5 и др.] Алгоритмическая модель A_Z^I функции f_Z^I приведена на рис. 1. Блоки *1, 2* — Z^I и *3* — Z^I представляют собой блоки нормализации, классификации и выходной соответственно [6], остальные блоки уточняющие и введены для разбора случаев омографии глагольных форм. Все уточняющие блоки, разработанные нами в результате исследований процессов классификации полностью и частично совпадающих форм, подробно описаны в [4].

В зависимости от конкретных условий функционирования можно менять состав, число и порядок соединения уточняющих блоков (чтобы множество выходных сигналов модели не менялось), варьируя тем самым сложность и точность алгоритма. Чтобы построить различные модификации модели A_Z^I, сформулируем правила объединения блоков в алгоритм, верные и для частных моделей классификации по признакам M, T, N и L.

1. Уточняющие блоки и блок нормализации (1, 1′ или 1″) всегда располагаются перед основным блоком классификации, который всегда стоит перед выходным блоком (общая схема алгоритмов приведена в работе [6]).

2. Блоки 2-*усь* и 2-*ли* должны находиться перед блоком нормализации, а блоки 2-*ешь* и $H_{\text{ем}}^I$ — после блока 1 (или 1″).

3. Блоки типа T подключаются перед блоком 1 (или 1″).

Блок 1 является блоком постфиксной нормализации (отбрасывает постфикс *ся* или *сь* при наличии его в слове), 1′ — частичной при наличии в слове частицы *-ка* (отбрасывает ее), 1″ — полной (отбрасывает *-ка* и *ся* или *сь*).

2-Z^I (рис. 2) — блок классификации и алгоритма A_z, определяющего все собственно грамматические значения личных форм

без учета омографии. Оператор $И_j (j=\overline{1,3})$ отбрасывает j последних букв слова. Любой распознаватель Φ_i из граф-схемы алгоритма 2-Z^1 (см. рис. 2) проверяет конец слова на совпадение с одним из буквосочетаний (букв), входящих в упорядоченное множество $P\Phi_i$ формальных (внешних) грамматических признаков словоформ.

В результате проведенных исследований получено, что $P\Phi_1 = \{$у, ю, дам$\}$, $P\Phi_2 = \{$т$\}$, $P\Phi_3 = \{$е, и, с$\}$, $P\Phi_4 = \{$м$\}$, $P\Phi_5 = \{$ете, ите$\}$, $P\Phi_6 = \{$ли$\}$, $P\Phi_7 = \{$л, з, с, б, п, р, к, х$\}$, $P\Phi_8 = \{$ла$\}$, $P\Phi_9 = \{$ло$\}$, $P\Phi_{10} = \{$г$\}$, $P\Phi_{11} = \{$ляг$\}$, $P\Phi_{12} = \{$ь$\}$, $P\Phi_{13} = \{$е$\}$, $P\Phi_{14} = \{$еш, иш$\}$, $P\Phi_{15} = \{$ест$\}$, $P\Phi_{16} = \{$сут$\}$, $P\Phi_{17} = \{$ся, сь$\}$, $P\Phi_0^1 = \{$ка$\}$.

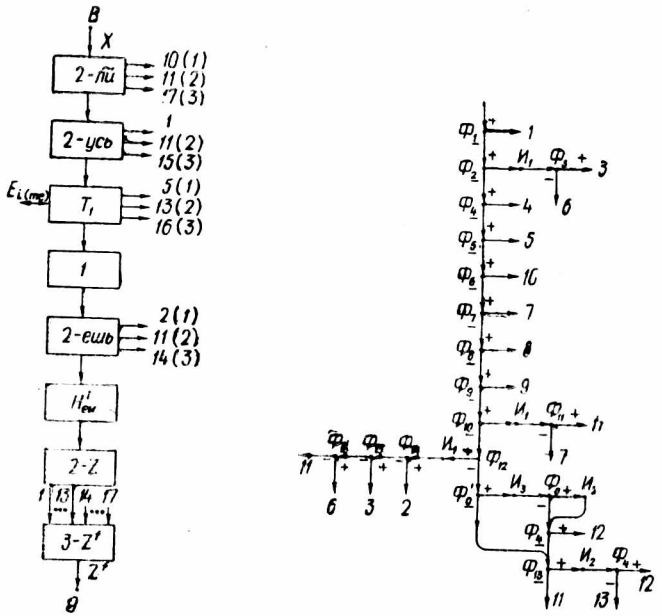

Рис. 1. Блок-схема общего классификатора личных форм (с учетом омографии).

