例句质量评估体系构建及大语言模型例句生成能力评估

方明炜[1,2]，朱君辉[1,2]，鲁鹿鸣[1,2]，杨尔弘[1,2]*，杨麟儿[1,2]
[1]北京语言大学国家语言资源监测与研究平面媒体中心，北京
[2]北京语言大学信息科学学院，北京
mingweifang808@gmail.com，nysyzxzjh@gmail.com，llm410402@gmail.com
yerhong@blcu.edu.cn，lineryang@gmail.com

摘要

本研究针对大语言模型（LLMs）生成例句的教学适用性问题，基于二语习得认知理论构建了多维例句质量评估体系，涵盖规范性、语境独立性、典型度、词汇适切性及句法复杂度五大核心维度。通过采集汉语词典与教材的优质例句作为基准语料，结合特征工程构建了机器学习模型（准确率为98.6%），验证了评估框架的有效性。在此基础上，本研究利用该评估框架对LLMs生成例句与传统人工编纂词典中的例句进行了系统对比分析。研究结果表明：LLMs在语法典型度、词汇难度、汉字笔画数方面展现出与传统词典例句相当的质量水平，而在语境独立性、语义典型度、词汇常用度方面仍存在一定不足。进一步研究发现，不同提示策略影响例句生成质量，其中融合语言特征约束型提示策略优化效果最佳。本研究首次实现LLMs生成例句教育适应性的量化评估，为智能语言教辅系统开发提供了兼具理论指导意义与实践应用价值的评估范式。

关键词： 例句；语料库；特征提取；机器学习；大语言模型

Construction of Example Sentence Quality Evaluation System and Evaluation of Example Sentence Generation Ability in Large Language Models

Mingwei Fang[1,2], Junhui Zhu[1,2], Luming Lu[1,2] , Erhong Yang[1,2]*, Liner Yang[1]
[1]National Language Resources Monitoring and Research Center for Print Media, Beijing Language and Culture University, Beijing
[2]School of Information Science, Beijing Language and Culture University, Beijing
mingweifang808@gmail.com，nysyzxzjh@gmail.com，llm410402@gmail.com
yerhong@blcu.edu.cn，lineryang@gmail.com

Abstract

This study addresses the pedagogical applicability of example sentences generated by large language models (LLMs) and proposes a multidimensional evaluation framework for sentence quality based on cognitive theories of second language acquisition. The framework encompasses five core dimensions: accuracy, contextual independence, typicality, vocabulary control and syntactic complexity. High-quality example sentences from Chinese dictionaries and textbooks were collected as benchmark corpora, and a machine learning model was constructed using feature engineering, achieving an accuracy of 98.6%, thereby validating the effectiveness of the proposed evaluation system. Based on this framework, the study conducted a systematic comparative analysis between LLM-generated Chinese example sentences and those compiled in traditional dictionaries. The results indicate that LLMs perform comparably to human-compiled examples in terms of syntactic typicality, lexical difficulty, and character stroke count, but still show deficiencies in contextual independence, semantic typicality, and lexical

* 通讯作者

第二十四届中国计算语言学大会论文集，第117页-第130页，济南，中国，2025年8月11日至14日。

frequency. Further research has found that different prompting strategies affect the quality of example sentence generation, with the fusion language feature constrained prompting strategy showing the best optimization effect. This study is the first to achieve a quantitative evaluation of the educational adaptability of LLMs in generating example sentences, providing an evaluation paradigm that combines theoretical guidance and practical application value for the development of intelligent language teaching assistance systems.

Keywords: Example sentence , corpus , feature extraction , machine learning , Large Language Models

1 引言

在二语习得过程中，教学例句作为知识载体需同时传递词汇的语义网络、句法规则及语用边界。随着汉语国际传播的深入推进以及数字化语言资源的不断丰富，语言教学对高质量例句语料的依赖日益增强。然而，当前汉语教学语料的例句呈现仍面临多重瓶颈：其一，部分例句词汇难度偏高，未能准确体现词目核心语义与典型用法；其二，例句数量有限，难以满足多样化教学与学习需求；其三，例句编写仍较多依赖传统纸质媒介，影响了其传播范围和适应性 (杨玉玲and 段彤彤, 2023)。以《当代汉语学习词典》[0]为例，可以看出，这些例句所用词语的难度显然高于词目的难度。

包：骑兵从两翼包过去。

办：办签证。行李要办托运。办完登机手续。贷款办不下来。上午跟他办了交接。

在语料库技术普及之前，教学例句多由编写者基于经验或直觉人工创设。此过程不仅耗费大量人力与时间资源，且对编写者的语言学素养及编写经验构成严峻挑战。如Okeeffe(2007)等学者指出，此类例句很大部分“是基于我们如何使用语言的直觉，而不是使用语言的实际证据”，导致其主观性偏高。相比而言，英语等语种的学习词典编纂已广泛应用真实语料库与计算方法支持例句生成与筛选，而汉语学习词典起步较晚，仍易受传统“重释义、轻示例”编纂观念影响，造成例句质量参差不齐、教学辅助功能不足。

近年来，自然语言处理技术的长足发展，特别是大型语言模型（Large Language Models, LLMs）的兴起，为二语学习任务的自动化（如试题生成、释义生成等）开辟了广阔前景。LLMs 在通用文本生成中表现出色，有望缓解教学例句编写中存在的高成本、高主观性问题，提升例句资源的可获取性与适应性。然而，尽管LLMs在通用文本生成任务中展现出卓越性能，其在教学例句生成等特定语言教育场景下仍面临诸多挑战。因此，亟需开展系统、深入的质量评估研究，以衡量并提升其在实际教学中的适用价值。

尽管如此，当前针对汉语自动例句质量评估的研究仍处于起步阶段，尚缺乏面向汉语语言特性和教学需求的系统性评估框架。考虑到汉语在语法结构、词汇搭配与教学规律等方面的独特性，构建一套可量化、有效且具备可推广性的自动评估体系，对于提升汉语教学效率与词典编纂水平具有重要意义。

基于上述背景，本文聚焦于大语言模型生成教学例句的质量评估问题，提出一套多维度、结构化的评估框架，涵盖规范性、语境独立性、典型度、词汇适切性与句法复杂度五大核心维度。本文通过构建高质量实验语料集，结合特征工程与机器学习建模，系统验证评估体系在例句筛选任务中的可行性与有效性，并在此基础上对比分析LLMs与传统词典例句在多个维度上的表现差异。研究旨在为LLMs在汉语教学中的应用提供理论支持与实证依据，同时为例句质量评估与教学资源建设提供实践参考。

2 相关研究

例句的质量与例句应具备的相关特征紧密相连。英语相较于中文作为第二语言进行系统性教学的时间跨度更长，已经积累了较多宝贵的经验和认识。VanPatten (2004)从语言输入的角

国家语委重大项目：大语言模型的评测技术和方法研究；项目号：ZDA145-17
[0]张志毅.当代汉语学习词典.北京:商务印书馆,2020.

