失语症话语中的任务依赖型语言测量：叙事性、描述性和程序性言语的比较分析

Emma AlKubaisi斯德哥尔摩大学，瑞典斯德哥尔摩

ORCID： https://orcid.org/0009-0003-0689-0710

通讯作者：Emma Al Kubaisi，emma.alkubaisi@gmail.com

文章类型：研究报告

词数：6,974（摘要和正文，包括文内引用；不包括表格、图注和参考文献）

摘要

背景：基于转录文本的测量指标，如词数、词汇多样性和不流畅率，被广泛用于描述失语症患者的语言能力。每项测量指标都来自特定的引出任务，而不同研究采用的任务各不相同，因此不同任务所得的数值可能无法进行比较。

目的： 检验话语任务是否会改变同一说话者的这些指标，以及这些指标是否会按照从程序性到描述性再到叙事性话语的梯度变化。

方法与程序： 从 AphasiaBank 中的 Fridriksson-2语料库（POLAR试验，ClinicalTrials.gov NCT03416738）获取了 61 名慢性卒中后失语症成年人的基线转录文本。每位说话者在一次会话中分别完成一个程序性（三明治）、描述性（破窗）和叙事性（灰姑娘）样本。基于规则的处理流程提取了 13 项指标，其中 10 项足够常见，可用于检验。研究采用 Friedman、Page 和 Wilcoxon检验，并进行错误发现率校正；同时使用按说话者分组的随机森林分类器进行辅助验证。

结果： 总词数和不同词数呈现完整梯度，但从程序性到描述性的变化幅度较小。对于大多数其他指标，叙事性话语与其他话语类型明显不同，而描述性与程序性话语之间则无差异。在每项任务中至少说出 25 个词的 31 名说话者中，叙事性话语的填充词率、重复率和总不流利密度均高于另外两种任务。分类器的准确率为 0.61（若始终猜测最常见的任务，准确率为 0.41），其主要依据是识别叙事性样本。

结论：话语任务不能互换，按每个词计算言语不流畅现象也无法消除任务效应。报告每项测量结果时都应注明任务，且仅应比较同一任务中的数值。

关键词：失语症；话语；任务效应；言语不流畅；自动化语言分析

1. 引言

话语或连贯言语展现了失语症患者如何在接近日常交流的情境中使用语言。它揭示了单词测试无法反映的词语提取困难、自我修正、重复和犹豫（Dutta 等人，2025）。手动分析话语需要耗费大量时间，而时间也是临床医生和研究人员最常提及的障碍（Stark & Dalton, 2024）。自动化方法减轻了这一负担。在 AphasiaBank 中，转录员使用 CHAT转录格式的固定代码标记重复、修正、填充词、词片段和不可理解言语，因此可以自动统计这些行为（MacWhinney, 2000；MacWhinney 等，2011；Fromm 等，2024）。以这种方式提取的测量指标如今已被用于例如估计标准化失语症测试分数（Riccardi 等，2024）。

然而，每项测量指标都是从特定任务中提取的，而不同研究所使用的任务各不相同。自动化流利度和宏观语言学分析仅使用了灰姑娘故事（Fromm 等人，2024；Lee 等人，2026）；一项关于词语信息量的研究使用了单一的图片描述任务（Pittman 等人，2025）；还有一项关于失语症严重程度的研究结合了三项任务（Riccardi 等，2024）。如果任务会改变测量结果，就不能将来自不同任务的数值视为相等并进行比较。这样一来，两项研究之间的差异可能反映的是任务而非说话者，而两次评估之间的变化也可能被误认为语言能力发生了变化。在患有失语症的同一批说话者中，图片描述与对话仅呈中度或弱相关（Leaman & Edmonds，2023）；而将来自不同体裁的样本合并，可能会掩盖其中一种体裁中的困难（Bryant 等人，2016）。

人们认为，不同任务的认知语言负荷有所不同。程序，例如解释如何制作三明治，是一种熟悉的固定程序，其步骤顺序是固定的。与所有话语类型相比，程序的词汇多样性最低（Fergadiotis 等人，2011），且使用简单的名词密集型表达（Stark & Fukuyama，2021）。在图片序列任务中，图片始终呈现在眼前，充当外部记忆，并提示所需的词语。故事复述是负荷最高的任务：说话者必须在没有图片的情况下召回一个较长的故事，在脑中保持其整体结构（Kintsch & van Dijk，1978），并将事件按时间和因果关系联系起来。与图片描述相比，故事复述对工作记忆和执行功能提出了更高的要求（Dutta 等人，2025）；而失语症患者的叙事表现与言语执行技能相关（Dutta 等人，2024）。这些

观点表明，其顺序为程序性 < 描述性 < 叙事性。这一顺序是一种假设，而非测量结果；它结合了两个不一定一致的维度：内容的固定程度，以及是否必须从记忆中召回内容。

任务要求也会影响言语不流畅现象。在典型说话者中，言语不流畅现象通常表明说话者在规划信息、选择词语或构建句子时遇到了困难（Didirková，2024）。诸如“呃”和“嗯”之类的填充词表明说话者预计会出现延迟（Clark & Fox Tree，2002）；话语越长、表达形式越缺乏结构，言语不流畅现象就越多（Oviatt，1995）。因此，对规划和词语提取要求更高的任务应会产生更多填充词、修正和言语中断。

早期研究表明，任务很重要。在 AphasiaBank 中，话语类型会影响失语症说话者和非失语症说话者的大多数语言学指标（Stark，2019），而同一体裁的任务在微观结构方面最为相似（Stark & Fukuyama，2021）。在普通话和克罗地亚语中，典型说话者的词汇多样性因任务而异，但失语症说话者之间的差异并不显著（Deng 等人，2024；Jozipović 等人，2026）。最近，一项分析了 TalkBank 中近 10,000 个样本、并自动提取了近 300 项特征的研究发现，任务类型会影响每个临床组的语言输出（Stark & Themistocleous，2026）。该研究未纳入不流畅度指标，而且并非所有参与者都完成了所有任务。关于言语不流畅现象的证据则较为有限。与图片序列描述相比，灰姑娘故事中的重复和回溯更频繁，但与三明治制作程序相比则并非如此（Culbertson，2022）。健康成人在开放式回答中的每词言语不流畅程度高于图画书叙事（Walters 等人，2026）；而在极轻度失语症中，自动化流利度测量指标在叙事性话语的重复测试中最为可靠（Stark 等人，2026）。

另一个问题是样本长度。不同任务的长度差异很大，而总词数和不同词数（NDW）等指标会随言语量增加。移动平均形符类符比（MATTR）旨在消除此依赖性（Covington & McFall，2010年），言语不流畅现象则可按每100词计算。目前尚不清楚，在控制长度后，任务对言语不流畅现象的影响是否依然存在；据我们所知，也没有研究检验过在所有说话者都完成每项任务的情况下，各项特征是否呈现假设中的有序梯度。本研究使用 Fridriksson-2语料库探讨这些问题。该语料库中的同一批失语症说话者完成了程序性、描述性和叙事性任务。

2. 研究问题与假设

本研究探讨话语任务的选择是否会改变从失语症话语中提取的语言学指标。本研究围绕以下一个研究问题展开：

研究问题： 具有不同认知语言需求的话语任务是否会产生不同的语言特征谱？这些指标是否会在程序性、描述性和叙事性话语之间呈现系统性的复杂度梯度？

根据引言中回顾的文献，提出了两项预测：

- **输出与词汇：** 反映言语数量和多样性的特征（总词数、不同词数和平均话语长度）将呈现完整的复杂度梯度：程序性 < 描述性 < 叙事性。

- **言语不流畅与中断：** 每百词的言语不流畅和中断比率（填充词、修正和不可理解言语）在叙事性话语中将高于描述性和程序性话语，而不会在三项任务中逐步递增。

此外，本研究还探讨了长度校正词汇多样性（移动平均词汇类型比，MATTR）、总不流畅密度及其他不流畅率在不同任务间是否存在差异。这项分析属于探索性分析，未提出方向性预测。

