ЦИФРОВЫЕ МЕТОДЫ В ЛИНГВИСТИКЕ | DIGITAL METHODS IN LINGUISTICS

От российской синологической лингвистики XX в. к нейросетевым моделям XXI в.: градуальная природа китайского слова

From Russian Sinological Linguistics of the 20th Century to 21st-Century Neural Models: The Gradience of the Chinese Word

Кожа Ксения Анатольевна

Кандидат филологических наук, Институт востоковедения Российской академии наук, Отдел языков, научный сотрудник

ksenia.kozha@gmail.com

Koža Ksenia

PHD (linguistics), Institute of Oriental Studies of Russian Academy of Sciences, Languages Department, Researcher

ORCID: 0000-0003-2717-6156

В статье рассматривается проблема выделимости слова в китайском языке в контексте верификации теоретических разработок советской востоковедной лингвистики с применением современных методов обработки естественного языка. Анализируется теоретическое наследие советских синологов (Е .Д. Поливанова, Ю. В. Рождественского, В. М. Солнцева), которые одними из первых отказались от европоцентричного подхода и предложили концепцию континуума между морфемой, словом и словосочетанием в изолирующих языках. Показано, что их идеи о кросс-уровневой природе китайского слова и градуальности словности находят эмпирическое подтверждение в современных корпусных исследованиях и нейросетевых моделях сегментации. Опыт конфликта стандартов разметки, успехи мультикритериального обучения и эффективность методов *Whole Word Masking* демонстрируют, что слово в китайском языке представляет собой не дискретную единицу, а многомерный узел признаков с градиентными значениями. Делается вывод о продуктивности обращения к классическому теоретическому наследию для решения актуальных задач вычислительной лингвистики.

Ключевые слова: китайский язык, слово, градуальность словности, изолирующие

This article examines the problem of word delimitation in Chinese in the context of continuity between classical linguistic approaches and modern natural language processing methods. The theoretical legacy of Soviet sinologists (E. D. Polivanov, Yu. V. Rozhdestvensky, V. M. Solntsev) is analyzed; they were among the first to reject the Eurocentric approach and propose the concept of a continuum between morpheme, word, and phrase in isolating languages. It is shown that their ideas about the cross-level nature of the Chinese word and the gradience of wordhood find empirical confirmation in contemporary corpus studies and neural segmentation models. The experience of conflicting annotation standards, the success of multi-criteria training, and the effectiveness of *Whole Word Masking* methods demonstrate that the word in Chinese is not a discrete unit but a multidimensional feature bundle with gradient values. The conclusion is drawn that returning to classical theoretical heritage is productive for solving current computational linguistics challenges.

Keywords: Chinese language, word, wordhood gradience, isolating languages, text

языки, сегментация текста, корпусная лингвистика, NLP, нейросетевые модели, лихэци

segmentation, corpus linguistics, NLP, neural models, liheci

Введение

Теоретическое наследие советских китаистов-языковедов продолжает оказывать влияние на современную лингвистику. Многие идеи, такие как теория «инкорпораций» Е. Д. Поливанова, взгляды на природу китайского слова Ю. В. Рождественского, построение теории о частях речи А. А. Драгунова, типологические изыскания В. М. и Н. В. Солнцевых в обширном поле изолирующих языков, диахронические исследования С. Е. Яхонтова находят подтверждение и развитие в современных подходах, включая корпусные методы, компьютерную лингвистику и языковой анализ с применением моделей искусственного интеллекта. Настоящая статья имеет целью продемонстрировать значимость обращения «к истокам» на примере одной из классических проблем востоковедной лингвистики – проблемы выделения слова.

Вопрос о выделимости слова в китайском языке на протяжении последнего столетия остается одной из центральных методологических проблем общих и востоковедных исследований. Отсутствие графических пробелов между словами в китайской письменности, преимущественно моносиллабическая морфология с высокой степенью лексикализации и аналитизма, широкая зона переходных явлений между словом и словосочетанием, а также функционально-прагматическая вариативность единиц на стыке морфемы и слова – все это системно усложняет определение статуса слова как универсальной единицы языка.

В современных прикладных областях (корпусная лингвистика, автоматическая сегментация текста, *lexicon induction*) эта проблема дополнительно обостряется, поскольку алгоритмам необходимо устойчивое множество признаков для сегментации и токенизации. Лингвистическая теория в ответ предлагает многокритериальные модели «словности» (*wordhood*), комбинирующие фонологические, морфологические, синтаксические, семантические, а также дискурсивно-прагматические или психолингвистические критерии, и, применительно к китайскому, неизбежно сталкивается с тем, что эти критерии расходятся. Для лингвистического анализа используются статистические и нейросетевые модели, которые, строго говоря, лишь имитируют человеческую метаязыковую интуицию, а не извлекают слово как лингвистическую универсалию. Для описания принципов работы таких моделей и полученных с их применением выводов необходимо снова и снова обращаться к теоретическим основаниям, и пересмотр этих оснований в свете новейших данных позволяет по-новому осознать теоретическую интуицию наших соотечественников – лингвистов-синологов прошлого столетия.

Китайское слово в трактовке советских лингвистов.

Примечательно, что в своих работах языковеды советского периода стремились не к теоретическому обоснованию идеального «слова» по меркам, выработанным для индоевропейских языков, а к построению универсальной типологической рамки, с помощью которой можно было бы адекватно описать специфику изолирующих языков. Классические подходы, основанные на материале флективных языков, часто приводили либо к отрицанию существования слова как такового, либо к его искусственному обнаружению через механическую интерпретацию статистически частотных двусложных сочетаний. В этом контексте отечественная школа синологической лингвистики сыграла революционную роль, предложив принципиально иную методологию.