Рис. 2. Граф-схема блока классификации общего анализатора и классификатора личных форм (блок 2-Z).

Алгоритм A_z, состоящий из последовательно соединенных блоков: постфиксной нормализации, классификации и выходного [6], описывает функцию $f_Z = \langle F_Z, X, Y_Z \rangle$ (общий анализатор личных форм), где $Y_Z = \{z_1, \ldots, z_{13}\}$, $z_i = y_i$ [3] (индекс f отбрасывается). f_Z является приближенной к функции, которую человек реализует на уровне контекста, при этом омографичным словоформам ставится в соответствие однозначный ответ. Действующая модель такой функции, как показано далее, может представить интерес наряду с моделью A_Z^1. Приведем здесь также теоретико-множественное описание функции f_Z [3].

Задавать график $F \subset X \times Y$ функции f, реализуемой человеком при обработке словоформ (X — множество словоформ, Y — признаков), простым перечислением элементов этого множества нецелесообразно и невозможно, поэтому зададим его описанием [7]

$$F = \mathfrak{E}\{\langle x, y\rangle \mid K(x, y)\}, \qquad (2)$$

$K(x, y)$ — двухместная высказывательная форма, при подстановке в которую пары $\langle x, y \rangle \in F$ получим истинное высказывание. Записывая $K(x, y)$, используем некоторые понятия теории множеств, математической логики и основываемся на анализе правил грамматики и процессов обработки словесной информации.

Известно, что «грамматическое значение имеет в языке обязательное специализированное воплощение — форму слова», причем в качестве формального показателя значения выступают флексии, суффиксы и постфиксы [8, с. 303]. Человек при определении многих признаков также основывается на формальной структуре слова. Поэтому задание $K(x, y)$ базируется на том, что каждому внутреннему (морфологическому) признаку можно поставить в соответствие некоторую совокупность внешних (формальных) признаков. $K(x, y)$ фактически представляет собой систему формальных правил отыскания морфологических признаков слова на основании внешних.

Для определения собственно грамматических значений глагольной формы обычно достаточно исследовать одну или несколько ее последних букв, на чем и основано предложенное нами в [3] описание $K(x, y)$. При необходимости можно пользоваться и другими множествами букв, например, множеством $P_{l \div k}$ из $(l-k+1)$ букв слова, начиная с l-й и кончая k-й ($l \geqslant k$), и т. д.

Рассмотренное теоретико-множественное описание функций, реализуемых при классификации словесной информации (или другом виде обработки такой информации), не является единственно возможным. Способ задания функции зависит от ее вида, конкретных целей исследования и т. п. $K(x, y)$ удобно задавать в виде формул исчисления предикатов; вместо множеств, в записи $K(x, y)$ можно использовать их характеристические функции и т. п. Предложенное теоретико-множественное описание [3] является, на наш взгляд, несколько более компактным и наглядным. Теоретико-множественное описание категорий [3] также полезно и целесообразно; оно, в частности, облегчает построение формальных категорий.

График F_Z представим в виде (2) (или в виде (3) [3]), где $K_Z(x, y) = [\bigwedge\limits_{\substack{j=\overline{1,6} \\ 8,10,12,13}} (P(x) \cap P_j \neq \varnothing \Rightarrow y = z_j) \wedge ((P(x) \cap P_7 \neq \varnothing) \wedge$

$\wedge (P(x) \cap P_{11} = \varnothing) \Rightarrow y = z_7) \wedge (P(x) \cap \bigcup\limits_{\substack{j=\overline{1,13} \\ j \neq 11}} P_j = \varnothing \Rightarrow y = z_{11})]$. $P_i = P_{\Phi_i}$ $(i = 1, 5, 7, 8, 9, 11$; последняя буква каждого признака в P_i отмечена знаком „~“, обозначающим возможность наличия постфикса *ся* или *сь* [3]), $P_2 = \{$ешь, ишь$\}$, $P_3 = \{$ет, ит, ст, есть$\}$,

$P_4 = \{$ем, им$\}$, $P_6 = \{$ат, ят, ут, ют, суть$\}$, $P_{10} = \{$ли$\}$, $P_{12} = \{$мте, м-ка, мте-ка$\}$, $P_{13} = \{$ите, ьте, йте, ите-ка, ьте-ка, йте-ка$\}$.