3. 方法

本研究采用确定性的自然语言处理流程，从话语转录文本中提取语言特征，随后进行重复测量统计分析和机器学习检验。

3.1 数据来源

数据取自 Fridriksson-2语料库，该语料库存档于 TalkBank 系统中的 AphasiaBank 数据库（MacWhinney 等，2011）。AphasiaBank 是一项共享的多媒体资源，用于研究失语症患者的语言产出。Fridriksson-2语料库包含卒中后失语症患者的话语转录文本，这些文本是在 POLAR（失语症语言康复结局预测）临床试验中收集的（ClinicalTrials.gov NCT03416738；Kristinsson 等，2021年、2023年）。POLAR 纳入了年龄为 21–80 岁、至少 20 年以英语为主要语言且患有慢性失语症的成年人（左半球缺血性卒中或出血性卒中后至少 12 个月）。言语产出严重受限的人

或听觉理解严重受限者（西方失语症成套测验修订版 [WAB-R]任一分测验得分为 0–1），或双侧卒中患者均被排除在外。测试在南卡罗来纳大学和南卡罗来纳医科大学进行；COVID-19 暴发后入组试验的参与者接受了线上评估和治疗（Kristinsson 等，2021年）。参与者接受了为期 3 周的语音学侧重命名治疗和为期 3 周的语义侧重命名治疗，顺序随机；并在治疗前、两个治疗阶段之间及治疗结束后，以及治疗后 1 个月和 6 个月接受评估（Kristinsson 等，2021年、2023年）。每个测试时间点，参与者均完成 AphasiaBank 标准话语方案中的三项任务：破窗图片序列、灰姑娘故事和花生酱果酱三明治程序。所有转录文本均遵循 CHAT转录标准（MacWhinney，2000）。

由于试验在各测试时间点之间实施了治疗，因此仅使用治疗前的基线转录文本（时间点1）。这避免了将任务效应与治疗相关变化混淆：在同一队列中，治疗后《灰姑娘》复述发生了变化，平均话语长度（平均话语长度）、每分钟词语数和核心词汇均有所增加（den Ouden 等，2026）。在说话者内分析中，仅纳入基线时完成全部三项任务的参与者，因为 Friedman 检验和 Page 检验要求每位说话者具有完整的观测数据集。语料库描述列出68名参与者（Fridriksson，2023），但实际提供了67名参与者的转录文本。其中，64名参与者有基线转录文本：其中一名说话者的基线录音故障过多，无法转录；另两名说话者从第二个测试时间点开始。64份基线转录文本中有3份缺少一项任务（一份由于视频故障缺少破窗任务，一份仅包含破窗任务，另一份缺少《灰姑娘》任务）。因此，最终样本包括61名说话者和183份话语样本。

这 61 位说话者是 POLAR 队列的一个子集。表 1 展示了已发表的队列特征（Kristinsson 等，2023年）；该子集的特征未单独报告。

表 1已发表的 *POLAR* 队列特征（*Kristinsson* 等，*2023*）

特征	均值（标准差）或人数（百分比）	范围
年龄（岁）	60.5（10.9）	29–80
女性	42 (41.2%)	
非西班牙裔白人	75 (73.5%)	
右利手	90（88.2%）	
受教育年限（年）	15.4（2.3）	12–20
卒中后时间（月）	46.2（48.1）	12–241
美国国立卫生研究院卒中量表评分	6.2 (3.7)	0–16
西方失语症成套测验修订版失语商	60.0 (22.7)	14.5–93.1
ASRS 言语失用严重程度	1.6（1.6）	0–4
糖尿病病史	23（22.5%）	
抑郁症病史	24（23.5%）	
失语症类型：布罗卡失语症	45 (44.1%)	
失语症类型：命名性失语	29 (28.4%)	
失语症类型：传导性失语	16 (15.7%)	
失语症类型：韦尼克失语症	7（6.9%）	
失语症类型：全面性	4（3.9%）	
失语症类型：经皮质运动性	1（1.0%）	

注。 这些数值描述了 Kristinsson 等（2023）报告的 102 名 POLAR 参与者，他们完成了所有治疗程序，并至少接受了一项结果评估。本研究的 61 名说话者样本取自这一队列。美国国立卫生研究院 = 美国国立卫生研究院（最高分 42）； WAB-R= 西方失语症成套测验修订版（最高分 100）； ASRS = 言语失用症评定量表（最高严重程度评分 4）。

伦理。本研究是对现有去标识化转录文本进行的二次分析。POLAR试验已获南卡罗来纳大学和南卡罗来纳医科大学的机构审查委员会批准，所有参与者均同意研究程序（Kristinsson 等，2023年）。我们通过 AphasiaBank，依据 TalkBank 的数据访问和使用规则获取了这些转录文本。下载的仅为文本转录文本；未访问任何音频或视频录音，转录文本中不含姓名或其他直接身份标识信息，也未联系任何参与者。由于未处理任何可识别个人身份的数据，因此无需依据《瑞典伦理审查法》（2003:460）另行获得伦理批准。

3.2 话语任务

三项任务及其假定的要求总结于表2中。任务说明遵循标准的 AphasiaBank 协议（MacWhinney 等，2011）。选择这三项任务，是因为它们属于三种不同的体裁（Riccardi 等，

2024；Stark，2019；Stark & Fukuyama，2021），由同一批说话者在同一会话中完成（每份基线转录文本均关联一段单独的视频录制），且给予说话者的支持度各不相同。所有61位说话者都按相同顺序完成任务：破窗、灰姑娘，然后是三明治。在这个语料库中，三明治任务配有一张花生酱和果冻三明治的照片；对于难以完成该任务的说话者，标准的 AphasiaBank 协议提供的并非这张图片（Fridriksson，2023年）。

表 2话语任务及其假定的认知语言需求

Task	体裁	Task 描述	视觉支持	主要需求	假定的要求
三明治	程序性	解释如何制作一个花生酱和果酱三明治	一张三明治的照片	对熟悉的固定程序的召回率；小型并重复词汇；固定步骤顺序	最低
损坏 Window	描述性（图片序列）	讲述一系列事件中图片	图片保留在 view	描述并连接可见事件；内容受到限制并由图片提示	中级
灰姑娘	叙事性（故事复述）	浏览无字图画书，然后复述故事，之后将书收起来 away	无（书已移除）	长篇故事的召回率依靠记忆；追踪多个角色和情节；时间和因果联系	最高

最后一列的排序依据是《引言》中回顾的文献。这是Page趋势检验所检验的假设，并非对任务难度的独立测量。

3.3 数据预处理

转录文本以CHAT（.cha）格式存储，并遵循失语症库标注标准。这些文件包含多个层级，包括参与者言语、检查者言语和元数据。

我们实施了基于Python的预处理流程，以提取并清理参与者言语。该处理流程提取参与者言语层（*PAR:）及其续行，并移除非语言材料，包括检查者言语、媒体标头和CHAT元数据，从而确保所有分析仅基于参与者产出的语言。

我们使用嵌入CHAT文件中的@G（gem）标记，将每份转录文本拆分为任务特定样本。这些标记用于区分描述性（Window）、叙事性（灰姑娘）和程序性（三明治_图片）任务，从而为每位参与者的每项任务生成一个样本。我们未分析灰姑娘_Intro gem，因为说话者在该环节中只是翻看书本。

处理流程采用正则表达式（regex）实现，以确保所有文件的解析一致、确定且可复现。词数是在移除所有方括号中的 CHAT代码、填充词（&-uh）、词片段（&+fr）、其他 &-代码（如声音标记（&=laughs））和其他说话者的感叹词、占位符 xxx、yyy 和 www、省略词代码（0-words）以及时间戳后统计的。特殊形式标记会从其所属词语中移除（例如，son@k 计为 son），而以下划线连接的词语（you_know）计为一个词。重复词和修订词均计入词数。带有后置代码 [+ exc], 的话语会被排除；转录员使用该代码标记不属于任务内容的评论（例如“I don't know”或“okay”），这与 CLAN程序的默认分析方式一致（MacWhinney，2000）。平均话语长度（MLU）的计算方法是用词数除以至少包含一个词的话语数，因此，仅由填充词、词片段或不可理解言语构成的话语不会被计为长度为零的话语。