Столетие тому назад Евгений Дмитриевич Поливанов (1891—1938), выдающийся советский лингвист и востоковед, внес значительный вклад в изучение китайского

языка, предложив оригинальный подход к анализу его базовых единиц. В частности, в «Грамматике современного китайского языка» (1930), созданной в соавторстве с А. И. Ивановым, Поливанов выдвинул тезис о том, что соответствие языковых единиц разных уровней (фонема/морфема/слово/словосочетание) между европейскими и китайским языками носит «далеко не эквивалентный характер» [2, с. 7]. Как справедливо отмечает в своём исследовании В. М. Алпатов, концепция Поливанова подразумевала «нетрадиционный подход к разграничению морфологии и синтаксиса в китайском языке» [1, с. 184]. При этом советский языковед широко задействовал акцентуационные (просодические), формальные (аффиксация, анализ структур предикации), диахронические признаки.

Поливанов одним из первых в отечественной лингвистике отказался от прямого переноса критериев, выработанных на основе анализа индоевропейских языков. Его ключевой тезис заключался в том, что словарная единица и синтаксическая конструкция образуют непрерывный спектр или континуум. Для китайского языка определяющим признаком являлась семантическая цельность и смысловая целостно-оформленность значения. Он последовательно сближал словообразование и синтаксис, фиксируя в качестве базовых описательных единиц устойчивые двусложные комплексы (чуть позже И. М. Ошанин назовет их «биномами», этот термин разовьёт и А. А. Драгунов[1]).

Новизна интерпретации соотношения языковых уровней проявляется, например, в разделе о так называемых раздельно-слитных словах китайского языка 离合词 *líhécí* лихэци – глагольно-именных комплексах, обладающих признаками как слова, так и словосочетания в виду возможности постановки различных элементов между компонентами комплекса. Поливанов рассматривал их в числе «инкорпораций»[2] – сочетаний двух и более лексических морфем. В частности, он отмечал, что для слов типа 吃饭 *chīfàn* «принимать пищу», 睡觉 *shuìjiào* «спать», 说话 *shuōhuà* «говорить», 知道 *zhīdào* «знать» вторая морфема как бы «привлекается» в виде «пустого, формально лишь обусловленного, а логически излишнего объекта», стремясь к компенсации двусложной нормы китайского слова, которая «является (на правах принципиального минимума) организующей нормой китайской морфологии» [2, с. 8]. Таким образом, «инкорпорация из 2х лексических морфем соответствует, с одной стороны, «простым единым словам русского языка («есть», «спать», «сказать» и т. д.), а с другой стороны – и словосочетаниям (комплексам из двух слов) русского языка» [2, с. 9]. Иначе говоря, «классификационные пороги в области морфологических величин (каковыми являются слово, с одной стороны, и словосочетание с другой) опять-таки принципиально не совпадают между русским и китайским языками, точно так же как не совпадают в этих языках и фонетические критерии элементарной фонетической величины» [2, с. 9].

---

[1] В своем докладе на XXV Международном конгрессе востоковедов Н. Н. Коротков представлял категорию «бинома» так: «Внутри категории бинома, как мы видели, выделяются две полярные противоположности – бесспорное слово и бесспорное словосочетание; между ними располагаются комплексы, в которых это разграничение принципиально невозможно. Однако, поскольку бином функционирует в предложении как единое целое, для китайской речи такое разграничение, видимо, и не является существенным» [6, с. 106]. И далее, рассуждая о экспонентах значимого однослога и бинома: «Между этими двумя структурными единицами, характеризуемыми одновременно и количественными и качественными признаками, наблюдается известный параллелизм: значимый однослог может выступать в качестве: а) полного слова, б) неполного слова (величины, пограничной между словом и морфемой), в) части слова (морфемы). Бином может выступать в качестве: а) единого слова, б) протяженного слова (величины, пограничной между словом и словосочетанием), в) соединения слов (словосочетания лексического или синтаксического). Обе эти основные структурные единицы представляют, таким образом, более общие категории, чем слово и морфема (в первом случае), слово и словосочетание (во втором), и включают их в себя вместе со всеми промежуточными явлениями» [6, с. 107]

[2] Термин «инкорпорация» у Поливанова – это эвристический ярлык для композитов, так он называл все сложные слова вообще; в современной типологии «инкорпорация» имеет более узкое значение (включение имени в глагольный комплекс).

Совсем недавно проблема *лихэци* была детально исследована П. О. Кисель в её кандидатской диссертации [4], а вопрос о «пустом дополнении» рассматривался, в частности, в отдельной статье чуть ранее [3]. Автор опиралась, главным образом, на новейшие разработки китайских и других зарубежных лингвистов об этом дискуссионном классе единиц, а также на эмпирические данные, полученные на основе работы с корпусом китайского языка Пекинского университета [7]. В заключении этого исследования высказана мысль о том, что «вопрос уровневой принадлежности *лихэцы* звучит в контексте общетеоретической проблемы универсального определения слова и идей европейской лингвистики, в результате чего он до сих пор не получил однозначного решения, хотя сущность двусложных комплексов в китайском языке очевидно не соответствует представлениям о слове в языках иной структуры. Возможность дистантного функционирования компонентов двуслога в современном китайском языке, для которого не актуален словоцентрический подход, не представляет собой уникальный феномен, а, напротив, является характерной чертой данной языковой системы» [3, с. 121].

Важно подчеркнуть, что феномен *лихэци* не исчерпывает проблематику оценки природы слова в китайском языке, а лишь наглядно иллюстрирует ее. На современном этапе исследований различных параметров слова и его функционального поведения для большинства авторов очевидна общая закономерность: в китайском языке слово представляет собой градуальную и контекстно-зависимую категорию, реализуемую через совокупность разноуровневых признаков, а не через набор морфологических критериев. Такая градуальность или континуум постулируются, например, такими исследователями китайской морфологии, как Arcodia 2012 [15]; Bisang 2004 [17]; Myers 2022 [24]. Так, например, Аркодия, ссылаясь на предшественников, считает, что «словообразование [в китайском] это категория с размытыми границами, как в плане флексий, так и в плане словосложения. Как отмечают Науманн и Фогель [25, с. 929], «словоизменение, словообразование и лексика как таковая, по-видимому, представляют собой лишь центральные точки на протяжении более общего континуума, простирающегося от грамматики к лексике»; в плоскости такого континуума словообразование ближе к лексике, тогда как словоизменение – к грамматике (cf. Bybee 1985: 82)» [15, с. 11; 18, p. 82].

