ЛЕКСИКО-ГРАММАТИЧЕСКИЙ КЛАССИФИКАТОР СВОЙСТВ ЯПОНСКОГО ЯЗЫКА

С. Ю. Кравченко

Минский государственный лингвистический университет
Минск, Беларусь
E-mail: crux57005@tut.by

В данной работе представлен лексико-грамматический классификатор свойств японского языка. Классификатор разработан с целью применения в системах автоматического лингвистического анализа текстов на японском языке. Предлагаемый подход апробирован в рамках лингвистического процессора экспертной системы Goldfire.

Ключевые слова: японский язык, лингвистический процессор, лексико-грамматический анализ.

В настоящее время существует большое количество систем автоматического анализа естественных языков. Известно, что разработка математико-алгоритмического и лингвистического компонентов таких систем начинается с анализа свойств обрабатываемого естественного языка. Так, например, при обработке текстов русского языка система должна эффективно работать с развитой системой словоизменения, богатой системой флексий и спряжений, а также со свободным порядком слов в предложении. При обработке текстов, например, английского языка особое внимание уделяется лексико-грамматической многозначности слов и синтаксической структуры предложения. Другими словами, особенности и свойства языка являются определяющим критерием в организации лингвистического процессора. Особую роль при этом играет выбор классификатора свойств языка, поскольку именно этот компонент лингвистического процессора является показателем степени формализации лингвистических явлений, присутствующих в обрабатываемых текстах.

Особый интерес с точки зрения организации лингвистического компонента системы обработки текстов представляет собой лексико-грамматический строй японского языка. Японская письменность состоит из трех основных частей, кандзи (иероглифов, заимствованных из Китая), и двух слоговых азбук – каны, созданных в Японии на основе кандзи – катаканы и хираганы. Смешанное написание из чередующихся иероглифов и каны без пробелов является нормой современного японского письма. По грамматическому строю – это агглютинативный язык с преимущественно синтетическим выражением грамматических значений.

При моделировании лексико-грамматического содержания текста первоочередной задачей является определение границ и свойств минимальных лексических единиц (МЛЕ). Отсутствие индикаторов границ слов в японском языке значительно усложняет эту задачу. Предлагаемый подход к ее решению не относится к традицион-

ным японским [8–10] или постсоветским [4–7] лингвистическим теориям, а является их переосмыслением с целью создания формальной модели лексико-грамматического анализа, по возможности, лишенной типичных лингвистических противоречий, связанных с описанием лексико-грамматического строя японского языка. Основным отличием данного подхода можно назвать выделение морфем в качестве минимальных лексических единиц содержания текста. Определение границ морфем и их типов в японском языке, хотя и не лишено некоторых противоречий, представляется задачей более тривиальной, чем выделение границ и типов слов.

Японские морфемы можно поделить на 3 типа – корневые, суффиксальные и служебные. Отталкиваясь от этих базовых типов, был разработан лексико-грамматический классификатор, который мог бы наиболее полно отображать лексико-грамматический строй японского языка, и при этом быть предельно простым в понимании и применении с целью организации ресурсов ЛБЗ и ручной разметки базового корпуса текстов.

Классификатор включает 151 лексико-грамматический код, из которых 34 кода используются для описания корневых морфем, 98 кодов – для суффиксальных морфем, и 19 кодов – для служебных морфем.

Корневые морфемы можно условно разделить на корни существительных (3 кода), корни прилагательных (3 кода), корни наречий (1 код), корни местоимений (8 кодов), корни глаголов (17 кодов), корни числительных (1 код) и междометия (1 код).

К корням существительных относятся:

1) слова японского языка, обычно записываемые при помощи кандзи или хирагана (код NN) – 先生 (СЭНСЭЙ, учитель), 大学 (ДАЙГАКУ, университет), 教室 (КЁ:СИЦУ, аудитория) и т. д.;

2) слова, заимствованные из других языков, которые обычно записываются при помощи катакана, ромадзи или с использованием любого не японского письма (код FW) – テーブル (ТЭ:БУРУ, стол), プリンター (ПУРИНТА:, принтер), table (англ. стол), printer (англ. принтер) и т. д.;

3) любые имена собственные, сокращения или условные обозначения (код NP) – 山本(Ямамото), ベラルーシ (Беларусь), Manchester (Манчестер), DVD, LCD и т. д.