3.4 自动特征提取

3.4.1 言语不流畅检测（正则表达式引擎）

为了将转录文本转化为可测量的特征，本研究使用了确定性规则驱动正则表达式引擎。该引擎统计转录员在转录文本中插入的 CHAT言语不流畅代码。因此，除了标准的 CHAT 转录之外，无需进行人工编码，但其输出取决于转录的质量和一致性。

该引擎提取八种言语不流畅类型：重复、修正、重新表述、填充词、音位片段、停顿、截断和不可理解言语。每个标记均通过预定义的正则表达式模式进行识别，该模式对应于一项 CHAT规范。这些类别及其模式列于表 3 中。言语不流畅现象可以通过多种方式进行分类，例如分为犹豫和自我修正，或分为类似口吃的言语不流畅现象与其他言语不流畅现象（Penttilä等，2022）；不同研究所采用的术语也存在很大差异

（Didirková，2024）。遵循 CHAT 规范意味着，这些类别是根据共享的转写代码定义的，而不是根据本研究自行选择的标签定义的。

表 3自动提取言语不流畅现象的正则表达式模式

言语不流畅 Type	语言学 / 临床 Term	正则表达式模式	描述
重复	重复（回溯未进行修正）	\[/\]	重复词或短语
修正	自我修正（回溯带修正）	\[//\]	说话者修正或更改话语
重新表述	带回溯的重新表述	\[///\]	说话者重新开始，采用不同的表达方式
填充词	填充停顿	&-\w+	非词汇填充词（例如，“呃”，“嗯”）
音位片段	词片段	&\+\w+	不完整或部分词语产出（例如 &+fr）
停顿	停顿	\((?:\.{1,3}|[\d:]*\d\.?\d*)\)	停顿代码 (.)、(..)、(...) 和计时停顿，例如 (1.5)
截断	截断词	\b[A-Za-z]+-(?=\s|$)	突然被截断的词语（例如，"go-"）
无法听清	不可理解言语	\b(?:xxx|yyy)\b	无法被理解

3.4.2 语言学指标

除言语不流利计数外，处理流程还会计算言语产出、话语长度、词汇多样性和总体不流利度等指标。完整的特征集汇总于表4。

- **总词数（TWC）**衡量总体言语产出。
- **平均话语长度（MLU）**指每个话语的平均词数，反映话语长度和复杂度。

- **不同词数（NDW）**统计样本中的独特词语数量，反映词汇丰富度。NDW 在很大程度上取决于样本长度。
- **移动平均形符类符比（MATTR）**计算固定大小的移动 Window 内的型符比，并对其在整个样本中的数值取平均（Covington & McFall，2010年）。与 NDW 不同，它基本不受样本长度影响。由于许多程序性和描述性样本都很短，因此使用了包含25个词语的 Window。

- **不流畅率**表示每100个词语中的不流畅次数，因此不会随着言语量的增加而增加。

- **总不流利密度（TDD）**是检测到的不流畅标记总数除以总词数：

$$TDD = \text{（不流畅标记）} / \text{（总词数）}$$

对于非常短的样本，比率、总不流利密度和移动平均词汇类型比都不可靠。因此，少于25个词的样本在这些指标上不赋值；对于每项指标，只有当说话者的三个样本都至少包含25个词时，才将其纳入分析。所有样本均计算总词数、不同词数和平均话语长度。

表4分析的语言特征

特征	**领域**	**定义**	**随……增长 样本长度?**
总词数（TWC）	输出	产出的词数	Yes
平均话语长度 （平均话语长度）	输出 / 句法	每个话语的词数	否（平均每个话语）
不同词数 （NDW）	词汇 多样性	独特词语数量	Yes
移动平均词 汇类型比	词汇 多样性	a 中的平均型符比 25词移动窗口中的平均型符比	否（设计如此）
每100词填充词数	言语不流畅	每100词填充停顿数	否（比率）
每100词的重复次数	言语不流畅	[/]每100词标记数	否（比率）
每100词的修订	言语不流畅	[//] 每100词的标记数	否（比率）
每100词的不可理解词数	言语中断	每100个片段中有xxx / yyy个词语	数量（比率）
每100词的音位片段数 词语	言语不流畅	&+ 每100词的片段数	否（比率）
每100词的停顿数	言语不流畅	每100词的停顿代码数	否（比率）
每100词的重述数	言语不流畅	[///] 每100词的标记数	否（比率）
每100词的截断数	言语不流畅	每100词的截断词数	否（比率）
总不流利密度 （TDD）	言语不流畅	所有不流畅标记 / 总词数	否（每词比率）

注。 停顿、重新表述和截断过于少见，无法进行检验（超过90%的样本均未出现；参见第4.1节）。

3.5 统计分析

由于每位说话者都完成了全部三项任务，因此该设计完全属于被试内设计。这一点很重要，因为不流畅行为具有很强的个体差异。即使是典型说话者，也可以仅凭其不流畅模式区分彼此（Braun等，2023）。将每位说话者与其自身进行比较，可以消除这一变异来源。特征分布呈偏态，且许多不流利计数都很低，因此

采用了参数重复测量检验。只有当某项测量中为零的样本比例低于 90% 时，才对其进行检验。分析分五个步骤进行。

步骤 1：弗里德曼检验。对于每个特征，均采用弗里德曼检验（Friedman, 1937）检验三项任务整体上是否存在差异。弗里德曼检验会对每位说话者的三个数值进行秩排序，因此能够在控制说话者之间差异的同时比较各项任务。采用肯德尔 W（Kendall & Babington Smith, 1939）作为效应量，其中 0 表示不存在一致的任务效应，1 表示所有说话者对各项任务的排序方式完全相同。

步骤 2：Page趋势检验。弗里德曼检验只能检验各项任务是否存在差异。为直接检验有序假设，按照预测顺序程序性 <描述性 < 叙事性，采用了 Page趋势检验（Page, 1963）。当备择假设指定了顺序时，Page趋势检验的效力高于弗里德曼检验。显著的 Page趋势检验结果表明存在朝预测方向发展的趋势，但不能说明该顺序中的每一步都成立。

步骤 3：成对威尔科克森符号秩检验。为定位差异，使用双侧威尔科克森符号秩检验（Wilcoxon, 1945）比较每一对任务（叙事性 vs. 程序性、叙事性 vs. 描述性，以及描述性 vs. 程序性）。效应量为匹配对秩双列相关 r，其范围为 −1 至 1；正值表示该对比较中的第一个任务具有更高的数值。其 95% 置信区间通过对说话者进行自助法估计（2,000 次重抽样）。

步骤 4：错误发现率校正。由于评估了许多特征和检验，因此按照 Benjamini–Hochberg错误发现率（FDR）程序（Benjamini & Hochberg, 1995），分别对每类检验的 p 值进行校正：所有弗里德曼检验、所有 Page检验以及所有两两 Wilcoxon检验。当 q <.05 时，结果被视为显著。

步骤 5：响应模式分类。 根据步骤 1–4，按照以下顺序所列规则，将每个特征归入四种模式之一：

- **完整梯度：** Page检验趋势显著，且三组两两差异均沿预测方向显著（程序性 < 描述性 < 叙事性）。
- **叙事性较高：** 叙事性显著高于描述性和程序性，而描述性与程序性之间没有显著差异。

- **没有任务效应的证据：** 经 FDR校正后，Friedman效应不显著。由于检验结果不显著并不能说明两项任务相同，因此不将这些特征描述为稳定。

- **其他：** 任何其他模式。

所有分析均使用 Python 3.14 完成；统计检验使用 SciPy 1.18（Virtanen等人，2020），分类器使用 scikit-learn 1.9（Pedregosa等人，2011）。

3.6 机器学习验证

统计检验每次考察一个特征。作为补充检验，我们使用了一个监督分类器，测试综合考量这些特征时，它们所包含的信息是否足以识别样本来自哪项任务。该分类器并非主要实验，其作用是提供支持性证据，表明三项任务具有不同的语言特征。

我们训练了一个包含500棵树的随机森林分类器（Breiman，2001），将每个样本分配到三项任务之一（节点分裂所需的最少样本数为5，每次分裂考虑的特征数为特征总数的平方根）。所用特征包括总词数、不同词数、平均话语长度、总不流利密度和五种经过检验的不流畅率；这些特征均以原始值使用，未进行缩放。由于短样本无法计算移动平均词汇类型比，因此未纳入该指标；同时排除了存在缺失值的样本，最终保留了来自52位说话者的126个样本。我们还训练了第二个模型，去除总词数和不同词数，以检验在无法获取言语量的情况下，是否仍能区分这些任务。未对任何设置进行调优。