Эта мысль естественным образом переадресует нас к трудам Е.Д. Поливанова, который, подводя итоги теоретических рассуждений во вступительной части своей «Грамматики», писал: «вместо двух принципиально отличных систем формально-семантического оформления материала – *морфологии* и *синтаксиса*, как в европейских языках (русском, например), в китайском мы имеем нечто *одно* – одни и те же общие схемы сочетания (соположения) значимых единиц, будь то единицы порядка *морфемы* или единицы порядка *слов* (и в том числе именно многосложные инкорпорации, доходящие даже до нашего понятия предложения)» [2, с. 22].

Несколько десятилетий спустя, Ю. В. Рождественский в своих работах развивал компромиссную теорию, сочетавшую структурные и функциональные аспекты анализа китайского языка. Предвосхищая современные интегративные подходы, он подчеркивал, что морфологическая специфика китайского должна рассматриваться во взаимосвязи с фонологией, синтаксисом и семантикой. Природа слова долгое время оставалась ключевой темой его исследований. В своей работе, посвященной понятию формы слова в китайском, он резюмирует: «в результате исторического развития научных представлений в современном китаеведении, в общем говоря, складываются два различных понятия формы слова: 1) форма слова как его морфологическая членимость, то есть способность распадаться на знаменательную и формальную части, где формальная часть служит базой для выражения грамматического значения слова, 2) форма слова как его категориальная отнесенность, принадлежность к какой-либо грамматической категории. В последнем случае базой для выражения грамматического значения служит все

слово в целом. Категориальное значение слова проявляется в разных формальных приметах, в основном синтаксических и отчасти морфологических»[3] [8, с. 136].

Иными словами, интерпретируя этот вывод Рождественского в контексте основной темы данного обзора, можно говорить о кросс-уровневом распределении формы слова в китайском – на морфологическом и синтаксическом уровнях одновременно. Примечательно, что, говоря об условности границ морфемного и синтаксического поля в китайском, и Поливанов, и Рожественский одновременно стремились подчеркнуть универсальный статус слова и значимость самого понятия для сохранения возможности системного подхода к описанию языка.

Наиболее мощный толчок языковой теории в направлении кросс-уровневого анализа китайских языковых единиц в широком типологическом контексте дал В. М. Солнцев – один из выдающихся архитекторов общей теории изолирующих языков в отечественной традиции. Для нашей темы важны два его принципа: 1) слово – универсальная единица языка, но его, говоря языком современной лингвистики, «признаковый пучок» (*feature bundle*) варьирует типологически; 2) критерии «словности» (*wordhood*) многомерны и лежат на разных уровнях: фонетическом/просодическом, морфологическом, синтаксическом, семантическом[4]. В изолирующих языках, по Солнцеву, доминируют синтактико-семантические критерии при редукции собственно морфологических. Наблюдения над китайским, вьетнамским, тай-кадайскими и др. послужили материалом к формированию общей идеи «размытости» границ между уровнями в языках аналитического строя: «Преобладание в индоевропейских языках производных слов, которые, как мы знаем, имеют четкое отличие и от морфемы, и от словосочетания, делает проблему отграничения слова в этих языках от других единиц не столь острой, как в формоизолирующих языках, где как раз преобладают те структурные разновидности слов (простые слова и сложные слова), которые далеко не всегда лучшим образом отграничиваются от смежных единиц. <…> Именно в силу преобладания в этих языках структурных категорий слов, которые находятся как бы по краям (верхнему и нижнему) словесной сферы, т.е. словесного уровня, где возможна размытость границ слов, у многих, чаще всего у неспециалистов по формоизолирующим языкам, создается впечатление размытости категории слова в формоизолирующих языках вообще» [12, с. 132].

«Словность» по Солнцеву это поле признаков с градиентными значениями. Это означает наличие широких промежуточных зон между морфемой, словом и словосочетанием в китайском языке. Такая экстраполяция понятия формы слова на смежные уровни позволяет выйти за рамки классической дискуссии о правомерности наличия слова в китайском[5].

И Рождественский, и Солнцев развивали идею о том, что определение границ сложного слова в китайском должно осуществляться в речевой цепи, а не в языковой статике.

---

[3] В работах по китайской грамматике середины XX в. подход Рождественского поддерживался. В лексикографической практике реализация этой концепции не была доведена до конца. В ряде советских словарей середины XX века границы между словом и словосочетанием колебались, что связано не столько с теорией, сколько с методикой ее применения.

[4] Интересно привести здесь сравнение с наиболее современным подходом Паккарда, который выделяет для китайского восемь различных уровней/определений слова: орфографический, социологический, лексический, семантический, фонологический, морфологический, синтаксический, психолингвистический [26, p. 7-13].

[5] Заметим, что эта позиция сегодня, как и столетие назад, разделяется не всеми лингвистами-типологами. Например, норвежский исследователь Х. Эйфринг (Halvor Eifring) радикально подошел к вопросу о неприменимости традиционного понятия «слова» к китайскому языку как к терминологическому ярлыку, ограничивающему применение альтернативных подходов. В работе 2013 г. он выдвинул гипотезу, что китайская речевая цепь может обходиться без слов как обособленных единиц, а состоит преимущественно из морфемных комбинаций разной степени спаянности. По наблюдениям Эйфринга, многие традиционные критерии слова (ударение, изменение формы, возможность изолированного употребления) в китайском «не срабатывают», что вызывает необходимость пересмотра подходов к описанию китайского языкового слова вне «словности» как таковой [20].