Алгоритмы A_Z и A'_Z доведены до уровня действующих (реализованы на ЭЦВМ «Урал-14Д» с помощью транслятора АЛГОЛ-ЦЭМИ). A_Z принят в Республиканский фонд алгоритмов и программ [9]. Модель A'_Z, в силу метода своего построения, безошибочно функционирует на любой синтетической личной форме глаголов из словаря [10] на 104 тыс. слов. При практическом функционировании A'_Z ошибок также не было обнаружено. Вероятность p_{A_Z} достоверного предсказания алгоритма A_Z, определенная (с вероятностью $p(t)=0{,}95$ и точностью $\varepsilon=0{,}002$) на данных частотного словаря [11], составляет $0{,}991 \pm 0{,}002$. Если глаголы на *ите* относить к повелительному наклонению, что целесообразно делать при действии модели на текстах разговорного стиля, то для полученного анализатора $A^{\text{п}}_Z$ — $P_{A^{\text{п}}_Z} = 0{,}994 \pm 0{,}002$.

При функционировании A_Z на выборках (сформированных с помощью таблиц случайных чисел) $n_1 = 2200$ словоформ из газетных текстов, $n_2 = 2200$ — из литературных, $n_3 = 1000$ — из математических и $n_4 = 1000$ — из технических не было обнаружено ни одной ошибки. Результаты подсчета вероятностных характеристик моделей (разработанных в статье [3]) также свидетельствуют о высоком качестве последних.

Описанные общие анализатор и классификатор гораздо проще, чем модели, полученные в результате простого объединения частных алгоритмов классификации по признакам *M, T, N, L* и *G*. Такое представление A_Z и A'_Z оказалось возможным благодаря тому, что в личных формах русского глагола грамматические значения некоторых категорий выражаются одинаковыми формальными признаками (флексиями, суффиксами и т. п.). Использованный принцип упрощения общих моделей, основанный на анализе грамматических фактов, назовем грамматическим.

Можно предложить и другие принципы упрощения, пригодные не только для моделей классификации глагольных форм, например, принцип учета иерархических отношений между категориями.

Среди категорий можно обнаружить определенную иерархию. Для анализа отношений между категориями в заданном классе словоформ введем некоторые определения. Будем говорить, что категория S^0 (обозначения соответствуют обозначениям в работе [3]) сильнее категории S^0_c и обозначать $S^0 > S^0_c$, если $S' \supset S'_c$. В случае $(S' \supseteq S'_c) \wedge (S'_c \supseteq S') \wedge (S' \cap S'_c \neq \varnothing)$ назовем категории S^0 и S^0_c нейтральными, в случае $S' \cap S'_c = \varnothing$ — независимыми, если же $S' = S'_c$ — равносильными. Будем называть категорию S^0_c подчиненной категории $S^0 = \{S^0_i, \ldots, S^0_k\}$ и обозначать $S^0 \gg S^0_c$ при выполнении одного из условий $S^0_i \supseteq S'_c$ $(i = \overline{1,k})$. Если выполнено $S^0 \gg S^0_c$, то обязательно выполнено и $S^0 > S^0_c$ $(S^0 \gg S^0_c \Rightarrow S^0_i \supseteq$

$\supseteq S_c^0 = \rangle S' \supset S_c' S^0 > = \rangle S_c^0$), обратное утверждение верно не всегда). Будем говорить, что категории S_{c1}^0, S_{c2}^0, ..., S_{cm}^0 находятся на 1-й 2-й..., m-й ступени иерархии подчинения, если $S_{c1}^0 \gg S_{c2}^0 \gg ... \gg \gg S_{cm}^0$.