采用按说话者分组的10折交叉验证评估表现，确保同一说话者不会同时出现在训练数据和测试数据中。按记录划分数据会使模型能够识别个别说话者，从而可能夸大准确率（Saeb等人，2017）。表现以总体准确率、宏平均F1分数，以及各类别的精确率、召回率和F1分数报告，并附上混淆矩阵。对于三分类任务，机会水平宏平均F1约为0.33。由于排除了含缺失值的样本，各类别样本数量并不均衡（34个程序性样本、40个描述性样本和52个叙事性样本），因此始终猜测叙事性类别的准确率将为0.41。

4. 结果

4.1 样本与描述性统计

最终样本包括61名说话者，每人各贡献一个程序性、一个描述性和一个叙事性样本（共183个样本）。剔除标记为 [+ exc]的话语后（3,338个话语中有364个被剔除），计算了每项特征和每种任务的描述性统计（表5）。由于大多数特征不符合正态分布，因此报告中位数。

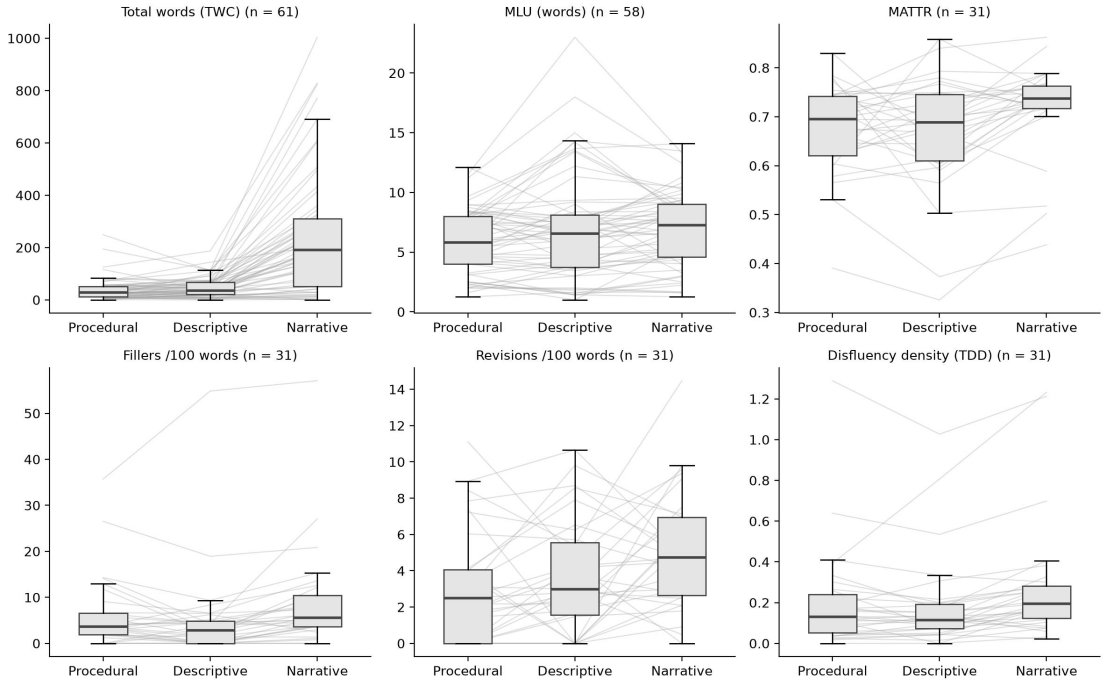
许多样本篇幅较短：27个程序性样本、21个描述性样本和9个叙事样本的词语少于25个。31位说话者在三项任务中均至少有25个词语，因此，不流畅率、总不流利密度和移动平均词汇类型比的分析均基于这31位说话者。因此，在这些分析中，产出极短样本的说话者（通常是失语症较严重者）代表性不足。58位说话者的数据可用于计算平均话语长度，因为有三位说话者的一项任务中没有任何包含词语的话语。在至少有25个词语的126个样本中，仅有11个样本标注了停顿（91%为零），一个样本标注了重新表述，没有样本标注截断。因此，这三项指标未进行检验。

表5按任务划分的描述性统计，中位数（四分位距）

特征	n	程序性（三明治）	描述性（破窗）	叙事性（灰姑娘）
总词数（TWC）	61	28（13–51）	35 (21–67)	191 (52–311)
NDW	61	19 (10–26)	22 (11–36)	81 (27–128)
MLU	58	5.80（4.00–8.00）	6.56（3.75–8.11）	7.27（4.58–9.02）
移动平均词汇类型比	31	0.70（0.62–0.74）	0.69（0.61–0.75）	0.74（0.72–0.76）
TDD	31	0.13（0.05–0.24）	0.11（0.07–0.19）	0.20（0.12–0.28）
每100词填充词数	31	3.61（1.89–6.59）	2.86（0.00–4.89）	5.56（3.61–10.43）
每100词的修订	31	2.50（0.00–4.06）	2.99（1.55–5.56）	4.73（2.64–6.94）
每100词的重复次数	31	0.00 (0.00–5.88)	1.85（0.00–6.38）	3.63（1.82–6.06）
每100词的不可理解词数	31	0.00（0.00–0.00）	0.00（0.00–0.00）	0.00（0.00–0.82）
每100词的音位片段数词语	31	1.82（0.00–5.26）	2.17（0.00–5.40）	3.78（2.04–5.27）

注。 n = 各项分析中的说话者数量。言语不流畅率、TDD 和移动平均词汇类型比仅纳入在三项任务中词数均至少为 25 的说话者。IQR = 四分位距；NDW = 不同词数；MLU = 平均话语长度；MATTR =移动平均形符类符比；TDD=总不流利密度。

图1三项任务中各说话者的数值



注。 每条线连接同一说话者的三个样本，箱线图显示中位数和四分位距。各面板展示总词数（TWC）、平均话语长度（MLU）、移动平均词汇类型比（MATTR）、填充词率、修正率和总言语不流畅密度（TDD）。MATTR、各项比率和TDD展示的是在每项任务中词数均至少为25的31位说话者的数据。

4.2 总体任务差异

表6展示了Friedman检验和Page检验的结果。这些是本研究的主要结果，因为它们直接检验任务是否会改变各项指标，以及这种变化是否符合预测顺序。

表6*Friedman* 和 *Page* 趋势检验结果（预测顺序：程序性 < 描述性 <叙事性）

特征	n	Friedman $\chi^2(2)$	q（错误发现率）	Kendall's W	Page 的 L	q（错误发现率）
总词数（TWC）	61	73.23	< .001*	0.60	821	< .001*
NDW	61	68.00	< .001*	0.56	817	< .001*
MLU	58	9.93	.010*	0.09	728	.003*
移动平均词汇类型比	31	9.55	.011*	0.15	392	.007*
TDD	31	19.93	< .001*	0.32	400	< .001*
每100词填充词数	31	22.50	< .001*	0.36	392	.007*
每100词的修订	31	8.69	.014*	0.14	394	.004*
每100词的重复次数	31	11.86	.004*	0.19	395	.003*
每100词的不可理解词数	31	12.12	.004*	0.20	386	.048*
音位片段词数	31	3.50	.174	0.06	383	.081

注。 q = 在分别对弗里德曼检验和Page检验进行Benjamini–Hochberg错误发现率（FDR）校正后的p值。* q < .05。NDW = 不同词数；MLU = 平均话语长度；MATTR = 移动平均形符类符比；TDD = 总不流利密度。

FDR校正后，弗里德曼检验显示，任务对10项特征中的9项有显著影响。影响最大的是总词数和NDW（W =0.60和0.56），这意味着大多数说话者在这些指标上对三项任务的排序相同。对存在显著任务效应的言语不流畅指标，影响程度中等（W = 0.14–0.36）；对MATTR和MLU的影响较小（W= 0.15和0.09）。Page趋势检验对10项特征中的9项也达到显著，但正如两两检验所示，其中大多数特征的趋势仅由叙事任务造成。唯一没有任务效应的特征是音位片段率（弗里德曼检验 q= .174）。