度指出，例句应注意满足学习者关于实际语义的需求；Atkins和Rundell (2008)在《牛津词典编纂指南》中对于词典例句的标准进行了全面的探讨，他们认为，一个良好的例句应当兼具典型性、信息丰富性和可读性。Harras (1989)提到好例句应满足的四个基本标准包括：例句应体现目标词所代表的对象的典型特征、应呈现与目标词通常同时出现的单词、应是真实的、并且包含与目标词语义上相关的词。而关于汉语学习词典例句的编写，已有研究从配例原则、配例功能、学习者的认知规律和实际汉语教学经验等角度出发，为例句的选用和编写原则提出了一些可供参考的建议，且各有侧重。如孙全洲 (1986)、徐玉敏等 (2005)、张 (2005)、鲁健骥等 (2006)、李禄兴 (2006)、刘川平 (2006)、李泉等（2008） (2008)、徐茗 (2009)、蔡永强 (2011)、刘若云 (2012)等,主要提出的编写原则有科学性、规范性、典型性、可学性、语境自明性、实用性、趣味性等。

随着汉语热在全球不断升温，与之相关的汉语教学问题越来越受到关注，相关研究者对汉语例句的编排与呈现如何凸显这些特征做出了一些有益的尝试。如臧宁 (2022)提出量词释义和例句的编写要增加语义特征说明,主要需显示词条的搭配关系、提供词条的语境信息、体现词条的语用文化，以便于学习者理解。李淑云 (2023)发现现有的同义词词典配例存在搭配不够典型等问题，提出可通过借助语料库考察同义词的典型搭配、多元序列、类联结等方式选择合适配例。郭小娜 (2023)指出应完善词典中的提示信息、注重搭配对象的频率、选取配例时考虑用词难度，针对易混淆词,设置辨析栏、设置反例、提供完整语境信息等。杨玉玲等 (2023)对《当代》的配例用词进行了定量统计，发现在不同难度等级的被释词上，配例用词的难度控制意识不强，基于定量统计结果，提出了外向型汉语学习词典配例难度控制的¡=i原则，也就是配例用词的难度不应高于被释词的难度。

这些研究从不同角度对例句进行了特征分析和编排建议，然而仍旧缺乏较为系统的定量分析。要将例句选用和编写经验应用于例句自动评估，需要将这些经验和理论进行系统化和量化。其他语言上已有相关量化指标的研究以供中文借鉴。Segler (2007)的研究旨在为德语二语阅读课文中的疑难单词选择合适的例句，从(a)降低句法复杂性，(b)句子相似性，提供(c)重要的共现词和(d)语义相关词这四个方面出发构建例句排名模型，评估结果表明例句的排名与教师和学生评价的高低相符。现有应用于Sketch Engine的例句评估工具GDEX (Husák, 2010; Kilgarriff et al., 2008)，能够自动从语料库中检索好的词典例句。GDEX最初是为英语学习者词典《麦克米伦英语词典》的电子版添加例句设计的，通过分析包含目标索引的所有语料库数据并根据特定的指标进行评分，最终形成包含例句分数的列表。评分模型是一个基于规则的公式，对候选句的句长、词频、照应词、目标词位置、标点符号等特征赋予不同的权重。一些其他语言也相继推出了GDEX评分模型。例如，Kosem等 (2011)分别测试了不同的特征在斯洛文尼亚语例句评分模型构建中的效果，发现句长、句子中相对关键字位置、关键字是否重复、单词长度是否过长和句子长度是否过长在分类中发挥了更大的作用，并最终选择效果最优的分类器作为斯洛文尼亚语的GDEX模型并在Sketch Engine中实现。Srdanović I和I.Kosem (2016)根据日语学习和词典编纂需求提出了日语的GDEX评分模型，纳入的特征包括标点符号出现次数、句长、非法字符黑名单、词长、专有名词、搭配典型度等，还将日语能力测试词表作为词汇难度的参考。这些基于计算机的研究为汉语例句质量自动评估提供了可供参考的思路和路径，因此，本文将进一步探索中文例句的量化和评价方法。

3 例句质量评估体系构建

例句质量评估的意义在于从大规模中文语料库中选择质量良好的例句来辅助词汇的学习。为此，需要对“质量良好”这一范畴进行界定，并将其转化为可操作的标准，从而应用于实际的语料提取中。本文对二语教学中例句相关理论和例句语料进行了广泛研究和调研，认为“质量良好”的例句除了其中应出现学习者检索的内容之外，还应满足以下几个条件。

（1）例句的结构完整，用语规范；
（2）例句的语义独立、完整，不依赖于上下文语境；
（3）例句应体现目标语言点的典型语义及用法；
（4）例句的可读性与学习者的水平相符。

其中第一、二项是对例句作为汉语教学载体的基本要求，第三项突出体现目标语言点在例句中的主体地位，第四项则将例句的呈现与学习者的能力层级紧密结合。基于此，本文构建了包括规范性、语境独立性、典型度、词汇适切性、句法复杂度在内的例句质量评估体系。

第二十四届中国计算语言学大会论文集，第117页-第130页，济南，中国，2025年8月11日至14日。

3.1 规范性(Normativeness)

规范性指例句在语法上应符合汉语普通话的语法规范，没有歧义、语病等现象；在语义上应合乎逻辑，思想积极向上，避免出现文化禁忌等内容。“内容规范”是对教材科学性的基本要求之一 (2008)，Kilgarriff (2008)等、Husák (2010)等指出语料库中好的候选句子应该在结构和词汇上都格式良好。只有在例句准确规范这一前提下，我们才能从更为广泛和纵深的方面评估例句的质量。

3.2 语境独立性(Context Independence)

语境独立性是指例句在脱离其原有上下文语境之后的可理解程度。即在没有更大的上下文的情况下，例句是否能向学习者提供一个独立的情境以及完整的语义，从而更好地促进学习者对词语的理解 (张, 2005)。所以优秀的例句就是对上下文依赖度低，语境相对自足的句子。但目前尚未有较为完整的计算方法。本文从“句间衔接”、“对话回应”、“直接引语”、“语气”和“回指”这五个方面对配例的语境独立性进行评判，其中“句间衔接”和“对话回应”通过关联词和应答词表识别，每出现一处符合的特征则进行相应的减分（当词目词恰好在词表中时，可以去除相应的限制）。“直接引语”和“语气”通过字符匹配识别，“回指”通过依存树位置识别。

句间衔接：判断句首是否出现具有衔接不同分句作用的关联词。关联词起到表示复句中分句间关系的作用，这些关系包括“假设关系”、“条件关系”、“因果关系”等，如果句首只出现“而且”、“都”、“所以”等关联词时，就会脱离原有语境，如例1：而且两人一般不在外面吃饭。

对话回应：表示对话回应的句子是对前一方所说的话的回应和衔接，在没有前文的情况下，缺乏语境独立性。我们通过句首是否存在“应答词”来判断语料中表示对话回应的句子。所谓“应答词”是一种功能词，用来表示听话人对说话人所说的话的一种回应。它们具有某些特殊的语用功能 (邵敬敏and 朱晓亚, 2005)。如“对”、“好啊”等。如例2：没错儿，是在我们这儿买的。

直接引语：判断句中是否包含直接引语。直接引语是指在句中引用他人话语，通常出现在引号内。这种引用必然涉及到对前文或特定人物的依赖，容易导致语境不完整。学习者在脱离具体语境时，往往无法理解引语的背景与含义。如例3：丁广泉笑着问：“你敢扎吗？”

语气：判断句中是否出现疑问句、反问句和感叹句。这种情况一般还需要对感叹、疑问或反问的的解释才能构成一个完整的语境。如例4：会不会是你吃了其它的东西?