К корням прилагательных относятся:

1) корни предикативных прилагательных (код JJI), т. е. корни прилагательных, которые в начальной форме присоединяют суффикс い (И) и имеют соответствующую форму спряжения, а также форму прошедшего времени – 面白[い] (ОМОСИРО[Й], интересный), 楽し[い] (ТАНОСИ[Й], веселый) и др.;

2) корни полу-предикативных прилагательных (код JJN), т. е. корни прилагательных, которые в начальной форме не имеют суффикса, но присоединяют суффиксы な или の, когда используются в атрибутивной роли – 便利(БЭНРИ, удобный), 静か (СИЗУКА, тихий) и др.;

3) корни непредикативных прилагательных (код PRA), т. е. корни неизменяемых прилагательных, или короткие формы предикативных прилагательных – いわゆる (ИВАЮРУ, так называемый), ある (АРУ, некоторый), 高(КО:, высокий), 低(ТЭЙ, низкий) и др.

Среди примеров корней наречий можно назвать ゆっくり (ЮККУРИ, медленно), 時々 (ТОКИДОКИ, время от времени, иногда), 更に (САРАНИ, к тому же), 再び (ФУТАТАБИ, снова) и др.

К корням местоимений относятся:

1) местоимения 1-го, 2-го и 3-го лица (коды PP1, PP2, PP3 соответственно) – 私 (ВАТАСИ, я), 俺 (ОРЭ, я), 僕 (БОКУ, я), 貴方 (АНАТА, ты, вы), 彼 (КАРЭ, он), 彼女 (КАНОДЗЁ, она) и др.;

2) указательные местоимения (коды PS, PD), т. е. субстантивные местоимения, а также местоимения места и направления – これ (КОРЭ, это), それ (СОРЭ, то), ここ (КОКО, здесь), そこ (СОКО, там), こっち (КОТТИ, сюда), そっち (СОТТИ, туда), и др.;

3) вопросительные местоимения (код PQ) – 何(НАНИ, что), 誰(ДАРЭ, кто), 何故 (НАЗЭ, почему), どの (ДОНО, какой), どこ (ДОКО, где), どっち (ДОТТИ, который, куда) и др.;

4) определительные местоимения (код PDT) – この (КОНО, этот), その (СОНО, тот), あの (АНО, тот) и др.;

5) возвратные местоимения (код PREF) – 自身 (ДЗИСИН, сам (одушевленный)), 自分 (ДЗИБУН, сам, свой), 本体 (ХОНТАЙ, сам (неодушевленный)) и др.;

Корни глаголов включают:

1) корни глаголов 1-й формы спряжения [2], которые в начальной форме присоединяют суффиксы ぶ, ぐ, く , む, ぬ, る, す, つ, う (коды VB1B, VB1G, VB1K, VB1M, VB1N, VB1R, VB1S, VB1T, VB1W соответственно) – 行 [く] (И [КУ], идти), 話 [す] (ХАНА [СУ], разговаривать) и др.;

2) корни глаголов 2-й формы спряжения [2] (код VB2) – 食べ [る] (ТАБЭ [РУ], есть), 見[る] (МИ [РУ], видеть) и др.;

3) корни глаголов особой формы спряжения (коды VBS, VBK) – す [る] (СУ [РУ], делать), 来[る] (КУ [РУ], приходить);

4) корни глаголов «наличия» или «обладания», которые часто используются как вспомогательные (коды VBA, VBO, VBI, VBDE) – あ [る] (А [РУ], быть, иметь), お [る] (О [РУ], быть, иметь), い [る] (И [РУ], быть, иметь), であ [る] (ДЭА [РУ], быть);

5) корень вспомогательного глагола ま [す] (МА [СУ]) (код VBMA), который используется в вежливой устной речи.

К числительным относятся как слова, записанные иероглифами, так и цифрами (арабскими или римскими) – 三(САН, три), 三つ (МИЦУ, три), XVI, 3.14 и др.