Так, Рождественский отмечал: «Возникает вопрос: не объясняется ли неудача в определении границ китайских сложных слов тем, что они вольно или невольно приравнивались к словам индоевропейских языков, которые по грамматической организации являются единицами языка? Ведь применявшиеся до сих пор критерии – ударение, морфемный анализ, подстановка и т. п. – суть критерии для выделения единицы, грамматически организованной как индоевропейское слово. Если же признать, что китайское сложное слово по своей грамматической форме – единица речи, значит, нужны иные критерии» [9, с. 180]. Солнцев позже системно обобщит это наблюдение: «Поиски границ сложного слова и словосочетания становятся бессмысленными уже лишь по установлении факта, что во многих языках при определенных условиях сложное слово и словосочетание принципиально неразличимы в силу того, что в самом языке между этими единицами различия нет. Максимально, что можно сказать по поводу таких образований, – это являются ли они единицами языка или единицами речи» [11, с. 173].

Таким образом, представление о границах сложного слова в китайском языке, выработанное на этой теоретической базе, подчеркивает необходимость ориентироваться на речевую цепь, а не только на грамматические критерии, часто перенесённые из индоевропейской языковой теории. Это создало предпосылки для современных исследований, развивающих понимание слова как динамической единицы речи, не всегда совпадающей с единицей языка.

Проблема слова в общетеоретической лингвистике начала XXI в.

Важно отметить, что к началу «цифровой эпохи», когда применение корпусных методов приняло глобальные масштабы, идеи о градуальности и многомерности слова уже активно обсуждались в мировой теоретической лингвистике. Дискуссия вышла за рамки типологического противопоставления языков и сфокусировалась на универсальности самого понятия «слово».

Представляется естественным, что конфликт различных параметров для выделения слова не является уникальной чертой китайского языка. Рассогласование морфологических, просодических и синтаксических критериев «словности» наблюдается в самых разных языковых семьях. Так, классические примеры включают немецкие композиты, которые сохраняют внутреннюю синтаксическую структуру фразы[6] или французские местоимения-клитики, которые фонологически примыкают к глаголу (аффиксы), но синтаксически являются самостоятельными аргументами (отдельными словами). В тюркских языках длинные агглютинативные цепочки морфем могут синтаксически вести себя как единое целое, но фонологически делиться на несколько просодических единиц и т.д.

По мере увеличения объёма данных о языках различной структуры нарастала тенденция фундаментальной критики «слова» как кросс-языковой универсалии. Мартин Хаспельмат (Martin Haspelmath) в своей ставшей флагманской работе [23] последовательно доказывает, что не существует единого набора необходимых и достаточных критериев (будь то «потенциальная пауза» (potential pause), «свободное употребление» (free use), «мобильность» (mobility) или «непрерываемость» uninterruptibility [23, p. 39–45] для универсального определения морфосинтаксического слова. Он приходит к выводу, что граница между морфологией и синтаксисом является нечеткой и, возможно, представляет собой континуум (Morphology-Syntax Continuum) без явных кластеров [23, p. 31–80].

Развивая эту идею, другие исследователи, например Адам Толман (Adam Jr. Tallman), подчеркивают необходимость выходить за рамки бинарного противопоставления «грамматического» и «фонологического» слова и использовать многофакторные

[6] Напр., Flughafensicherheitskontrolle «служба безопасности аэропорта» и др.

модели (multi-factor model), в которых несовпадение разных критериев рассматривается не как исключение, а как системная характеристика языка: «многофакторная модель выделения слова, в которой статус слова (wordhood) является кластерным понятием, определяемым совокупностью типологически независимых признаков» [30, p. 297]. Основной вклад Толмана заключается в том, что он систематизирует множество независимых критериев (мобильность, непрерываемость, просодическая целостность, автономное употребление и т.д.), демонстрируя, что они редко совпадают. Эти критерии являются не просто инструментами для выделения слова, а типологически независимыми признаками, каждый из которых вносит свой вклад в полноту «словности» определенной единицы. Это позволяет отойти от поиска идеальной, универсальной единицы и сосредоточиться на том, как эти разные свойства распределяются в языках мира, образуя пучки свойств (bundles of properties) вокруг концепта слова. Толман приходит к замечательному выводу: «объект с неопределенными границами является нормой для языков, а идеально определяемое слово – исключением» [30, p. 317].

Таким образом, теоретическая лингвистика первой четверти XXI века подтверждает и независимо развивает те выводы, к которым интуитивно пришли советские синологи на материале китайского языка: «слово» – это не дискретная единица, а, скорее, «пучок признаков», узел на пересечении различных уровней, и китайский язык является одним из наиболее ярких системных примеров этой общей закономерности.

С развитием цифровых технологий и применением методов обработки естественного языка (NLP), корпусной лингвистики и ИИ, вопрос о слове в китайском языке получил практическое измерение: автоматизация и создание алгоритмов требуют чётких критериев для идентификации слова без пробелов на письме и при высокой вариативности конструкции. В связи с этим представленные выше теории о необходимости поиска новых, речевых критериев, обрели новый смысл, получив развитие в современных прагматических подходах к обработке и описанию китайского языка.

Критерии «словности» в контексте современных методов сегментирования

В последние два десятилетия вопрос о слове в китайском языке переместился из области теоретической типологии в конкурирующую среду систем обработки естественного языка (NLP), корпусной лингвистики и ИИ-моделирования. Практические задачи (создание словарей, разметка корпусов, автоматическая сегментация текста) продолжают стимулировать выработку рабочих определений слова. Отсутствие пробелов на письме вынуждает специалистов задавать критерии «словности» явно, а масштабные корпуса и предобученные модели предоставили инструменты для проверки этих критериев в автоматизированной обработке. Большинство современных лингвистов сходятся на том, что понятие слова в китайском необходимо для описания языка – хотя бы в лексикографическом и когнитивном плане [21; 33; 19].