回指：回指是重复指出上文出现过的人或事物，指示代词如“这”、“那”等，常常用来表示这种指称功能。如例5：这件事能不能成功全看你了。

3.3 典型度（Typicality）

典型度包括语法典型度和语义典型度两个特征。语法典型度特征探讨例句能否反映词语的典型用法，一般情况下可通过一个词语的常用搭配来体现。例句有反映搭配信息的作用，且由于汉语缺乏形态变化，词的搭配关系则更为重要。对于搭配信息的呈现，应提供较为典型或高频的搭配词语 (李禄兴, 2006; 李禄兴, 2015; 徐玉敏et al., 2005; 付娜, 2010)。要评估例句中的搭配是否典型，就需要对语料的搭配频次进行统计，本文参考了胡韧奋和肖航 (2019)在短语层面对八种搭配类型的界定，采用基于依存的方法抽取搭配，形成词典和教材语料的词表和搭配表，计算句子中各个词汇的搭配频率之和，同时参考Root TTR指标的计算方法，使用字数的平方根来减少文本长度对结果的影响。

语义典型度特征则表示例句能否体现词语的典型语义，对于表现语义的程度可以用整句的语义特征向量和词向量之间的关系来衡量。在语义典型度的量化上，本文采用BERT模型计算全句的语义CLS token和词语本身的词向量，计算两者之间的cosine相似度分数作为语义相关性，相关性越大，语义典型度得分越高。

3.4 词汇适切性（Vocabulary Control）

词汇适切性是指在特定教学或学习场景中，例句所使用的词汇与学习者的语言水平、认知能力以及学习目标相匹配的程度。通过控制例句中的词汇属性，可以有效辅助学习者对目标语言点的理解与掌握。本研究选择相对词汇难度、词汇常用度和汉字笔画数这三项特征，以量化例句的词汇适切性。

在相对词汇难度上，对外汉语教学中所教授的内容要循序渐进，例句的呈现同样要遵循从易到难、由浅入深的过程，且例句围绕词目服务，需要对例句的难度有所控制。对于目标词

目，我们计算例句中除词目外其他部分的难度高于词目的比例。令l表示词目的等级，相对词汇难度得分越高，表示例句相对于词目的难度越低，从而更适切学习者。

词汇常用度指词汇在文本中的出现频率，词汇出现的频率越高，人对其反应的时间就越短，就越容易被理解 (Forster and Chambers, 1973)。本文参照《现代汉语语料库词频表》，通过计算平均对数词例频数衡量例句的词汇常用度。平均对数词例频数越高，词汇常用度越高，例句的词汇适切性越强。

汉字笔画数作为构成汉字字形的最小单位数量，体现了汉字的视觉复杂程度。在汉语学习中，汉字的视觉复杂度是衡量词汇（特别是包含复杂汉字的词）学习难度的一个重要指标。我们统计了句中汉字的笔画总数，并除以使用词数的平方根，以减少文本长度对结果的潜在影响，从而获得基于汉字笔画数的词汇复杂度评价指标。

3.5 句法复杂度（Syntactic Complexity）

句法复杂度是从句子层面衡量句法结构的复杂程度。本节将从句子长度和依存句法分析两个维度探讨例句的句法复杂性。

句子长度是衡量句法复杂度的重要因素。过长或过短的句长均可能影响例句对词汇的教学效果，应将其控制在一个适当的范围内。现有研究对不同语言的例句长度进行了规范：英语模式的GDEX对于例句长度的限制为10到15个单词 (2021)；Kosem (2011)等通过WEKA工具确定斯洛文尼亚语首选句子长度为8-30个单词。对于中文例句的长度范围，《国际汉语教学通用课程大纲》中指出，中级教材文本的平均句长参考值为12.88-32.59字/句。本研究从字符个数的角度统计句长，通过对样本语料中级词例句的计算，得出平均句长为39.43字、句长中位数为34字、句长众数25字作为参考。我们将句长范围设置为5-15个字，超出或不足该范围的句子进行减分。

通过依存句法分析（Dependency Parsing）可以分析句子中词汇之间的相互依存关系，形成的依存句法关系树及其中的依存距离可以反映句子的句法复杂性。句法树越高，表示句法结构越复杂。而依存距离就是具有句法关系的词之间的线性距离，依存距离越长，表示句法关系越丰富。此外，认知语言学的相关研究表明，两个句法相关词之间的线性距离会影响记忆的存储，两个词语之间的距离越长，认知成本就会越大。主要动词前的最大词数可以反映这个认知距离。因此，我们选取主要平均句子依存距离和最大句子依存距离这两个特征分析例句的句法复杂性。最终构建的例句质量评估体系及计算方式如表1所示。

指标		描述	公式
规范性	规范性	在语义和语法上符合现代汉语普通话的规范	-
语境独立性	语境独立性	在脱离上下文后仍能够被理解的程度	构建关联词、回应词词表，同时识别特殊符号和依存回指
典型度	语法典型度	例句能否反映词目的典型用法	$\frac{\Sigma(\text{freq-coll})}{\sqrt{\Sigma(character)}}$
	语义典型度	例句能否体现词目的典型语义	$sim(CLStoken, W_i)$
词汇适切性	相对词汇难度	例句整体难度不高于词目难度	$\left[1-\frac{\Sigma(word>l)}{\Sigma(word)}\right]\times 10$
	词汇常用度	例句中词汇的出现和使用频率	$log(w_{freq})$
	汉字笔画数	汉字维度体现例句的复杂程度	$\frac{\Sigma(S)}{\sqrt{\Sigma(character)}}$
句法复杂度	例句长度	例句的长短情况	$\Sigma(character)$
	最大依存距离	例句依存距离的最大值	$\max\|DD_i\|$
	平均依存距离	例句依存距离的平均值	$\text{avg}\|DD_i\|$

注: $\Sigma(character)$ 表示句子的总字数; $\Sigma(word)$ 表示句子的总词例数,$\Sigma(word > l)$ 表示难度等级高于词目难度等级l的词例数; $|DD_i|$ 表示第i个依存关系的距离;Σ(freq-coll) 表示整句中所有搭配的总频次; W_i表示词w_i的词向量; $log(w_freq)$表示对句子中的词例计算频数后取平均值，再取对数便于比较; $\Sigma(S)$ 表示句子中所有汉字的笔画数之和。

Table 1: 例句质量评估指标体系

4 基于机器学习的评估模型构建

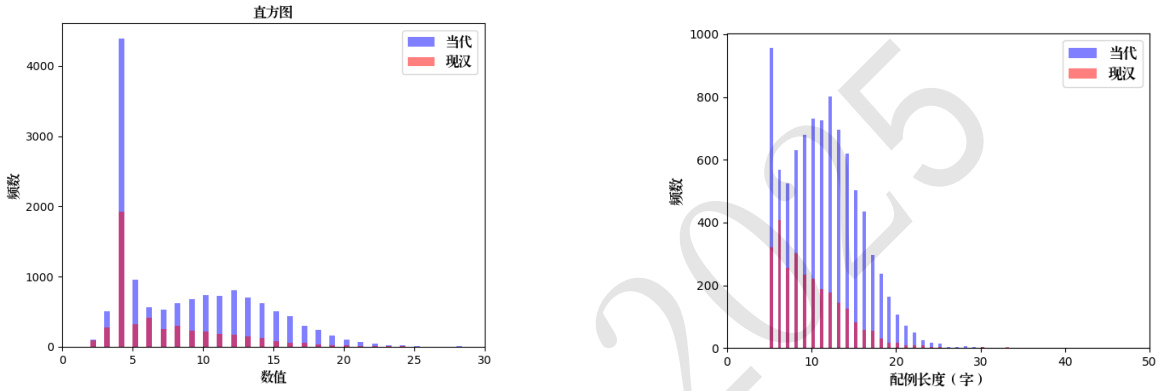
为了验证前文构建的例句特征体系具备科学性和适用性，本章首先选择汉语学习词典和汉语教材中的相关语料，对语料进行标注和处理；再进行特征合理性描述统计实验，分析不同语