Суффиксальные морфемы представлены наиболее многочисленной группой кодов. Условно, их можно поделить на следующие подклассы – суффиксы форм и спряжений глаголов и прилагательных (63 кода), словообразующие суффиксы – (12 кодов), семантические суффиксы (5 кодов), суффиксы числительных (3 кода), показатели падежей (11 кодов), именные послелоги (1 код) и частицы (3 кода). Суффиксальные морфемы являются закрытой частью лексико-грамматического классификатора – их количество невелико и строго ограниченно.

К суффиксам форм и спряжений глаголов и прилагательных можно отнести:

1) все суффиксы глагольных основ, которые являются частью спряжения японских глаголов (коды SV1BA, SV1BI, SV1GA, SV1GI, SV2U, SV2E, SV3U, SV3E и др.);

2) залоговые суффиксы (коды SPRA, SPRE, SPSA, SPSE) – например, в глаголе 食べられる (ТАБЭРАРЭРУ, быть съеденым) присутствуют суффиксы ら (SPRA) и れ (SPRE), или в глаголе 食べさせる (ТАБЭСАСЭРУ, заставить съесть) – さ (SPSA) и せ (SPSE);

3) суффиксы временных и деепричастных форм (коды SPTE и SPTA) – например, в глаголе 食べた (ТАБЭТА, съел), た (SPTA) – суффикс прошедшего времени;

4) суффиксы спряжения прилагательных (коды SPI, SPKU), как, например, суффикс い (И) в 面白い (ОМОСИРОЙ, интересный);

5) суффиксы отрицательных форм (код SPNEG) – например, в слове 面白く ない (ОМОСИРОКУНАЙ, неинтересный), которое состоит из 面白 (ОМОСИРО, корень прилагательного JJI), く (КУ, суффикс спряжения SPKU), な (НА, отрицательный суффикс SPNEG) и い (И, суффикс спряжения SPI);

6) модальные суффиксы (коды SPM, SPMTA, SPMIMP), как, например, そう (СО:) в 面白そう (ОМОСИРОСО:, казаться интересным); た (ТА) в 食べたい (ТАБЭТАЙ, хотеть съесть); ろう (РО:) в 食べろう (ТАБЭРО:, ешь!).

Словообразующие суффиксы включают суффиксы, которые, присоединяясь к корневым морфемам, образуют новые слова (коды SJJ, SRB, SNNBI, SNNGI, SNNKI, SNNNI, SNNMI, SNNRI, SNNSI, SNNTI, SNNWI, SNNJJ). Так, например, суффикс し (СИ, код SNNSI), присоединяясь к корню глагола 話[す] (ХАНА [СУ], разговаривать), образует слово 話し (ХАНАСИ, беседа)). Другим наглядным примером применения словообразующих суффиксов может служить наречие 運命的に (УНМЭЙТЭКИНИ, судьбоносно), которое состоит из корня существительного 運命 (УНМЭЙ, судьба), словообразующего суффикса 的 (ТЭКИ), в результате присоединения которого образуется прилагательное 運命的 (УНМЭЙТЭКИ, судьбоносный), и еще одного словообразующего суффикса に (НИ), в результате присоединения которого образуется собственно наречие 運命的に.

К семантическим суффиксам относятся:

1) семантические префиксы (код PRD), такие как 不 (ФУ, не-), 未 (МИ, не-), 半 (ХАН, полу-);

2) семантические постфиксы (код POC), такие как 化 (КА) в слове アクチブ化 (АКУТИБУКА, активизация), или 性 (СЭЙ) в слове 危険性 (КИКЭНСЭЙ, опасность);

3) суффиксы множественного числа (код POPL), например суффикс 達 (ТАТИ) в слове 先生達 (СЭНСЭЙТАТИ, учителя);

4) префиксы вежливости (код PRHO), такие как お (О) или ご (ГО), которые наиболее часто употребляется в устной речи;

5) звательные постфиксы (код POPE), такие как, например, さん (САН), 君 (КУН), ちゃん (ТЯН), которые присоединяются к существительным для обозначения одушевленности и социальной принадлежности.