4.3 成对比较

成对Wilcoxon检验（表7）显示了任务差异所在的位置。

表 7成对威尔科克森符号秩检验：效应量 *r* *[95% 置信区间]* 和经错误发现率校正的 *q*

特征	叙事性 vs. 程序性	叙事性 vs. 描述性	描述性与 程序性
总词数 （TWC）	r = 0.99 [0.96, 1.00] q <.001*	r = 0.96 [0.89, 0.99] q <.001*	r = 0.37 [0.08, 0.63] q =.029*
NDW	r = 0.98 [0.94, 1.00] q <.001*	r = 0.95 [0.88, 0.99] q <.001*	r = 0.36 [0.07, 0.61] q =.029*
MLU	r = 0.58 [0.33, 0.79] q <.001*	r = 0.22 [−0.08, 0.53] q =.193	r = 0.24 [−0.06, 0.51] q =.165
移动平均词 汇类型比	r = 0.54 [0.18, 0.82] q =.020*	r = 0.65 [0.35, 0.89] q =.004*	r = −0.04 [−0.43, 0.38] q =.870
TDD	r = 0.69 [0.37, 0.93] q =.002*	r = 0.84 [0.60, 0.98] q <.001*	r = −0.22 [−0.61, 0.22] q =.353
每100词填充词数	r = 0.53 [0.17, 0.82] q =.020*	r = 0.89 [0.68, 1.00] q <.001*	r = −0.38 [−0.75, 0.01] q =.126
每100词的修订	r = 0.59 [0.24, 0.88] q =.013*	r = 0.32 [−0.08, 0.67] q =.172	r = 0.33 [−0.09, 0.76] q =.182
每100词的重复次数 词语	r = 0.50 [0.12, 0.80] q =.029*	r = 0.64 [0.31, 0.91] q =.007*	r = 0.23 [−0.25, 0.62] q =.382
每100词的不可理解词数 词语	r = 0.43 [−0.15, 1.00] q =.182	r = 0.72 [0.22, 1.00] q =.029*	r = −0.33 [−1.00, 0.79] q =.479
音位片段 /100 个词语	r = 0.18 [−0.22, 0.58] q =.409	r = 0.47 [0.07, 0.80] q =.041*	r = −0.23 [−0.66, 0.28] q =.394

注。 r = 匹配对秩双列相关；正值表示该组中的第一个任务数值较高。说话者经 2,000 次自助法重抽样得出的 95% 置信区间。q =经 Benjamini–Hochberg校正后、针对全部 30 项两两检验的 p值。* q < .05。CI = 置信区间；NDW = 不同词数；MLU =平均话语长度；MATTR =移动平均形符类符比；TDD = 总不流利密度。

最明显的对比存在于叙事性话语与程序性话语之间。与三明治制作程序相比，《灰姑娘》复述产生了更多词语（r = 0.99）、更高的不同词数（r = 0.98）、更长的话语（r = 0.58），并且每100个词语中有更多填充词（r = 0.53）、更多修正（r = 0.59）和更多重复（r = 0.50），以及更高的总不流利密度（r = 0.69）。这两项任务之间的不可理解语音率没有可靠差异（r = 0.43，q = .182）。

叙事性话语与描述性话语的大多数特征在比较时呈现出相同的方向。《灰姑娘》样本比破窗样本长（r = 0.96），且填充词率（r = 0.89）、重复率（r = 0.64）、不可理解言语率（r = 0.72）和音位片段率（r = 0.47）更高，总不流利密度也更高（r = 0.84）。这两项任务的平均话语长度和修正率没有显著差异（q =.193 和 .172）。

描述性话语与程序性话语之间的比较显示，差异最少。破窗样本的词语数和不同词数略高于三明治样本（r = 0.37 和 0.36），但两项任务在任何不受样本长度影响的指标上都没有差异。描述性任务中的填充词率若有差异，则反而较低（r = −0.38，q = .126）。

叙事性话语中的移动平均词汇类型比略高于程序性话语（r = 0.54，q = .020）和描述性话语（r = 0.65，q = .004），而描述性话语与程序性话语之间没有差异（r = −0.04）。因此，在控制样本长度后，尽管叙事复述长得多，其词汇多样性并未降低；如果说有变化，反而是略有提高。

4.4 哪些指标遵循复杂度梯度?

根据第3.5节中的判定规则，为每项特征分配了一种响应模式。结果汇总于表8。

表8各任务中语言学指标的响应模式

模式	定义	特征	解读
完整梯度	程序性 < 描述性 < 叙事性，所有配对显著	总词数、不同词数	言语量有所增加贯穿各项任务，但大部分增幅来自叙事任务
叙事性较高	叙事性 > 描述性 ≈ 程序性	每100词填充词数，每100词的重复次数，TDD、移动平均词汇类型比	每词不流利现象增加，而词汇多样性并未下降，当说话者必须从记忆中记忆
没有证据表明存在任务效应	Friedman 检验未达到显著水平	每100词的音位片段数	片段率并未随任务发生稳定变化
其他	任何其他模式	词语平均话语长度和每100词的修订（叙事性 > 仅程序性）；每100词的不可理解词数（叙事性 > 描述性仅）	叙事性中较高，但与两者相比并无明显差异其他任务
未测试	超过90%为零	停顿、重新表述、中断——offs	在这些转录文本中编码频率过低转录文本

注。 NDW = 不同词数；MLU =平均话语长度；MATTR = 移动平均型符比；TDD = 总不流利密度。

完整梯度。 从程序性话语到描述性话语，再到叙事性话语，总词数和不同词数均有所增加，且三个步骤均达到显著水平。这些测量结果支持了关于输出和词汇的预测。然而，该梯度非常不均衡：这一步

从三明治到破窗的变化很小（r = 0.37 和 0.36），而到灰姑娘的变化则非常大（r ≥ 0.95）。平均话语长度并未遵循完整梯度。叙事性话语中的平均话语长度高于程序性话语，但描述性任务与二者之间的差异均未达到可靠水平。因此，对于输出和词汇的预测，总词数和不同词数得到了支持，但平均话语长度未得到支持。

叙事性较高。 叙事性话语中的填充词率、重复率和总不流利密度均高于另外两种任务，但描述性话语与程序性话语之间没有差异。由于这些指标均按词数计算，叙事性话语中的升高并非仅仅是样本较长所致。填充词率支持了关于言语不流畅和言语中断的预测。该预测中的另外两个指标则呈现部分支持：叙事性话语的修正率仅高于程序性话语，而不可理解语音率仅高于描述性话语。因此，言语不流畅预测中的三个指标有一个得到完全支持，另外两个则得到部分支持。移动平均词汇类型比也符合叙事性较高的判定规则，但效应较小。

没有任务效应的证据。 音位片段率总体上未随任务发生显著变化，不过在两两比较中，叙事性话语的音位片段率高于描述性话语。

4.5 机器学习验证

随机森林分类器经过训练，可根据语言特征识别任务。其表现见表9。

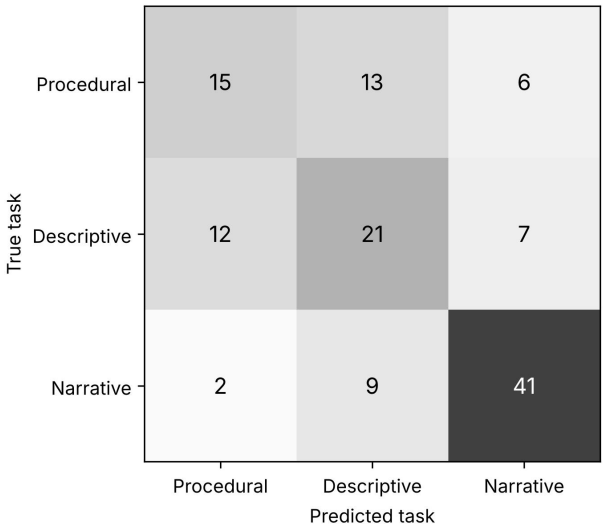
表9话语任务的随机森林分类（按说话者分组的*10*折交叉验证）

Task	精确率	召回率	F1分数	支持度
程序性（三明治）	0.52	0.44	0.48	34
描述性（破窗）	0.49	0.53	0.51	40
叙事性（灰姑娘）	0.76	0.79	0.77	52
总体准确率			0.61	126
宏平均	0.59	0.58	0.59	126