料在相同特征上的得分分布；最后以二分类的方式标注语料，构建了基于机器学习的例句质量评估模型，以进一步细化不同特征在衡量例句质量时所占权重，并选取语料库中的真实语料对模型的有效性进行验证。

4.1 语料来源及数据处理

选择《国际中文教育中文水平等级标准》词汇表中的四到六级词汇（即中等次词汇，共3211词）为词目词，查找《现代汉语词典》[1]和《当代汉语学习词典》相应词目的例句。对例句进行标注和整理，去掉一些例句中带括号的解释和注音等（如："三五次翻了一番(数量加了一倍)。"、"非分(fen4)"），共得到例句27788句。对于一般语料，本文选取了《标准中文》、《博雅汉语》、《成功之路》、《发展汉语》、《体验汉语》、《新标准汉语》等共16套76册教材的主课文，通过OCR和人工校对的方法，构建了国际中文教材语料库。这些教材文本涵盖了初级到高级共2000篇文本，去除诗歌、表单等特殊题材和重复出现的文本后，得到1929篇文本（共44731个句子）。匹配其中含有中级词的语料共18599句。由于这些句子并不是专门为相关词汇编写的例句，则可以用这些句子来模拟一般的语料。

通过统计整体语料的句长情况，我们绘制了句长分布的直方图，如图1（a）所示。



(a) 词典语料整体句长分布　　(b) 词典语料5字及以上句长分布

Figure 1: 词典语料配例长度分布直方图

观察该分布图可以发现，两部词典中的例句在长度上呈现出明显的偏态分布，其中5字以下的短句占据了相当大的比例。进一步分析发现，这些短句大多由短语或非完整句构成，例如"很好""快跑""他的书"等。由于本研究聚焦于完整的句子层面（即具有独立句调、能表达完整语义的语言单位），而短语在语法结构、语义复杂度及学习难度等方面与完整句存在显著差异，因此本研究将5字以下的配例排除在分析范围之外。经过上述筛选和整理，我们最终得到的有效例句数量如表2所示,5字及以上的例句分布情况如图1(b)所示。

语料名称	等级范围（级）	词目词数（词）	例句数（句）	例句数（大于等于5字）
《当代汉语学习词典》	4、5、6	2080	19352	8873
《现代汉语词典》	4、5、6	2291	8436	2681
教材语料	4、5、6	3165	18599	14216

Table 2: 语料具体信息

数据的处理流程分为三步。对实验语料进行预处理，调用自然语言处理工具Stanford CoreNLP将每个句子进行分词、词性标注和依存分析，形成相应的标注语料；调用Stanford CoreNLP抽取语料中的搭配，形成搭配频次表和搭配库。对应特征中的每条指标进行计算并形成相应的分数。

[1]中国社会科学院语言研究所词典编辑室. 现代汉语词典(第7版).北京:商务印书馆.2016.

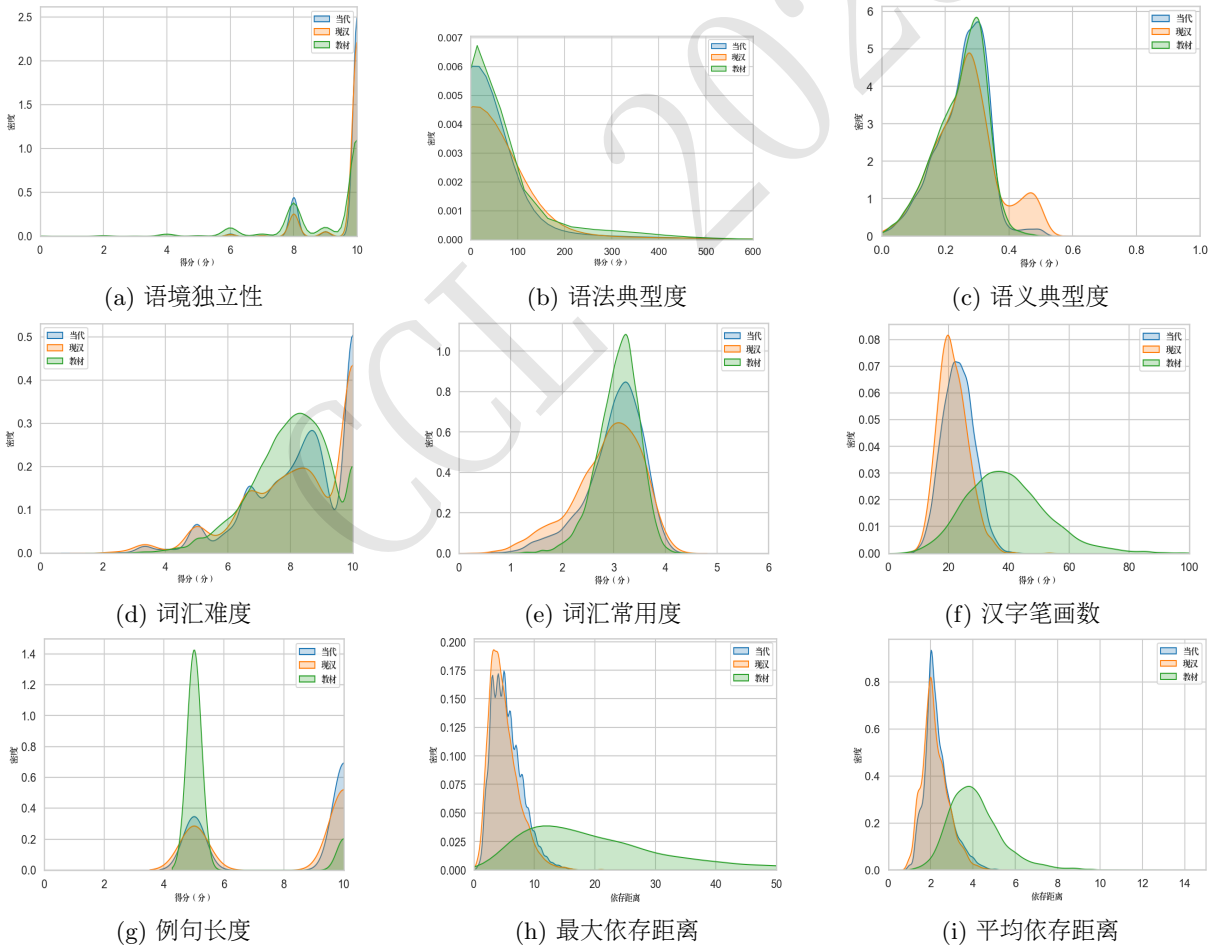
4.2 特征合理性描述统计实验

根据计算所得的分数，表3展示了不同语料分别在不同指标上的平均得分（保留3位小数，加粗部分表示每个指标上的最优得分）。可以看出，除在语法典型度上教材语料平均分为4.514，高于词典语料平均分外，在其他指标上词典语料的平均分均高于教材语料的平均分。而在两部词典之中，最高平均分的分布则较为均衡，且二者之间分值相差较小。