К суффиксам числительных относятся морфемы для обозначения счетных или порядковых числительных (коды PRNUM и PONUM), такие как, например, 第 (ДАЙ, код PRNUM) или 目 (МЭ, код PONUM) в словах 第二番 (ДАЙНИБАН, второй) или 二番目 (НИБАММЭ, второй), а также счетные суффиксы и общепринятые единицы измерения (код NNUM), как 本 (ХОН, суффикс для счета цилиндрических предметов) в 三本 (САНБОН, три штуки) или m2 в 10 m2.

Показатели падежей включают 11 подклассов морфем (коды CHA, CGA, CNO, CNI, CHE, CWO, CDE, CKARA, CTO, CYORI, CMADE), каждый из которых соответствует падежу японского языка.

Именные послелоги (код POPR) включают морфемы, зачастую совпадающие по написанию с некоторыми корнями существительных, но использующиеся в роли обстоятельственных показателей – 時 (ТОКИ, время (сущ.) или во время чего-л. (им.послелог)), 為 (ТАМЭ, причина (сущ.) или по причине чего-л. (им. послелог)), 結果 (КЭККА, результат (сущ.) или в результате чего-л. (им. послелог)).

Частицы включают морфемы, обозначаемые кодами PRTMO (для частицы も (МО), которая часто является индикатором общего (бессуфиксального) падежа), PRKTA (для частицы か (КА), которая используется при образовании неопределен-

ных местоимений и вопросительных предложений), и PRT (для всех остальных частиц).

Служебные морфемы включают знаки препинания – восклицательный знак, запятая, точка, многоточие, вопросительный знак, открывающая круглая скоба, закрывающая круглая скобка, открывающая квадратная скобка, закрывающая квадратная скобка, отрывающие кавычки, закрывающие кавычки, тире, двоеточие, точка с запятой (коды PME, PMC, PMFS, PMEL, PMQ, BRO, BRC, BSO, BSC, QO, QC, DA, :, ; соответственно), части математических формул и другие специальные обозначения (код FO), сочинительные союзы, подчинительные союзы, и связки (коды CC, CS, LNK соответственно), и японские субстантиваторы (код SUBST).

Предлагаемый классификатор был успешно апробирован в рамках системы лексико-грамматического анализа экспертной системы Goldfire. Точность лексико-грамматического анализа составила 97–98 %, что позволяет говорить о практической обоснованности выбранного подхода.

ЛИТЕРАТУРА

1. *Старостин, С. А.* Алтайская проблема и происхождение японского языка / С. А. Старостин. М., 1991. 298 с.

2. *Лаврентьев, Б. П.* Практическая грамматика японского языка / Б. П. Лаврентьев. М., 2002. 352 с.

3. *Алпатов, В. М.* Грамматика японского языка / В. М. Алпатов, С. А. Старостин, И. Ф. Вардуль. М., 2000. 149 с.

4. *Модина, Л. С.* Принципы анализа японских текстов при моделировании японской лексико-морфологической системы / Л. С. Модина, З. М. Шаляпина // Тр. междунар. семинара ДИАЛОГ '96 по компьютерной лингвистике и ее приложениям [под ред. А. С. Нариньяни]. М., 1996. С. 169–174.

5. *Пашковский, А. А.* Слово в японском языке / А. А. Пашковский. М., 2006. 208 с.

6. *Алпатов, В. М.* Структура грамматических единиц в современном японском языке / В. М. Алпатов. М., 1979. 148 с.

7. *Ефимова, С. К.* Лексикология японского языка / С. К. Ефимова, Е. С. Руфова // Электронный учебник на сайте ЯГУ, 2005. 1,5 п. л.

8. *Takeuchi, K.* HMM Parameter Learning for Japanese Morphological Analyser / K. Takeuchi, Y. Matsumoto // Procs. Of 10th Pacific Asia Conference. Language, Information and Computation. 1995, P.163–172.

9. *Matsumoto, Y.* Morphological Analysis System ChaSen version 2.2.9 Manual / Y. Matsumoto, [et al]. Nara, 2002.

10. *Uchimoto, K.* The unknown word problem: a morphological analysis of Japanese using maximum