注。 特征：总词数、不同词数、话语平均长度、总不流利密度，以及每100个词中的五项不流畅率。少于25个词的样本将被排除（来自52位说话者的126个样本）。机会水平宏平均F1约为0.33；如果始终猜测叙事性，准确率将为0.41。

该分类器识别任务的准确率为 0.61，宏平均 F1 为 0.59。叙事样本的识别效果较好（52 个中识别出 41 个），而程序性样本和描述性样本经常彼此混淆：34 个程序性样本中有 13 个被标记为描述性，40 个描述性样本中有 12 个被标记为程序性（图 2）。程序性样本与叙事样本之间的混淆较少：34 个程序性样本中有 6 个被标记为叙事性，52 个叙事样本中有 2 个被标记为程序性。去除总词数和不同词数后，准确率降至 0.52，宏平均 F1 降至 0.48。叙事样本仍能被识别（召回率为 0.79），但程序性样本和描述性样本的分类表现接近随机水平（召回率分别为 0.35 和 0.33）。研究未进行与随机水平的正式检验，因此这些结果仅作描述。

图 2所有特征的随机森林分类器混淆矩阵



注。 行表示真实任务，列表示预测任务。126 个样本中的每一个都由一个在训练期间未见过该说话者的模型进行预测 ed 由一个在训练期间未见过该说话者的模型。

该分类器并非本研究问题的主要检验手段。它的价值在于同时考察所有特征，并且是在模型未见过的说话者上进行评估的。其模式与统计结果一致：即使去除言语量，叙事任务仍呈现出清晰的语言学特征，而描述性任务和程序性任务则难以区分。

5. 讨论

本研究考察了话语任务的选择是否会改变从失语症话语中提取的语言学指标。在同一批说话者中，确实会。言语量沿着完整的复杂度梯度变化，但大多数指标都将叙事任务与另外两种任务区分开来，而

描述性任务和程序性任务在任何独立于样本长度的指标上都没有差异。

5.1 发现1：话语任务不可互换

主要发现是，话语任务不可互换。同一说话者在同一会话中录制的语料，其大多数测量指标会因任务不同而呈现系统性差异。因此，说话者的总词数、不同词数或填充词率并非该说话者的固定属性，而是说话者与任务共同作用的产物。

这一结果与早期研究一致：这些研究表明，话语类型会影响失语症患者通过 CLAN 提取的测量指标（Stark, 2019）；失语症患者和非失语症患者的话语具有体裁特异性微观结构（Stark & Fukuyama, 2021）；同一说话者的图片描述与对话之间仅呈中等或较弱的一致性（Leaman & Edmonds, 2023）；此外，在临床组中，任务效应还会影响近300项自动提取的特征（Stark & Themistocleous, 2026）。本研究在三方面拓展了这些工作：研究设计确保每位说话者都完成全部三项任务；采用自动提取的言语不流畅标记，而这类标记正越来越多地用于自动化评估；检验了任务要求的有序梯度，而不仅仅是总体差异。此外，本研究还表明，即使按词数计算言语不流畅现象，任务对其产生的效应依然存在。分类器得出了相同的趋势，并且对叙事样本的识别最为可靠。

5.2 为什么叙事性话语与程序性话语存在根本差异

差异最大且最一致的是《灰姑娘》复述与三明治制作程序之间的差异，这两项任务至少在四个方面有所不同。第一，移除书本后，必须从记忆中召回灰姑娘故事，说话者还必须在逐集讲述的过程中保持其宏观结构（Kintsch & van Dijk，1978）；而三明治任务要求讲述一种大多数成年人已经做过许多次的简短日常程序。第二，程序具有固定的步骤顺序，每个步骤都可以用简短的名词短语来表达，例如“面包”。这会形成简单、名词密集型的表达以及较低的词汇多样性（Fergadiotis 等人，2011；Stark & Fukuyama，2021）。故事则要求说话者追踪角色、标示事件顺序并表达因果关系，也需要使用更多过去时形式（Riccardi 等，2024）。第三，故事鼓励说话者进行

较长的复述；对于典型说话者而言，言语不流畅度会随着话语长度增加以及结构化程度降低而上升（Oviatt，1995）。失语症患者的叙事表现与执行功能有关（Dutta等人，2024）。因此，《灰姑娘》复述并不只是“更多语言”：它同时调动记忆、规划和监控，这正是它产生内容最丰富且最不流利的话语的原因。

5.3 发现二：叙事性即使按每个词计算也独树一帜

第二个发现是，任务效应并不是平滑的梯度。只有言语量（总词数和不同词数）逐步上升，而且即使在这方面，从三明治制作程序到破窗的增幅也很小。在按每个词计算的指标上，描述性任务和程序性任务没有差异，而叙事任务则更高。有三点有助于解释这一模式。

言语量遵循任务的内容。 总词数和不同词数反映了需要表达的内容有多少。破窗图片展示了一段简短的事件序列，因此可描述的内容比三明治制作程序略多，但远少于完整的童话。平均话语长度只从程序性任务到叙事任务有所上升，因此话语长度并未随着假定的任务要求而稳步增加。

当必须从记忆中召回故事时，每词不流利现象会增加。 叙事任务中的填充词、重复和总不流利密度均高于另外两个任务，按每词计算时也是如此。填充词反映了规划延迟（Clark & Fox Tree，2002）。一种解释是，灰姑娘任务是唯一要求说话者在没有图片的情况下从记忆中提取内容的任务。在破窗任务中，图片充当外部记忆；而在本语料库中，三明治任务配有三明治照片（第 3.2 节）。因此，两个不存在差异的任务也正是两个具有视觉支持的任务，这一结果更符合记忆支持解释，而不是单一的需求梯度。不过，一张三明治照片所提供的支持少于图片序列，因此应直接检验这一解释。该解释与 Culbertson（2022）的发现一致：与图片序列描述相比，灰姑娘故事中出现了更多重复和回溯。健康成人中也有类似效应：与根据图片书讲故事相比，当他们必须自行组织回答时，每词的不流利现象更多（Walters 等人，2026）。叙事性话语中的修正率高于程序性话语，但并未可靠地高于描述性话语；不可理解言语

仅在叙事性话语中高于描述性话语，因此，自我修正和言语中断可能不像犹豫那样与任务的记忆负荷密切相关。还应考虑一种更简单的解释：两个较短任务中的不流畅率较低，而较低的计数使得差异难以被检测出来。

较长的复述中，词汇多样性并未下降。 不同词数（NDW）在本研究中表现出最大的任务差异之一，而用于独立于长度衡量同一属性的移动平均词汇类型比（MATTR）仅显示出较小的效应，方向上体现为叙事性话语中的词汇多样性略高。因此，叙事性话语在不同词数上的显著优势主要是长度效应，而这正是移动平均词汇类型比旨在解决的问题（Covington & McFall, 2010年; Fergadiotis 等人, 2015年）。与此同时，移动平均词汇类型比也并非不受任务影响：复述的词汇多样性略高，这可能是因为包含多个角色和事件的故事，比简短的程序说明或图片序列需要更多样的词汇。

5.4 发现 3：未来话语研究的方法学启示

这些结果对从失语症话语中提取语言学指标的计算研究和临床研究具有直接启示。

话语任务报告。 未报告任务的研究，其结果难以进行比较。口语话语研究的报告指南已要求作者详细描述诱发程序（Stark 等人, 2023年）。本研究结果说明了这一要求为何至关重要：同一指标在不同任务中可能取值迥异。因此，每项特征和每项分析都应报告所用任务。

跨任务汇集样本。 将不同任务的样本合并为一个特征向量，或在一项研究中提取某项任务的特征，并与另一项研究中另一项任务的相同特征进行比较，都会混淆任务与说话者的影响。Bryant 等人（2016年）提醒，将不同体裁合并为单一样本，可能会使某一体裁中的困难被另一体裁中较轻松的表现掩盖。当使用多项任务时，应将任务纳入模型，例如作为因子或分层变量。