			《当代》	《现汉》	教材语料
平均分	语境独立性	语境独立性	9.646	**9.726**	9.084
	典型度	语法典型度	3.164	2.921	**4.514**
		语义典型度	0.254	**0.266**	0.248
	词汇适切性	相对词汇难度	**8.368**	8.345	8.017
		词汇常用度	**3.094**	2.874	3.082
		汉字笔画数	23.181	**21.394**	39.352
	句法复杂度	例句长度	**8.333**	8.228	5.625
		最大依存距离	5.525	**4.996**	22.042
		平均依存距离	2.331	**2.218**	4.130

Table 3: 各指标平均得分统计图

为进一步观察不同语料上量化指标的分布情况，本文结合数据处理得到的分数绘制了核密度分布图，结果如图2的(a)-(i)图所示。在语境独立性上，得分越高，表示句子脱离上下文后的可理解程度越高。可以看出，教材在满分10分上的密度（约1.1）远低于词典的密度，而在其他分值上的密度则大多高于词典的密度。这说明相比词典，教材在语境独立性的五个维度存在更多的扣分，词典的语境独立性总体优于教材。



(a) 语境独立性 (b) 语法典型度 (c) 语义典型度

(d) 词汇难度 (e) 词汇常用度 (f) 汉字笔画数

(g) 例句长度 (h) 最大依存距离 (i) 平均依存距离

Figure 2: 词典与教材语料各指标得分核密度分布图

典型度特征如图(b)和(c)所示。在语法典型度上，得分越高，表示句子中出现的搭配越典

型。在0-100的低分区间教材的密度高于词典尤其是《现汉》语料，在高分区间两者密度则相差不多，说明词典在语法典型度上差距较小，分布较为稳定。在语义典型度上，在0-0.3分不同语料的语义典型度分布曲线大致相当，在0.3-0.4分《现汉》的密度曲线低于另外两种语料，在0.4分以上词典的密度高于教材，说明词典的语义典型度总体优于教材。

词汇适切性特征如图(d)-图(f)所示。可以看出，在满分10分的密度分布上，词典较教材更为集中，说明词典的词汇难度要优于教材。在平均对数词例频数3.5以上，词典的密度比教材更高，说明词典中词汇常用度总体优于教材。在汉字笔画数上，笔画数越小，阅读的难度就越小。由图可知词典在20-30分密度最高，分布最集中，教材则在40分左右分布最集中，且波峰远低于词典，说明词典例句的阅读难度比教材低。

句法复杂度特征如图(g)到图(i)所示。在句长上，教材语料和词典语料的峰值都集中在5分和10分，在满分10分的分布上，词典语料较教材语料更为集中，说明词典语料其句长更合适用于教学例句。依存距离反映句法关系的丰富度，可以看出，对于最大依存距离，教材语料的分布较为均匀，对于平均依存距离，词典集中在1-3词，教材集中在3-5词，且词典的波峰高于教材。以上数据均表明，教材的句法复杂性总体上高于词典。

通过以上分析可以看出，词典语料在总体上的分布表现均优于教材语料，这一表现也与前文两种语料在各个指标上的平均得分统计情况相符。结合上文的分析与讨论，本文所制定的系列指标具有合理性。

4.3 基于机器学习的例句质量评估模型

为进一步细化不同特征在衡量例句质量时所占权重，得出例句质量评估公式，本文基于机器学习算法构建例句质量评估模型。将词典语料中总体得分最高的600句例句构成正类，再从所有语料中，为每个相同的词目词匹配得分较低的句子作为负类，共1200句。按照二级分类进行人工标注，并通过调用随机森林训练模型、逻辑回归训练模型、SVM训练模型构建分类器，70%数据用于交叉验证，30%数据作为测试集对模型进行评估，得出实验结果如表4所示。

	随机森林	逻辑回归	SVM
Precision	0.9862	0.9612	0.9565
Recall	0.9861	0.9611	0.9556
F1-score	0.9861	0.9611	0.9555
Accuracy	0.9861	0.9611	0.9556

Table 4: 例句质量评估模型准确率

本文选择逻辑回归模型对系数公式进行拟合，最终得出的系数公式如下。公式中的系数能够体现各个指标的比重大小。可以看出，在所有特征中，最大依存距离所占比重最大（大于2.7），其次为语义典型度（大于2.4）和语法典型度（大于2.3），而占比最小的特征为汉字笔画数。这可以从一定程度上反映出依存距离、搭配频率以及例句对词汇语义的体现对于例句整体质量的重要性。

$$\begin{aligned}\text{logit}(p) = & -0.2558 + (0.3693 * \text{句长得分}) + (0.3834 * \text{语境独立性得分}) + (0.3660 * \text{平均对数词例频数}) \\ & + (1.0465 * \text{词汇难度得分}) + (2.3198 * \text{语法典型度得分}) + (2.4437 * \text{语义典型度得分}) \\ & + (0.1808 * \text{汉字笔画数得分}) + (-2.7989 * \text{最大依存距离}) + (-0.6907 * \text{平均依存距离})\end{aligned}$$

4.4 模型有效性验证

根据实验结果，本文随机抽取了DCC动态流通语料库中包含不同中级词的语料对评估模型进行验证，模型输出该句是否为目标词语高质量例句的概率，进而有助于筛选满足汉语学习需求的优质例句。因篇幅所限，本文以词语“皮球”为例，表5的数据展现了模型对其所对应的句子的评估结果（更多示例请见附录A）。

对于词语“皮球”，其高质量例句的概率呈现较为显著的差异。句子“皮球慢慢浮上来。”的概率为0.8965，明显较高，表明这句话对应于“皮球“而言质量较高，能够在保持词语难度较低和较简单的句法结构的同时，提供较为清晰的语境，交代“皮球”能够“浮上来”的特征，适合作为汉语教学的例句。而最后一句中虽然提到了“皮球”的抽象含义，但“八卦”、“高管”、“磨洋工”等词语难度过高，会对学习者的理解造成困难，故而被模型赋予了较低的概率值。

目标词	句子	概率
皮球	皮球慢慢浮上来。	0.8965
	有的打篮球，有的拍皮球，有的打乒乓球，有的手拉手做游戏。	0.4567
	我们去拍皮球吧！	0.3781
	国王一听，像泄了气的皮球，一下子瘫倒在地上。	0.3233
	相反．私下里的八卦中，外企高管踢皮球、磨洋工的现象屡见不鲜，而且和国企的领导一样，都披着日理万机的外表。	0.0142

Table 5: 词语“皮球”高质量例句的概率

为了进一步验证评估体系的合理性，本文采用5级李克特量表对相关例句进行打分，标注工作由十一位具有语言学研究生背景的标注员完成；标注一致性采用Cohen’s Kappa计算，整体的Kappa值为0.76，属于“较高一致”区间；对于存在明显冲突的情况，采用第三方裁定方式确定最终标注结果。计算出平均分数后与相应的概率值进行对比，以词语“活力”所得各句的平均分分布情况为例，如表6所示。可以看出，人工评分所得平均数与评估体系打分情况较为相符，测试中其他词语的例句也符合这一规律（更多评分情况请见附录B）。