В начале 2000-х годов сообщество разработчиков систем автоматической обработки китайского языка столкнулось с необходимостью унификации базовых представлений о том, что считать словом. Эта проблема стала особенно очевидной в рамках серии соревнований SIGHAN Bakeoff (Special Interest Group for Chinese Language Processing, Association for Computational Linguistics), где различные исследовательские группы сопоставляли результаты алгоритмов сегментации китайского текста. Несмотря на общую задачу – выделение словных границ, – участники использовали различные корпусные стандарты, каждый из которых базировался на собственных лингвистических предпосылках и правилах аннотации. Различия между этими стандартами касались, прежде всего, того, как трактовать сложные числительные и сочетания с классификаторами (например, 一只鸟 yī zhī niǎo *«одна-CLF-птица»*): для одних корпусов это одно слово, для других – три отдельных элемента. Ещё одно расхождение касалось лексикализованных сочетаний, прежде всего типичных конструкций типа глагол-объект (VO), таких как 睡

觉 shuìjiào *«спать»*, 吃饭 chīfàn *«есть»*. Некоторые корпуса аннотировали их как единые слова, другие – как два самостоятельных компонента [14, с. 40-43]. Аналогичные расхождения наблюдались в трактовке именных композитов и имён собственных. Очевидно, что модель, обученная на корпусе, использующем один стандарт словных границ, демонстрировала заметное падение точности при применении к данным, размеченным по другому стандарту. В теоретическом осмыслении это означало, что «словность» в китайском языке проявляется не как универсальное свойство языка, а как результат конкретного операционального решения – того, как исследовательская группа определяет границы слов в своём корпусе.

На следующем этапе определённую доминанту задали такие корпуса, как Chinese Treebank (CTB) и, особенно, Universal Dependencies (UD). Эти ресурсы закрепили синтаксический подход, согласно которому предпочтение отдаётся синтаксическим словам [33]. Даже устойчивые лексемы типа VO (睡觉 shuìjiào *«спать»*) разбиваются на глагол и зависимое имя. Такая стратегия облегчает синтаксическое аннотирование, поскольку делает структуру предложения более прозрачной, однако она вступает в противоречие с интуицией лексемной целостности часто употребляемых композитов. В инженерной системе корпусов именно эта «тонкая» сегментация стала де-факто стандартом. Она удобна для синтаксического анализа структур и совместима с универсальными аннотационными схемами, но не всегда отражает лексикализованные единицы и устойчивые выражения. Как отмечают в своей статье Xue и др., разработчики *The Penn Chinese TreeBank*: «Лингвистические основания для определения того, какой должна быть правильная сегментация, довольно условны. «Правильная» сегментация может быть определена только в контексте конкретного применения. Для лексических композитов, имеющих лингвистически обоснованные внутренние структуры, наш подход заключается в присвоении им иерархической структуры. Например, результативный глагольный композит типа 走 zǒu / 过来 guòlái «идти через» мы рассматриваем как состоящий из двух сегментов, но на этапе расстановки скобок мы присваиваем метку глагольному композиту в целом. Такой подход позволяет пользователю выбирать желаемый уровень детализации» [33, p. 219].

В результате на практике сосуществуют конкурирующие стандарты – лексико- и синтаксико-ориентированные, а практическая универсальность применения теоретических инструментов остаётся по-прежнему ограниченной.

С точки зрения типологии слова, опыт корпусного определения границ слова выявляет два важных фактора: во-первых, становится очевидно, что сегментация – это не нейтральный/линейный процесс обработки текста, а конкретный теоретический выбор; во-вторых, разногласия стандартов говорят не о недостатке данных, а отражают многомерность понятия слова как такового, когда разные критерии «подсвечивают» разные грани одного феномена. *Этот практический вызов, по сути, стал эмпирической верификацией теоретических положений о градуальности: именно потому, что слово в китайском – это «многомерный узел признаков», и стало возможным сосуществование нескольких лингвистически обоснованных, но несовместимых стандартов сегментации. Алгоритмы не могут найти «истинное» слово, поскольку его критерии расходятся, что и предсказывали теоретики.*

На этапе подключения нейронных моделей к проблеме сегментации (Chinese Word Segmentation, CWS) эта задача решается путем последовательной разметки: каждому иероглифу приписывается метка, обозначающая его положение в слове: начало (B), внутренняя позиция (I), конец (E) или одиночное слово (S). Таким образом, система должна определить, где именно проходят словные границы в непрерывной последовательности символов. С начала 2010-х годов стандартом стали нейронные модели, основанные на архитектуре BiLSTM (bidirectional Long Short-Term Memory) с условными случайными полями CRF (conditional random field), а позднее – на трансформерах, таких как

BERT, также с CRF-слоем на выходе. Эти модели обучаются на больших размеченных корпусах и предсказывают вероятностное распределение границ слов. Идея мультикритериального (multi-criteria) обучения состоит в том, чтобы тренировать одну базовую модель одновременно на нескольких корпусах, каждый из которых следует своему стандарту [32; 22; 27]. В архитектуре таких моделей выделяют общий слой, отвечающий за извлечение инвариантных признаков, то есть таких свойств последовательности знаков, которые устойчивы при смене стандарта (например, частые биграммы, морфологические шаблоны, типичные синтаксические контексты). Поверх него располагаются специализированные слои, адаптированные под конкретный стандарт. В результате образуется модель, способная переключаться между наборами данных и сохранять разумную точность даже на новых схемах аннотации. Алгоритм, обученный на множестве стандартов, *по сути, учится распознавать* устойчивые структурные сигналы, *которые не зависят от* смены аннотационного критерия (например, двусложные комбинации с фиксированным порядком следования компонентов и высокой частотностью совместного употребления).

В типологическом смысле такой опыт нейронной обработки текста указывает на то, что признаки «словности» в китайском языке обладают градуальной природой: они не располагаются линейно на оси сегментации, а распределены спектрально.

Скажем здесь несколько подробнее о предобученных языковых моделях для китайского, таких как Chinese BERT (*Bidirectional Encoder Representations from Transformers*). Они работают на уровне иероглифов или субсловных единиц: каждый иероглиф рассматривается как самостоятельный токен словаря. При этом такие модели демонстрируют высокие результаты в самых разных задачах даже без явного введения понятия слова. Однако модификация процедуры предобучения под названием *Whole Word Masking* (WWM) показала, что включение словных индикаторов способно повысить качество модели. При WWM система не маскирует случайные отдельные иероглифы, а заменяет маской все символы, составляющие одно слово согласно внешнему сегментатору. Таким образом, модель учится предсказывать сразу целую словную единицу, а не её фрагменты. Эмпирически оказалось, что такая стратегия повышает точность в задачах распознавания имён собственных, ответов на вопросы и даже самой сегментации [19, p. 6–7]. С теоретической точки зрения это означает, что добавление словного слоя как индуктивной подсказки способствует лучшему усвоению распределительных и морфосемантических связей между иероглифами.