Анализируя категории наклонения, времени, числа, лица и рода глагола [3], соответственно получим, что для множества X личных форм $S_м^0 > S_T^0$, $S_м^0 > S_L^0$, $S_м^0 > S_G^0$, $S_T^0 > S_G^0$, $S_N^0 > S_T^0$, $S_N^0 > S_L^0$, $S_N^0 > S_G^0$, S_T^0 и S_L^0 нейтральны, S_L^0 и S_G^0 независимы, $S_м^0$ равносильна S_N^0. Нетрудно убедиться в том, что $S_м^0 \gg S_T^0 \gg S_G^0$. Наличие между категориями отношений подчинения позволяет пользоваться при моделировании правилами I или II.

I. В первую очередь определяются значения категорий, находящиеся на высшей ступени иерархии подчинения.

II. Значения категорий определяются в порядке расположения их на ступенях иерархии подчинения. В случае, если какая-либо из рассматриваемых категорий S не связана ни с какой другой отношениями подчинения, то значения S определяются после значений категории, которая сильнее S. Если же категории нейтральны, независимы или равносильны, их значения определяют в произвольном порядке.

При использовании правил I и II внутри алгоритма как бы происходит переход от автоматического режима функционирования к полуавтоматическому (использующему дополнительные признаки), а это значительно облегчает процесс классификации. В результате использования этих правил были построены модификации общих моделей.

Разработаны и другие варианты алгоритмов A_Z и A_Z^f на основании принципа учета информативности грамматических категорий, которую можно оценить в результате логико-грамматического анализа. Рассмотрим кратко некоторые вопросы такого анализа.

Построенные модели основаны на способности человека опираться в процессе морфологической классификации на внешние признаки словоформ. Человек может также использовать при классификации внутренние признаки, упрощающие классификацию. Более того, психологические эксперименты и анализ данных грамматики показали, что грамотный человек без труда определяет некоторые внутренние признаки с помощью других, не обращаясь к словоформе. Например, если известно значение * рода (G_1, G_2 или G_3), то испытуемый легко отыщет значения всех ос-

---

* При анализе внутренних признаков нам удобнее пользоваться грамматическими категориями и значениями не в теоретико-множественном представлении, а в смысле признаков. Для конкретности описания будем производить логико-грамматический анализ на примере словоизменительных категорий и значений глагола (для личных глагольных форм). Если множество признаков категорий обозначить через CS, а множество признаков значений через C, то для выбранных категорий и значений получим, что $C^S = \{M, T, N, L, G\}$, $C = \{M_1, M_2, T_1, T_2, T_3, N_1, N_2, L_1, L_2, L_3, L_4, G_1, G_2, G_3, G_4\}$.

тальных словоизменительных категорий, не зная глагольной формы, и т. д. Это свидетельствует о тесной взаимозависимости внутренних признаков и о различных уровнях их информативности.

Если при решении задач по обработке словесной информации необходимо обращение к словоформе, назовем такие задачи грамматическими, если нет — логико-грамматическими. Целью последних является исследование взаимосвязей и взаимозависимостей внутренних признаков. Полезно изучить сочетаемость грамматических категорий и значений (о сочетаемости граммем см. в работе [12], разрозненные сведения имеются также в грамматиках). Одним из результатов логико-грамматического анализа можно считать данные об уровне информативности рассматриваемых признаков.

Уровни информативности значения и категории будем характеризовать соответственно функциями f и φ (принимающими целочисленные значения). $f(c)$ равняется числу значений (наличий значений), найденных на основании значения c без обращения к словоформе, т. е. длине кортежа $z(c)$, составленного из этих значений; а $\varphi(c^s) = \sum\limits_{c_i \in c^S} f(c_i)$.

Введя функции f^0 и φ^0, характеризующие уровни информативности грамматических категорий (S^0) и их значений, получим: $f^0(G_i)=3 (i=\overline{1,3})$, $\varphi^0(G)=9$, $f^0(T_1)=f^0(T_2)=1$, $\varphi^0(T)=2$; $f^0(L_3)=1$, $\varphi^0(L)=1$; остальные функции равны 0.

Аналогичным образом введены функции f^f и φ^f, которые определены на формальных категориях глагола и их значениях. Не представляет труда определить и функции информативности от нескольких аргументов. Значения введенных функций определены для собственно грамматических категорий и значений глагола.