句子	人工评分	概率
你很有活力，总是准备着长时间地工作。	4.45	1
吐鲁番明媚的阳光，充满活力的绿树，鲜艳的花朵，黄色的墙，以及维吾尔人从容不迫的生活艺术，让我想永远地生活在这座花园里。	3.45	0.6431
降低细胞活力和对抗感冒、感染和早期癌细胞的能力。	2.81	0.4092
年轻人的笑声格外响亮，年轻人的爱情充满活力。	2.72	0.0369
一般来说，每天6点 7点、10点 11点、18点 21点是三个最适合学习的时间段，在这些时间段里脑细胞活力最强，记忆力也最强，是学习的黄金时间。	2	0.0034

Table 6: 词语”活力”相关句子人工评测情况

通过上述分析，我们可以看到，评估体系模型能够较为准确地对例句质量进行评估，对于每个词语，概率得分较高的句子通常具有较为清晰的语境，能够有效地展示词语的典型意义和用法，同时能够保证整句的词汇难度和句法复杂度不至于过高，符合我们对汉语教学中高质量例句的认知。同时，人工评分所得平均数与评估体系打分情况较为相符，测试中其他词语的例句也符合这一规律。因此，本文所构建的例句质量评估体系模型在汉语教学中具备应用潜力，能够在实际应用中有效地对句子的合适性进行评分，为汉语教学提供质量更高的例句参考。

5 LLMs例句生成能力评估

5.1 LLMs例句生成实验

ChatGPT、Deepseek等大规模语言模型以其强大的语料生成能力为汉语教学带来了新机遇，也为高质量例句的来源提供了新的可能。但与此同时也引发了较多疑问。生成的例句质量如何、表达是否恰当、是否能够应用于实际的汉语学习等，还有待进一步的评估和分析。

由于LLMs作为概率语言模型的本质，提示工程（prompt-engineering）在LLMs的使用中发挥着较为关键的作用。本文结合朱奕瑾、饶高琦 (2023)的研究以及提示工程领域已有提示类型相关研究 (王东清et al., 2025)，采用对比实验设计构建五类提示词模板，核心要素通过表7呈现。

通过随机抽取《国际中文教育中文水平等级标准》中15个中级词汇（分别为：盗版、传言、动画、儿科、景点、盲人、外观、真相、战争、华人、安装、祝愿、关爱、打听、畅通），分别选择ChatGPT 与DeepSeek这两种目前具有代表性的大模型对这些词汇生成共150条例句，最终采用所构建的评估模型计算例句得分，计算各个提示的平均得分如表8所示。

整体来看，各模型在不同提示类型下的表现均较为稳定，平均得分均大于0.5，且两种LLMs均在“约束型”提示中表现最佳，一定程度上说明通过细化字数、结构、搭配等要求，能够有效避免模型生成内容偏离教学需求。

提示类型	核心指令结构	关键参数
指令型	“请分别对词语[X]造一个句子”	仅目标词语
身份赋予型	“假设你是词典编纂者，请分别对词语[X]造一个例句”	目标词语+专业身份限定
示例型	“以词语‘尝试’例句‘教师应敢于尝试新方法’为例，仿照其难度与风格生成[X]的例句”	目标词语+参考样例
约束型	“生成[X]的例句，要求：1.结构完整2.语义独立3.搭配典型4.字数合理5.难度控制6.词汇常用7.句法简单”	目标词语+7项量化规则
混合型	以上四种的结合	以上四种的结合

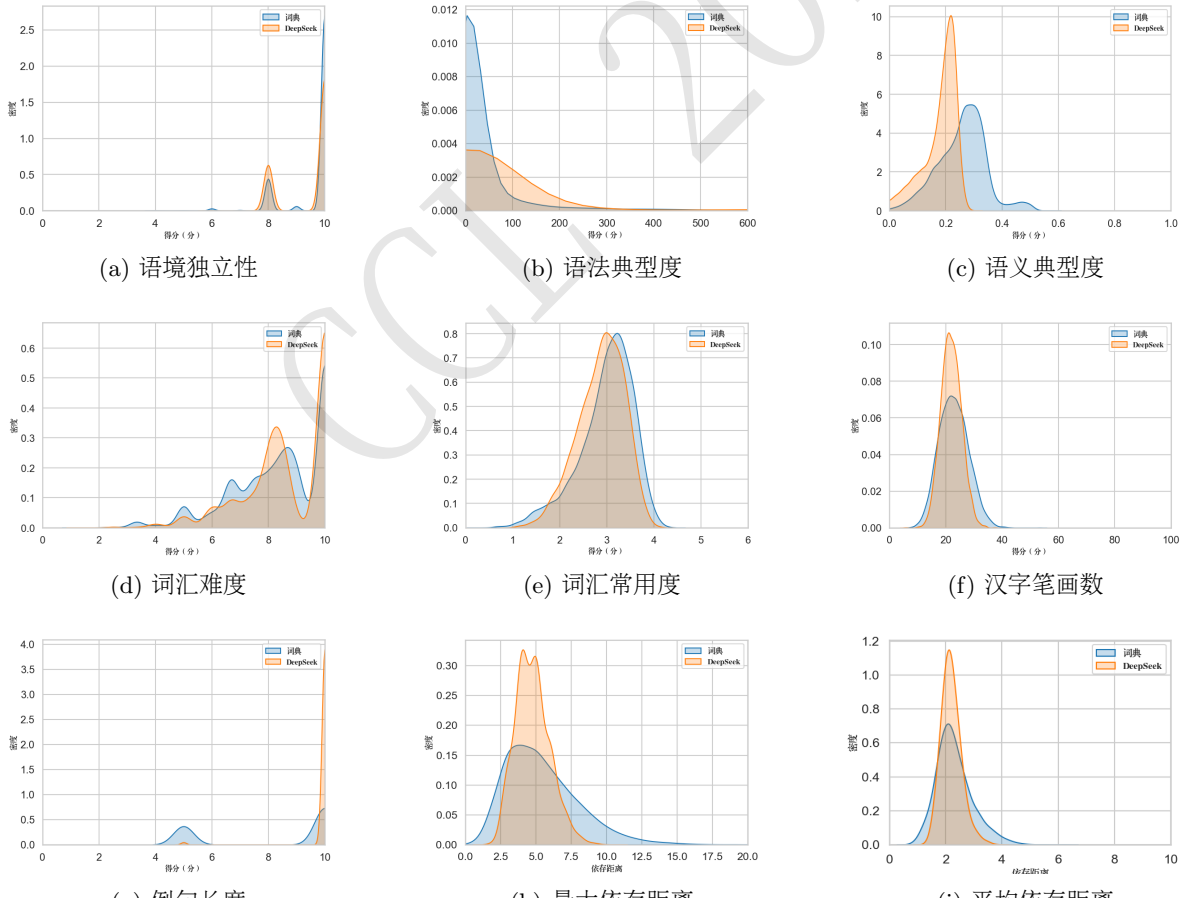
Table 7: 例句质量评估模型准确率

	指令型	身份赋予型	示例型	约束型	混合型
ChatGPT	0.734	0.709	0.654	0.995	0.964
DeepSeek	0.692	0.915	0.839	0.928	0.788