В то же время универсальные токенайзеры на основе субсловных алгоритмов (SentencePiece, BPE, Unigram) продолжают демонстрировать сильные результаты и без опоры на явные словные границы [29, p. 70]. Это показывает, что «слово» в классической трактовке не является единственным путём к высоким метрикам, однако WWM остаётся убедительным эмпирическим свидетельством того, что добавление словной информации помогает моделям улавливать те *лексико-семантические и дистрибутивные связи, которые участвуют в формировании лексемы* в языке.

В заключение

Безусловно, опыт последних двух десятилетий в сфере NLP не отменяет классических споров о слове в китайском. Он *не столько «доказывает» ту или иную теорию, сколько* операционализирует теоретические наработки и *предоставляет эмпирические аргументы* в пользу тех положений, верификация которых была ранее невозможна на доступных объёмах печатных текстов. Тот факт, что сегментация является «практической задачей», зависящей от цели, не противоречит этому выводу, а, напротив, подтверждает его: сама необходимость выбирать между разными (лексическими, синтаксическими) стандартами разметки является практическим отражением теоретической многомерности китайского слова.

Конфликт стандартов сегментации, успехи мультикритериального обучения, эффективность Whole Word Masking и психолингвистические данные о восприятии речевой цепи носителями сходятся в одном: **слово в китайском – не объект ступенчатой иерархии, а многоуровневый узел признаков**. Именно такую картину предвосхитили Поливанов, Рождественский, Коротков, Солнцев, настаивая на кросс-уровневой форме, учете спектра признаков и градиентной словности. Говоря на языке современной лингвистики, их практические рекомендации для сегодняшних систем могли бы быть условно сформулированы так: работать **со слоями (лексемным, морфосинтаксическим, семантическим), оценивать словность непрерывно и учиться** сразу на нескольких стандартах, подкрепляя решения диахроническими данными и когнитивными экспериментами, а в перспективе – и данными автоматической сегментации устной речи, что позволило бы верифицировать и просодические критерии, на важность которых указывали Поливанов и Солнцев. Таким образом, «возврат к истокам», к комплексному анализу природы китайского слова синологами-предшественниками, несмотря на свою кажущуюся консервативность, представляется продуктивным инструментом: теоретическая интуиция классиков становится схемой данных, архитектурой модели и протоколом эксперимента.

В своё время Н. В. Солнцева отмечала: «существующие типологические понятия являются тем стандартом, сопоставление с которым позволяет дать ту или иную оценку явлению или явлениям изолирующих языков. В то же время особенности изолирующих языков требуют уточнения в некоторых случаях и пересмотра самого этого стандарта. Иначе говоря, типологические стандарты, сложившиеся в условиях неравномерной изученности разносистемных языков, выполняя свою роль как меры сопоставления этих языков, должны видоизменяться и уточняться по мере выравнивания степени изученности сопоставляемых языков» [13, с. 7]. Очевидно, что сегодняшние разработки в области NLP, разметки корпусов, лингвистического анализа с применением ИИ-моделей в значительное мере способствуют выравниванию степени изученности разносистемных языков, внося существенные корректировки в принятые типологические стандарты.

На материале данного обзора мы стремились показать, что теоретическое наследие советских китаистов-языковедов 1930–1990-х годов по-прежнему актуально своей концептуальной смелостью и поразительной научной интуицией. Работы Е. Д. Поливанова, Н. Н. Короткова, И. М. Ошанина, Ю. В. Рождественского, А. А. Драгунова, В. М. и Н. В. Солнцевых, О. М. Готлиба, созданные в докомпьютерную эпоху, на основе скрупулезного анализа языкового материала, заложили основы неклассического, динамического понимания природы китайского слова. Они первыми системно описали его как единицу не столько морфологически монолитную, сколько функционально и семантически целостную, границы которой могут быть проницаемы и контекстуально зависимы.

Надеемся, что представленный анализ спектра теоретических подходов не только поможет начинающим и продолжающим лингвистам сориентироваться в калейдоскопе разрозненных воззрений на устройство китайского языка, но и подскажет пути дальнейших исследований с опорой на вдумчивое и внимательное прочтение работ выдающихся ученых отечественной синологической школы.

Библиография

1. Алпатов В. М. Е. Д. Поливанов о китайском и японском языках // Письменные памятники Востока. 2021. Т. 18. № 3. С. 179–186. DOI: 10.17816/WMO77362
2. Иванов А. И., Поливанов Е. Д. Грамматика современного китайского языка. М.: Инт востоковед., 1930. 304 с.
3. Кисель П. О. К вопросу о «пустом дополнении» в раздельно-слитных словах лихэцы глагольно-именной модели в современном китайском языке // Вестник