Информативность категорий связана с отношениями иерархичности среди них. Так, выполнение условия $\varphi(c_1^S) \leqslant \varphi(c_2^S)$ обязательно влечет $S_{c1} > S_{c2}$. Если же $\varphi(c_1^S) < \varphi(c_2^S)$, то $S_{c1} \gg S_{c2}$. Обратные утверждения также верны.

Логико-грамматический анализ внутренних признаков может представить интерес при использовании методов дополнения, обращения и ограничения [5], при синтаксическом анализе. Сказанное выше в одинаковой мере справедливо как для слов (глаголов и других частей речи), так и для псевдослов. Идеи логико-грамматического анализа можно использовать не только для русского, но и для других языков. Несмотря на свою несомненную пользу, логико-грамматический анализ не может и не должен (за исключением, быть может, редких случаев), заменять грамматический, так как это приведет к сужению круга задач и неполному их решению. Поэтому логико-грамматический анализ, который по существу является полуавтоматическим, целесообразно рассматривать наряду с автоматическим, основанным на исследовании формальной структуры слова.

Сущность принципа информативности для упрощения общих моделей сводится к тому, что значения категорий определяются

в порядке убывания информативности последних. Алгоритм, построенный на основании принципа информативности, будет существенно отличаться от алгоритма, построенного в результате применения принципа иерархичности, так как, например, наиболее информативная категория рода (значения которой, следуя принципу информативности, определяются первыми) находится в то же время на низшей ступени иерархии подчинения (и по принципу иерархичности ее значения будут определяться последними).

Принципы учета иерархичности и информативности категорий позволяют построить более простые алгоритмы, чем алгоритм, состоящий из частных моделей. Однако в данном случае эти алгоритмы сложнее моделей A_Z и A_Z^f, полученных на основании грамматического принципа.

Модели A_Z и A_Z^f являются универсальными, используют минимум (особенно A_Z) статической информации, действуют в автоматическом режиме, с высоким быстродействием и обладают хорошими вероятностными оценками качества.

С целью минимизации времени обработки словоформ мы получили (с помощью ЭЦВМ) сведения о частоте встречаемости букв и буквосочетаний в личных формах. Эти данные учтены в моделях для повышения их быстродействия.

СПИСОК ЛИТЕРАТУРЫ

1. Шульгин И. В., Лопатченко Б. К., Пильщиков Б. В. Математическое моделирование монокулярного зрительного восприятия. — В кн.: Проблемы бионики. Вып. 9, Харьков, 1972, с. 40—44.
2. Майстровская Л. М., Ольховский Ю. Г., Шабанов-Кушнаренко Ю. П. О некоторых бинарных отношениях. — В кн.: Проблемы бионики. Вып. 9. Харьков, 1972, с. 37—40.
3. Соловьева Е. А. К вопросу о построении общего алгоритма классификации глагольных форм русского языка. — В кн.: Проблемы бионики. Вып. 15. Харьков, 1975, с. 143—149.
4. Соловьева Е. А. Исследование процессов классификации омографичных глагольных форм. — В кн.: Проблемы бионики. Вып. 16. Харьков, 1976, с. 104—114.
5. Шабанов-Кушнаренко Ю. П., Соловьева Е. А. Бионические аспекты моделирования речевого поведения человека. — В кн.: Проблемы бионики. Вып. 13. Харьков, 1974, с. 59—66.
6. Соловьева Е. А. Автоматический морфологический анализ суженной парадигмы глагола. — В кн.: Проблемы бионики. Вып. 12. Харьков, 1974, с. 139—142.
7. Шиханович Ю. А. Введение в современную математику. М., «Наука», 1965. 376 с.
8. Грамматика современного русского языка. Отв. ред. Шведова Н. Ю. М., «Наука», 1970. 767 с.
9. Соловьева Е. А. Алгоритм автоматического определения значений числа у глагольных форм. — РФАП АН УССР. Справка № 70, Киев, 1974.
10. Орфографический словарь русского языка. Изд. 11-е. Под ред. Бархударова С. Г. и др. М., «Сов. энциклопедия», 1971, с. 520.
11. Штейнфельдт Э. А. Частотный словарь современного русского литературного языка. Таллин, 1963. 316 с.
12. Волоцкая З. М., Молошная Т. Н., Николаева Т. М. Опыт описания русского языка в его письменной форме. М., «Наука», 1964. 186 с.