Table 8: 大模型生成例句平均得分

5.2 LLMs生成例句与词典例句对比分析

为了将LLMs生成的例句与人工编写的例句进行对比分析，进一步评估LLMs的例句生成能力，本文采用上节效果最佳的约束型提示应用于DeepSeek模型，对《标准》中的五级词汇分别生成5条例句，经最终经标注获得4578条生成例句。以此为实验组，并选取前文整理的词典语料作为人工编写例句的对照组（去重后共11429句），使用所构建的评估体系对两组例句进行量化分析，结果如图3的图(a)-(i)所示。



(a) 语境独立性 (b) 语法典型度 (c) 语义典型度

(d) 词汇难度 (e) 词汇常用度 (f) 汉字笔画数

(g) 例句长度 (h) 最大依存距离 (i) 平均依存距离

Figure 3: 词典与DeepSeek例句各指标得分核密度分布图

如图(a)，在语境独立性上，DeepSeek在满分10分上的密度低于词典，而在其他区间则呈现出高于词典，说明词典在语境独立性上的表现总体优于DeepSeek。

在典型度方面，图(b)展示了语法典型度得分分布：词典例句在0–70区间密度明显高于DeepSeek，而在70分以上的高分区间，DeepSeek的密度更集中，说明DeepSeek在句法结构的典型性方面优于词典。图(c)展示的语义典型度则呈相反趋势：DeepSeek主要分布在0–0.3区间，而词典例句分布范围更广，集中在0–0.5区间，且高密度区域得分高于DeepSeek，表明词典在表达词语核心语义方面更具优势。

图(d)-(f)展现了词汇适切性特征。在词汇难度上，密度总体呈上升趋势，在满分10的分布上，DeepSeek较词典更集中，说明在词汇难度上DeepSeek的表现优于词典；在词汇常用度上，词典密度最集中的得分（约3.25）较Deepseek（约3）更高，说明在词汇常用度上词典的表现优于DeepSeek；在汉字笔画数上，DeepSeek与词典密度最集中的区域较为一致，但DeepSeek的波峰高于词典，且在区间30-40密度低于词典，说明在汉字笔画数上DeepSeek的分布更集中，优于词典的表现。

图(g)-(i)展现了句法复杂度特征。在例句长度中，可以看到，在满分10分的分布上DeepSeek较词典更为集中，表明在句长这一指标上DeepSeek比词典表现更好，这一定程度上也反映了提示的限制作用。在依存距离上，无论是最大依存距离还是平均依存距离，DeepSeek的分布都更集中，且在较长的依存距离区间密度均低于词典，说明词典的句法复杂度总体高于DeepSeek。

综上所述，以词典语料为代表的人工编写例句在语境独立性、语义典型度、词汇常用度上的总体分布表现优于LLMs，而在语法典型度、词汇难度、汉字笔画数等方面相较LLMs还有待提升。综合考量之下不难发现，DeepSeek为代表的LLMs在提示规定了明确的限制（如句子字数、词汇难度等）时能够较好地完成任务，在提供较为具体和量化的指标时能够表现出良好的生成水平，但在一些单一提示难以概括、较易涉及语感的指标时，表现则略有欠缺，尤其是语义典型度，与词典例句呈现较大差距。这也体现出提示工程对LLMs生成内容的重要性。合适的提示工程能够激发LLMs潜力，使其能更好地应用于语言教学领域的例句生成。

6 结语

针对汉语教学中例句的需求与例句质量之间的差异问题，本文探索汉语学习中例句质量自动评估的新路径，构建了例句质量评估体系，与此同时，通过机器学习方法拟合出系数公式，对LLMs的例句生成能力进行了评估。需要注意的是，虽然本文较为系统地构建了汉语外向型词典例句的质量自动评估体系，但仍存在着一些不足之处，如指标的数量仍较为有限，未来的研究可以考察更多汉语文本特征在例句模型构建中的作用，提升评估体系在更广泛教学情境中的适用性和准确度。同时，在多维度融合评估和生成式增强机制方面深入拓展，增强模型在复杂语言环境下的适应性与判别能力。此外，未来研究中可以注意提升LLMs例句在语义、上下文语境等方面的表现，尝试将评估指标转化为反馈信号，构建基于评估反馈的生成式增强机制，从而进一步提升模型性能。而站在外国学习者学习汉语、理解中华文化的角度上，如何让中国文化上的丰富性润物细无声地融入例句之中，也值得我们在未来进一步的分析和探讨。

参考文献

BT Sue Atkins and Michael Rundell. 2008. *The Oxford guide to practical lexicography*. Oxford University Press.

Kenneth I Forster and Susan M Chambers. 1973. Lexical access and naming time. *Journal of verbal learning and verbal behavior*, 12(6):627–635.

Gisela Harras. 1989. Zu einer theorie des lexikographischen beispiels. In *Wörterbücher. Ein internationales Handbuch zur Lexikografie. Erster Teilband*, pages 607–614. de Gruyter.

Milos Husák. 2010. Automatic retrieval of good dictionary examples. *Bachelor Thesis, Brno*.

Muhammad Yaseen Khan, Abdul Qayoom, Muhammad Suffian Nizami, Muhammad Shoaib Siddiqui, Shaukat Wasi, and Syed Muhammad Khaliq-ur-Rahman Raazi. 2021. Automated prediction of good dictionary examples (gdex): A comprehensive experiment with distant supervision, machine learning, and word embedding-based deep learning techniques. *Complexity*, 2021(1):2553199.

Adam Kilgarriff, Milos Husák, Katy McAdam, Michael Rundell, and Pavel Rychlỳ. 2008. Gdex: Automatically finding good dictionary examples in a corpus. In *Proceedings of the XIII EURALEX international congress*, volume 1, pages 425–432. Universitat Pompeu Fabra Barcelona.

Iztok Kosem, Milos Husák, and Diana McCarthy. 2011. Gdex for slovene. *Proceedings of eLex*, pages 151–159.

Anne O'keeffe, Michael McCarthy, and Ronald Carter. 2007. *From corpus to classroom: Language use and language teaching*. Cambridge University Press.

Thomas M Segler. 2007. *Investigating the selection of example sentences for unknown target words in ICALL reading texts for L2 German*. Ph.D. thesis, University of Edinburgh.

Irena Srdanović and Iztok Kosem. 2016. Gdex for japanese: automatic extraction of good dictionary example candidates. In *GLOBALEX 2016 Lexicographic Resources for Human Language Technology Workshop Programme*, page 57.

Bill VanPatten. 2004. *Processing instruction: Theory, research, and commentary*. Routledge.

付娜. 2010. 外向型汉语学习词典配例中搭配信息的呈现原则及实现条件. 辞书研究, (05):23–30.

刘川平. 2006. 对外汉语学习词典用例的一般原则. 辞书研究, (04):99–106+114.

刘若云, 张念, and 陈粉玲. 2012. 外向型汉语学习词典用例的语言与内容. 语言教学与研究, (4):9–15.

孙全洲. 1986. 《现代汉语学习词典》编纂中的探索. 辞书研究, (03):83–88.

张. 2005. 对外汉语教材例句研究. 硕士, 北京语言大学.

徐玉敏, 李禄兴, 杜健, and 张伟. 2005. 编纂汉语学习词典理论与实践之探索. 对外汉语学习词典学国际研讨会论文集.

徐茗. 2009. 对外汉语词汇教学中的例句设计. 安徽师范大学学报*(*人文社会科学版*)*, 37(04):462–466.