任务特异性参考值。 由于这些数值取决于任务，因此应针对每项任务分别设定参考范围和截断值。针对失语症口语话语评估的标准化工作，已在推动制定共享方案和结果指标（Stark 等人，2021年）。

长度校正。 总词数和不同词数等原始计数在很大程度上反映了说话者的发言量，而移动平均词汇类型比则消除了大部分任务效应对词汇的影响。然而，按词数计算的不流畅率在叙事任务与另外两项任务之间仍存在差异。因此，与原始计数相比，长度校正后的指标更适合用于跨任务比较，但尽可能仍应在同一任务内进行比较。原始计数仍然有用，因为一个人能够说出多少话本身就具有临床意义。

机器学习中的任务效应。 仅凭特征就能识别任务，尤其是叙事样本。如果不同组的说话者完成了不同的任务，模型因此可能学到的是任务，而非组别。在将模型描述为具有普遍适用性之前，应先用其他任务对在单一任务上训练的模型进行测试，并将同一说话者的所有样本保留在同一折中（Saeb等人，2017）。

随时间保持任务一致。 在纵向研究和治疗试验中，评估之间更换任务可能会被误认为语言能力发生了变化。每个时间点都应使用相同的任务。重复使用相同的刺激材料也有其风险，因为说话者在多次评估后可能会熟悉某个故事（Gupta等人，2025）。如果存在这种顾虑，选择同一体裁的平行刺激材料可能比更换体裁更合适。

5.5 发现4：临床意义

本研究的临床启示很简单。使用失语症话语的研究人员和临床医生应仔细控制引出任务，因为任务选择本身会影响这些测量指标的分布。

在实践中，应根据所要回答的问题以及当事人自身的沟通目标来选择任务（Leaman & Archer, 2023）。灰姑娘等叙事复述能产生最多的语言，也最能敏锐地反映规划和流利度方面的言语中断，但对受试者的要求也最高。三明治等程序性任务篇幅较短、内容熟悉，对于重度失语症患者可能更容易完成，但产生的语言有限且简单。在本样本中，破窗图片序列的单词级数值与三明治制作程序相近，但言语量略多。这些观察结果与 Stark 和 Fukuyama（2021）的建议一致：若目标是全面了解一个人的形态句法能力，叙事任务可能更为合适。在自动化

重复测试中，叙事性话语在极轻度失语症患者身上也能提供最可靠的流利度指标，而程序性任务提供的指标最不可靠（Stark 等人，2026）。这些结果也符合当前的临床指南：结构化任务有助于在不同时间从同一个人身上 elic出相似的样本；程序性话语可能很适合用于测量语速，因为其节奏通常比叙述更快（Dutta 等人，2025）。

当使用话语随时间监测某人时，每次评估都应使用相同的任务，并将该人的数值与来自相同任务的参考数据进行比较。《灰姑娘》复述中的填充词率高于较早进行的三明治制作程序，并不意味着该人的流利度下降了。

5.6 局限性

主要局限性如下。

- **固定任务顺序**。 所有说话者均在一次会话中按相同顺序完成三项任务（破窗、灰姑娘、三明治），因此无法完全排除疲劳或练习等顺序效应。不过，如果存在稳定的顺序效应，这些指标应会从第一项任务到最后一项任务逐渐变化。实际情况则是，处于中间位置的叙事任务表现突出，而第一项和最后一项任务在任何独立于样本长度的指标上都没有差异。

- **样本较短**。 逐词分析基于在每项任务中都至少产出25个词的31名说话者。因此，样本非常短的说话者（通常是失语症更严重者）代表性不足，对于任何单一言语不流畅指标的结果都应谨慎解读。

- **假定的任务要求**。程序性 < 描述性 < 叙事性的排序基于理论，并未经过测量。这些任务还同时存在若干差异，包括熟悉度、视觉支持和内容量。

- **对转录的依赖**。这些测量依赖人工转录员标注的不流利编码，而停顿的编码过少，无法进行分析。

- **推广性**。 数据来自一个英语成年人慢性卒中后失语症语料库，而灰姑娘故事和三明治任务都具有文化特异性（Dutta等人，2025年）。

5.7 未来方向

未来研究应在更大、更多样化的样本中重复这些发现，包括其他 AphasiaBank 语料库以及没有失语症的说话者。早期研究发现，有失语症和无失语症的说话者都存在任务效应（Stark，2019；Stark &Themistocleous，2026），但此处发现的模式是否为失语症所特有，仍有待探究。AphasiaBank 协议中的其他任务，例如救猫和拒绝雨伞图片任务以及个人叙事，可以检验叙事与其他任务的模式是否适用于更广泛的体裁。直接比较有图片和无图片时相同程序的表现，可以检验视觉支持是否解释了描述性和程序性任务表现相似的原因。以说话者作为随机效应的混合效应计数模型可以纳入样本较短的说话者，并检验任务效应如何与失语症类型和严重程度相互作用，因为失语症越严重，不同任务中的语言越相似（Stark & Fukuyama，2021年）。最后，随着自动语音识别日益取代人工转录，检验从自动生成的转录文本中提取特征时是否也会出现相同的任务效应，将变得十分重要。目前这还不是一件容易的事。近期一项基准测试对四种自动语音识别系统在失语症患者语音上的表现进行了评估，其中表现最好的模型的词错误率中位数约为21%，且四种系统都对动词识别不足（Kollapally与Akers，2025年）。针对典型自发言语，研究者已开发出将言语不流畅现象检测与语音识别相结合的模型（Futami等人，2023年），但这些模型尚未在失语症语音上得到验证。

6. 结论

本研究表明，话语任务选择会影响从失语症话语中提取的语言学指标。在同一组说话者中，从程序性任务到描述性任务再到叙事性任务，言语量逐渐增加，但大多数指标都能将故事复述与另外两项任务区分开来。叙事性任务中的填充词率、重复率和总不流利密度均高于另外两项任务，即使按词数计算也是如此；与此同时，长度校正词汇多样性并未下降，反而略有升高。机器学习分类器能够较好地识别叙事样本，但经常将描述性样本与程序性样本混淆。

这些发现表明，话语任务是一项重要的方法学变量，在未来的计算研究和临床研究中应明确报告并加以控制

这些研究针对失语症；此外，对样本长度进行校正并不能使不同任务所得的指标可以互换。

声明

致谢

作者感谢 Julius Fridriksson 及其同事将 Fridriksson-2语料库贡献给 AphasiaBank，并感谢参与 POLAR试验的失语症患者。

资金支持

本研究未获得任何资金支持。

利益披露

作者声明不存在需要披露的竞争性利益。

生成式人工智能使用声明

人工智能助手（ChatGPT、Claude 和 Gemini）被用作辅助工具，用于文献检索和参考文献核查、审计代码中潜在的运行时、实现及统计问题，以及改进语言表达。研究问题、代码、分析、研究结果、诠释和手稿均由作者撰写。作者对所有人工智能生成的建议进行了严格审查，并对本文的准确率、内容和结论承担全部责任。

数据可用性声明

本研究分析的转录文本属于 Fridriksson-2语料库（Fridriksson，2023年；https://doi.org/10.21415/E8FQ-ZY70）。注册用户可根据 TalkBank关于数据访问和使用的规则获取这些文本（https://aphasia.talkbank.org）；这些规则不允许作者分享这些文本。

代码可用性

本研究所使用的源代码可在 https://github.com/emmaalkubaisi-star/task-dependent-discourse-measures 获取。其中包括转录文本预处理、基于正则表达式的特征提取、统计分析和机器学习检验。