РГГУ. Серия «Литературоведение. Языкознание. Культурология». 2021. № 2. С. 32–63. DOI: 10.28995/2686-7249-2021-2-32-63
4. Кисель П. О. Раздельно-слитные слова лихэцы глагольно-именной модели в современном китайском языке: дис. ... канд. филол. наук. СПб., 2023.
5. Кисель П. О. Раздельно-слитные слова лихэцы в современном китайском языке: к вопросу уровневой принадлежности // Известия Восточного института. 2023. № 2. С. 138–147. DOI: 10.24866/2542-1611/2023-2/138-147.
6. Коротков Н. Н. К проблеме морфологической характеристики китайского литературного языка. – Труды XXV Международного конгресса востоковедов. Т. V. М., 1963., стр. 101–108.
7. Корпус китайского языка Пекинского университета [Электронный ресурс]. URL: http://ccl.pku.edu.cn:8080/ccl_corpus/ (дата обращения: 17.10.2025).
8. Рождественский Ю. В. Понятие формы слова в истории грамматики китайского языка: Очерки по истории китаеведения / Ин-т междунар. отношений. – Москва: Изд-во ИМО, 1958.
9. Рождественский Ю. В. О границах сложного слова в китайском языке. – Труды XXV Международного конгресса востоковедов. Т. V. М., 1963, стр. 176–181.
10. Рождественский Ю. В., Коротков Н. Н., Солнцев В. М., Сердюченко Г.П. Китайский язык. М.: Наука, 1969.
11. Солнцев В. М. Типологические свойства изолирующих языков (на материале китайского и вьетнамского языков) // Языки Юго-Восточной Азии. Проблемы морфологии, фонетики и фонологии. – М.,1970. С. 11–19.
12. Солнцев В. М. Введение в теорию изолирующих языков в связи с общими особенностями человеческого языка. – М.: Восточная литература, 1995.
13. Солнцева Н. В. Проблемы морфологии изолирующих языков в типологическом освещении: диссертация ... доктора филологических наук: 10.02.20. – Москва, 1984. – 419 с.
14. Huang Chu-Ren, Hsieh Shu-Kai, and Chen Keh-Jiann. Mandarin Chinese Words and Parts of Speech: A corpus-based study. London: Routledge. 2017.
15. Arcodia G.F. Lexical Derivation in Mandarin Chinese. Wenhe chuban youxian gongsi (文鹤出版有限公司), 2012.
16. Arcodia G.F., Basciano B. Morphology and the lexicon // Chinese Linguistics: An Introduction. Oxford Univ. Press, 2021.
17. Bisang W. Grammaticalization without coevolution of form and meaning. In: Bisang, Himmelmann & Wiemer (eds.) What makes grammaticalization? B. & W. Bisang, 2004.
18. Bybee J.L. Morphology: A Study of the Relation between Meaning and Form. Amsterdam; Philadelphia: John Benjamins Publishing Company, 1985.
19. Cui, Y. et al. 2021. Pre-Training with Whole Word Masking for Chinese BERT. IEEE/ACM TASLP, Nov. 2021. https://arxiv.org/pdf/1906.08101
20. Eifring H. The Chinese Word Reconsidered. Oslo: University of Oslo, 2013.
21. Emerson, T. 2005. The Second International Chinese Word Segmentation Bakeoff. In: Proc. 4th SIGHAN Workshop, pp. 123–133.
22. Gong, J. et al. 2019. Switch-LSTMs for Multi-Criteria Chinese Word Segmentation. In: Proc. 33rd AAAI Conference on Artificial Intelligence (AAAI-19), pp. 6457–6464.
23. Haspelmath, Martin. 2011. The indeterminacy of word segmentation and the nature of morphology and syntax. Folia linguistica 45(1). 31–80. DOI: 10.1515/flin.2011.002
24. Myers J. Wordhood and Disyllabicity in Chinese. In: Huang C-R, Lin Y-H, Chen I-H, Hsu Y-Y, eds. The Cambridge Handbook of Chinese Linguistics. Cambridge University Press, 2022: 47–73.
25. Naumann B., Vogel I. Prosodic Word Formation: Derivation and Lexicon. 2000.

26. Packard, J. L. The Morphology of Chinese: A Linguistic and Cognitive Approach. Cambridge University Press, 2000.
27. Qiu, X. et al. A Concise Model for Multi-Criteria Chinese Word Segmentation. Findings of EMNLP. 2020.
28. SIGHAN Bakeoff Reports (Special Interest Group for Chinese Language Processing, ACL) [Электронный ресурс].
29. Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Brussels, Belgium. Association for Computational Linguistics. pp. 66–71, https://doi.org/10.18653/v1/D18-2012, https://aclanthology.org/D18-2012.pdf
30. Tallman, Adam JR. 2020. Beyond grammatical and phonological words. Language and Linguistics Compass 14(2). e12364. DOI: 10.1111/lnc3.12364
31. Xia, F. 2000. The Segmentation Guidelines for the Penn Chinese Treebank (3.0). IRCS Tech. Report 00-06, Univ. of Pennsylvania: https://repository.upenn.edu/handle/20.500.14332/37641
32. Xinchi Chen, Zhan Shi, Xipeng Qiu, and Xuanjing Huang Adversarial Multi-Criteria Learning for Chinese Word Segmentation. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), . 2017, pp. 1193–1203, Vancouver, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-1110
33. Xue, N., Xia, F., Chiou, F-D., Palmer, M. 2005. The Penn Chinese Treebank: Phrase structure annotation of a large corpus. Natural Language Engineering. 11(2): 207–238. doi:10.1017/S135132490400364X
34. Zhang, Q. et al. Neural Networks Incorporating Dictionaries for Chinese Word Segmentation. 32nd AAAI Conference on Artificial Intelligence (AAAI-18). 2018. https://cdn.aaai.org/ojs/11959/11959-13-15487-1-2-20201228.pdf

References

1. Alpatov, V. M. (2021). E. D. Polivanov on Chinese and Japanese Languages. *Pis'mennye pamyatniki Vostoka* (Written Monuments of the Orient), *18*(3), 179–186. DOI: 10.17816/WMO77362
2. Ivanov, A. I., & Polivanov, E. D. (1930). *Grammar of the Modern Chinese Language*. Moscow: Institute of Oriental Studies Press.
3. Kisel, P. O. (2021). On Case of the "Empty Object" in Separable Words *liheci* of the Verb-Noun Model in Modern Chinese. *Vestnik RGGU. Seriya «Literaturovedenie. Yazykoznanie. Kul'turologiya»* (RSUH/RGGU Bulletin. "Literary Theory. Linguistics. Culturology" Series), (2), 32–63. DOI: 10.28995/2686-7249-2021-2-32-63
4. Kisel, P. O. (2023). *Separable Words 'liheci' of the Verb-Noun Model in Modern Chinese* [PhD Dissertation]. St. Petersburg.
5. Kisel, P. O. (2023). Separable Words *liheci* in Modern Chinese: On the Question of Level Affiliation. *Izvestiya Vostochnogo instituta* (Oriental Institute Journal), (2), 138–147. DOI: 10.24866/2542-1611/2023-2/138-147
6. Korotkov, N. N. (1963). On the Problem of the Morphological Characteristics of the Chinese Literary Language. In *Trudy XXV Mezhdunarodnogo kongressa vostokovedov* (Proceedings of the 25th International Congress of Orientalists) (Vol. V, pp. 101–108). Moscow.
7. Peking University Center for Chinese Linguistics. (n.d.). *CCL Corpus*. Retrieved October 17, 2025, from http://ccl.pku.edu.cn:8080/ccl_corpus/