朱奕瑾and 饶高琦. 2023. 基于chatgpt的生成式共同价值标准例句库建设. 云南师范大学学报*(*对外汉语教学与研究版*)*, (21(03)):71–80.

李泉and 金允贞. 2008. 论对外汉语教材的科学性. 语言文字应用, (4):108–117.

李淑云. 2023. 基于语料库的同义词词典配例方法研究. 硕士, 山东大学.

李禄兴. 2006. 单语外向型汉语学习词典的设例. 对外汉语学习词典学国际研讨会论文集（二）, pages 193–213.

李禄兴. 2015. 谈对外汉语词典被释词的搭配原则——以"交流"为例. 现代语文*(*语言研究版*)*, (11):68–71.

杨玉玲and 段彤彤. 2023. 外向型汉语学习词典配例难度控制原则及编写建议——以《当代汉语学习词典》为例. 国际中文教育*(*中英文*)*, 8(03):66–75.

王东清, 芦飞, and 张炳会等. 2025. 大语言模型中提示词工程综述. 计算机系统应用, (34(01)):1–10.

胡韧奋and 肖航. 2019. 面向二语教学的汉语搭配知识库构建及其应用研究. 语言文字应用, (1):135–144.

宁臧. 2022. 基于语义逻辑的外向型学习词典量词释义及其配例研究——以《汉语教与学词典》为例. 硕士, 辽宁师范大学.

蔡永强. 2011. 对外汉语学习词典编纂的用户友好原则. 辞书研究, (2):67–77.

邵敬敏and 朱晓亚. 2005. "好"的话语功能及其虚化轨迹. 中国语文, (05):399–407.

郭小娜. 2023. 《当代汉语学习词典》双音节动名兼类词配例研究. 硕士, 河北师范大学.

鲁健骥and 吕文华. 2006. 编写对外汉语单语学习词典的尝试与思考——《商务馆学汉语词典》编后. 世界汉语教学, (01):59–69.

第二十四届中国计算语言学大会论文集，第117页–第130页，济南，中国，2025年8月11日至14日。

A 例句评估示例

目标词	句子	概率
皮球	皮球慢慢浮上来。	0.8965
	有的打篮球，有的拍皮球，有的打乒乓球，有的手拉手做游戏。	0.4567
	我们去拍皮球吧！	0.3781
	国王一听，像泄了气的皮球，一下子瘫倒在地上。	0.3233
	相反．私下里的八卦中，外企高管踢皮球、磨洋工的现象屡见不鲜，而且和国企的领导一样，都披着日理万机的外表。	0.0142

Table 9: 词语“皮球”例句得分

目标词	句子	概率
操场	玛丽想去操场锻炼，但是她最怕冷。	0.9981
	操场上积满了白花花的水。	0.9954
	伟伟，这是操场，我们在这儿踢球、跑步。	0.7280
	我再也不用为那个像操场一样大的火车担心了。	0.2842
	我们在宿舍楼、教室、操场、食堂、图书馆、体育馆，每一个曾经去过的地方，都照了相、合了影，作为留念。	0.0142

Table 10: 词语“操场”例句得分

目标词	句子	概率
出色	我想你以后一定会成为一名出色的汉语专家。	1
	王羲之还是一个出色的书法教育家。	0.9963
	双方的技术都发挥得相当出色。	0.8959
	那是，要是没经过高考的历练和大学的教育，怎么可能有现在如此出色的晓雯呢?	0.1640
	高晓松表示，自己曾经担任过某全国性原创歌曲大赛的总策划，收到3000多首歌，竟挑不出一首非常出色的作品。	0.0545
	是他的狂放姿态，是他带泪的出色表演，让我们忽略了他内心的痛苦，忽略了他对东方文化的积极思考，忽略了他对这片土地命运的深切关注，也忽略了他曾做出的坚定而绝望的挣扎。	0.0426

Table 11: 词语“出色”例句得分

目标词	句子	概率
活力	你很有活力，总是准备着长时间地工作。	1
	吐鲁番明媚的阳光，充满活力的绿树，鲜艳的花朵，黄色的墙．以及维吾尔人从容不迫的生活艺术，让我想永远地生活在这座花园里。	0.6431
	降低细胞活力和对抗感冒、感染和早期癌细胞的能力。	0.4092
	年轻人的笑声格外响亮，年轻人的爱情充满活力。	0.0369
	一般来说，每天6 点7 点、10 点11 点、18 点21 点是三个最适合学习的时间段，在这些时间段里脑细胞活力最强，记忆力也最强，是学习的黄金时间。	0.0034

Table 12: 词语“活力”例句得分

第二十四届中国计算语言学大会论文集，第117页-第130页，济南，中国，2025年8月11日至14日。

B 人工评分与模型评分对比

句子	人工评分	概率
皮球慢慢浮上来。	3.73	0.8965
有的打篮球，有的拍皮球，有的打乒乓球，有的手拉手做游戏。	3.45	0.4567
我们去拍皮球吧！	3.18	0.3781
国王一听，像泄了气的皮球，一下子瘫倒在地上。	2.82	0.3233
相反．私下里的八卦中，外企高管踢皮球、磨洋工的现象屡见不鲜，而且和国企的领导一样，都披着日理万机的外表。	1.64	0.0142

Table 13: 词语”皮球”相关句子人工评分情况

句子	人工评分	概率
玛丽想去操场锻炼，但是她最怕冷。	4	0.9981
操场上积满了白花花的水。	3.33	0.9954
伟伟，这是操场，我们在这儿踢球、跑步。	3	0.7280
年轻人的笑声格外响亮，年轻人的爱情充满活力。	2.72	0.0369
我再也不用为那个像操场一样大的火车担心了。	2.33	0.2842
我们在宿舍楼、教室、操场、食堂、图书馆、体育馆，每一个曾经去过的地方，都照了相、合了影，作为留念。	1.67	0.0142

Table 14: 词语”操场”相关句子人工评分情况

句子	人工评分	概率
我想你以后一定会成为一名出色的汉语专家。	4.82	1
王羲之还是一个出色的书法教育家。	4.72	0.9963
双方的技术都发挥得相当出色。	4	0.8959
那是，要是没经过高考的历练和大学的教育，怎么可能有现在如此出色的晓雯呢?	3.67	0.1640
高晓松表示，自己曾经担任过某全国性原创歌曲大赛的总策划，收到3000多首歌，竟挑不出一首非常出色的作品。	2	0.0545
是他的狂放姿态，是他带泪的出色表演，让我们忽略了他内心的痛苦，忽略了他对东方文化的积极思考，忽略了他对这片土地命运的深切关注，也忽略了他曾做出的坚定而绝望的挣扎。	1.18	0.0426

Table 15: 词语”出色”相关句子人工评分情况

句子	人工评分	概率
你很有活力，总是准备着长时间地工作。	4.45	1
吐鲁番明媚的阳光，充满活力的绿树，鲜艳的花朵，黄色的墙，以及维吾尔人从容不迫的生活艺术，让我想永远地生活在这座花园里。	3.45	0.6431
降低细胞活力和对抗感冒、感染和早期癌细胞的能力。	2.81	0.4092
年轻人的笑声格外响亮，年轻人的爱情充满活力。	2.72	0.0369
一般来说，每天6点 7点、10点 11点、18点 21点是三个最适合学习的时间段，在这些时间段里脑细胞活力最强，记忆力也最强，是学习的黄金时间。	2	0.0034

Table 16: 词语”活力”相关句子人工评分情况