8. Rozhdestvensky, Yu. V. (1958). *The Concept of Word Form in the History of Chinese Grammar: Essays on the History of Sinology*. Moscow: IMO Publishing House.
9. Rozhdestvensky, Yu. V. (1963). On the Boundaries of the Compound Word in the Chinese Language. In *Trudy XXV Mezhdunarodnogo kongressa vostokovedov* (Proceedings of the 25th International Congress of Orientalists) (Vol. V, pp. 176–181). Moscow.
10. Rozhdestvensky, Yu. V., Korotkov, N. N., Solntsev, V. M., & Serdyuchenko, G. P. (1969). *The Chinese Language*. Moscow: Nauka.
11. Solntsev, V. M. (1970). Typological Features of Isolating Languages (on the data of Chinese and Vietnamese languages). In *Yazyki Yugo-Vostochnoy Azii. Problemy morfologii, fonetiki i fonologii* (Languages of Southeast Asia. Problems of Morphology, Phonetics, and Phonology) (pp. 11–19). Moscow.
12. Solntsev, V. M. (1995). *Introduction to the Theory of Isolating Languages in Connection with the General Features of Human Language*. Moscow: Vostochnaya literatura.
13. Solntseva, N. V. (1984). *Problems of Morphology of Isolating Languages in a Typological Perspective* [Doctoral Dissertation]. Moscow.
14. Huang Chu-Ren, Hsieh Shu-Kai, and Chen Keh-Jiann. Mandarin Chinese Words and Parts of Speech: A corpus-based study. London: Routledge. 2017.
15. Arcodia G.F. Lexical Derivation in Mandarin Chinese. Wen he chu ban you xian gong si, 2012.
16. Arcodia G.F., Basciano B. Morphology and the lexicon // Chinese Linguistics: An Introduction. Oxford Univ. Press, 2021.
17. Bisang W. Grammaticalization without coevolution of form and meaning. In: Bisang, Himmelmann & Wiemer (eds.) What makes grammaticalization? B. & W. Bisang, 2004.
18. Bybee J.L. Morphology: A Study of the Relation between Meaning and Form. Amsterdam; Philadelphia: John Benjamins Publishing Company, 1985.
19. Cui, Y. et al. 2021. Pre-Training with Whole Word Masking for Chinese BERT. IEEE/ACM TASLP, Nov. 2021. https://arxiv.org/pdf/1906.08101
20. Eifring H. The Chinese Word Reconsidered. Oslo: University of Oslo, 2013.
21. Emerson, T. 2005. The Second International Chinese Word Segmentation Bakeoff. In: Proc. 4th SIGHAN Workshop, pp. 123–133.
22. Gong, J. et al. 2019. Switch-LSTMs for Multi-Criteria Chinese Word Segmentation. In: Proc. 33rd AAAI Conference on Artificial Intelligence (AAAI-19), pp. 6457–6464.
23. Haspelmath, Martin. 2011. The indeterminacy of word segmentation and the nature of morphology and syntax. Folia linguistica 45(1). 31–80. DOI: 10.1515/flin.2011.002
24. Myers J. Wordhood and Disyllabicity in Chinese. In: Huang C-R, Lin Y-H, Chen I-H, Hsu Y-Y, eds. The Cambridge Handbook of Chinese Linguistics. Cambridge University Press, 2022: 47–73.
25. Naumann B., Vogel I. Prosodic Word Formation: Derivation and Lexicon. 2000.
26. Packard, J. L. The Morphology of Chinese: A Linguistic and Cognitive Approach. Cambridge University Press, 2000.
27. Qiu, X. et al. A Concise Model for Multi-Criteria Chinese Word Segmentation. Findings of EMNLP. 2020.
28. SIGHAN Bakeoff Reports (Special Interest Group for Chinese Language Processing, ACL) [Электронный ресурс].
29. Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Brussels, Belgium. Association for Computational Linguistics. pp. 66–71, https://doi.org/10.18653/v1/D18-2012, https://aclanthology.org/D18-2012.pdf
30. Tallman, Adam JR. 2020. Beyond grammatical and phonological words. Language and Linguistics Compass 14(2). e12364. DOI: 10.1111/lnc3.12364

31. Xia, F. 2000. The Segmentation Guidelines for the Penn Chinese Treebank (3.0). IRCS Tech. Report 00-06, Univ. of Pennsylvania: https://repository.upenn.edu/handle/20.500.14332/37641
32. Xinchi Chen, Zhan Shi, Xipeng Qiu, and Xuanjing Huang Adversarial Multi-Criteria Learning for Chinese Word Segmentation. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), . 2017, pp. 1193–1203, Vancouver, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-1110
33. Xue, N., Xia, F., Chiou, F-D., Palmer, M. 2005. The Penn Chinese Treebank: Phrase structure annotation of a large corpus. Natural Language Engineering. 11(2): 207–238. doi:10.1017/S135132490400364X
34. Zhang, Q. et al. Neural Networks Incorporating Dictionaries for Chinese Word Segmentation. 32nd AAAI Conference on Artificial Intelligence (AAAI-18). 2018. https://cdn.aaai.org/ojs/11959/11959-13-15487-1-2-20201228.pdf