参考文献

Benjamini, Y.，& Hochberg, Y.（1995）。控制错误发现率：一种实用且强大的多重检验方法。英国皇家统计学会杂志：*B*辑（方法论），*57*(1)，289–300。https://doi.org/10.1111/j.2517-6161.1995.tb02031.x Braun, A.、Elsässer, N.，& Willems, L.（2023）。重新审视言语不流畅现象——它们是否具有说话者特异性？*Languages*，*8*(3)，155。https://doi.org/10.3390/languages8030155 Breiman, L.（2001）。随机森林。*Machine Learning*，*45*(1)，5–32。https://doi.org/10.1023/A:1010933404324 L. Bryant、Ferguson, A.，& Spencer, E.（2016）。失语症话语的语言学分析：文献综述。*Clinical Linguistics & Phonetics*，*30*(7)，489–518。https://doi.org/10.3109/02699206.2016.1145740 Clark, H. H.，& Fox Tree, J. E.（2002）。在自发言语中使用*uh* 和*um* 。*Cognition*，*84*(1)，73–111。https://doi.org/10.1016/S0010-0277(02)00017-3 Covington, M. A.，& McFall, J. D.（2010）。解开戈耳狄俄斯之结：移动平均型符比（MATTR）。*Journal of Quantitative Linguistics*，*17*(2)，94–100。https://doi.org/10.1080/09296171003643098 Culbertson, E. A.（2022）。跨话语类型的失语症重复与回溯[硕士学位论文，密西西比大学]。eGrove。https://egrove.olemiss.edu/etd/2207/ Deng, B.-M.、Gao, J.、Liang, L.-S.、Zhao, J.-X.、Lin, F.、Yin, M.-Y.、Zheng, H.-Q.，& Hu, X.-Q.（2024）。命名性失语与健康对照组中的话语任务类型特异性语言特征：来自普通话中国失语症数据库的证据。《美国言语-语言病理学杂志》，*33*(2)，937–951. https://doi.org/10.1044/2023_AJSLP-23-00078 D. B. den Ouden、Giglio、Kristinsson、Bonilha、Schwen Blackett、Stark, B. C.、Wilmskoetter、& Fridriksson (2026)。失语症患者对命名干预反应的叙事性话语预测指标。《言语、语言与听力研究杂志》, *69*(7), 3272–3289. https://doi.org/10.1044/2026_JSLHR-25-00499 Didirková, I. (2024)。言语和语言障碍中的言语不流畅 [社论]。诊所语言学与语音学，*38*(4), 285–286. https://doi.org/10.1080/02699206.2023.2277122 M. Dutta、L. L. Murray、Park H.、Burklow E.、Bose A.、Kim H.、Greenslade K.、A. E. Ramage、Combs C.、Balasubramanian A. 和 Casilio M.（2025）。聊聊口语话语：在提供失语症临床服务时使用口语话语分析的教程。《美国期刊》

言语-语言病理学。在线先行出版。
https://doi.org/10.1044/2025_AJSLP-24-00534
Dutta, M., Murray, L. L., & Stark, B. C. (2024). 失语症中的执行功能与叙事语言能力之间的关系。《美国言语-语言病理学杂志》, *33*(5), 2500–2523. https://doi.org/10.1044/2024_AJSLP-23-00314 Fergadiotis, G., Wright, H. H., & Capilouto, G. J. (2011). 不同话语类型中的产出性词汇。《失语症学》, *25*(10), 1261–1278. https://doi.org/10.1080/02687038.2011.606974 Fergadiotis, G., Wright, H. H., & Green, S.B. (2015). 词汇多样性指标的心理测量学评估：考察长度效应。《言语、语言与听力研究杂志》, *58*(3), 840–852. https://doi.org/10.1044/2015_JSLHR-L-14-0280 Fridriksson, J. (2023). *Fridriksson-2*语料库 [数据集]。AphasiaBank, TalkBank. https://doi.org/10.21415/E8FQ-ZY70 Friedman, M. (1937). 使用秩避免方差分析中隐含的正态性假设。《美国统计学会杂志》, *32*(200), 675–701. https://doi.org/10.1080/01621459.1937.10503522

Fromm, D., Chern, S., Geng, Z., Kim, M., Greenhouse, J., & MacWhinney, B.(2024). 失语症患者流畅性行为的自动分析。《言语、语言与听力研究杂志》, *67*(7), 2333–2342. https://doi.org/10.1044/2024_JSLHR-23-00659

Futami, H., Tsunoo, E., Shibata, K., Kashiwagi, Y., Okuda, T., Arora, S.,& Watanabe, S.(2023). 流式联合语音识别与言语不流畅检测。载于*ICASSP 2023 – 2023 IEEE*声学、语音与信号处理国际会议 (第 1–5 页)。IEEE。https://doi.org/10.1109/ICASSP49357.2023.10094620

Gupta, A., Hudspeth, M., Stokes, P., Kurland, J.,& O'Connor, B. (2025年). 失语症叙事话语评估的自动主要概念生成。载于计算语言学协会年会论文集: *ACL 2025* （第24437–24451页）。计算语言学协会。Jozipović, M., Došen, A., & Hržica, G. (2026年). 克罗地亚语失语症患者与典型说话者口语话语中的词汇多样性和词汇错误：不同话语类型的比较。*ClinicalLinguistics & Phonetics,40*（7）， 670–692。https://doi.org/10.1080/02699206.2025.2585109 Kendall, M. G., & Babington Smith, B. (1939). *m* 排名问题。《数学统计年刊》， *10*（3）， 275–287。https://doi.org/10.1214/aoms/1177732186

Kintsch, W., & van Dijk, T. A. (1978). 论文本理解与生成模型。《心理学评论》，*85*（5），363–394。https://doi.org/10.1037/0033-295X.85.5.363 Kollapally, N. M., & Akers, C. (2025年). 失语症自动语音识别的基准测试：临床评估框架。载于*2025*年*IEEE*生物信息学与生物医学国际会议（*BIBM*）（第7599–7607页）。IEEE。https://doi.org/10.1109/BIBM66473.2025.11356945 Kristinsson, S., Basilakos, A., Elm, J., Spell, L. A., Bonilha, L., Rorden, C., D. B. den Ouden, Cassarly, C., Sen, S., Hillis, A. E., Hickok, G., & Fridriksson, J. (2021年). 卒中后语义性与语音性失语症治疗的个体化反应。《脑通讯》，*3*（3），fcab174。https://doi.org/10.1093/braincomms/fcab174 Kristinsson, S., Basilakos, A., D. B. den Ouden, Cassarly, C., Spell, L. A., Bonilha, L., Rorden, C., Hillis, A. E., Hickok, G., Johnson, L., Busby, N., Walker, G. M., McLain, A., & Fridriksson, J. (2023年). 预测语言康复结果：失语症治疗后近期和长期结果的预后因素。《言语、语言与听力研究杂志》，*66*（3），1068–1084。https://doi.org/10.1044/2022_JSLHR-22-00347 Leaman, M. C., & Archer, B. (2023年). 选择与以人为中心的失语症康复目标相一致的话语类型：临床教程。美国言语-语言-听力协会特别兴趣组的观点，*8*（2），254–273。https://doi.org/10.1044/2023_PERSP-22-00160 Leaman, M. C., & Edmonds, L. A. (2023年). 分析野餐场景图片和对话中的语言：我们选择的话语样本类型会影响对失语症患者的研究发现。《美国言语-语言病理学期刊》，*32*（4），1413–1430。https://doi.org/10.1044/2023_AJSLP-22-00279 Lee, S. H., Wang, S., Varkanitsa, M., & Kiran, S. (2026年). 用于跨诊断检测语言障碍的自动宏观语言学话语分析[预印本]。medRxiv。https://doi.org/10.64898/2026.05.19.26353614 MacWhinney, B. (2000). *CHILDES*项目：分析言语的工具（第3版）。Lawrence Erlbaum Associates。MacWhinney, B., Fromm, D., Forbes, M., & Holland, A. (2011). AphasiaBank：话语研究方法。《失语症学》，*25*（11），1286–1307。https://doi.org/10.1080/02687038.2011.589893 Oviatt, S. (1995). 预测人机交互过程中的口语不流畅现象。《计算机语音与语言》，*9*（1），19–35。https://doi.org/10.1006/csla.1995.0002

Page, E. B.（1963）。多种处理的有序假设：线性秩的显著性检验。《美国统计学会杂志》，*58*(301)，216–230。https://doi.org/10.1080/01621459.1963.10500843 Pedregosa, F.、Varoquaux, G.、Gramfort, A.、Michel, V.、Thirion, B.、Grisel, O.、Blondel, M.、Prettenhofer, P.、Weiss, R.、Dubourg, V.、Vanderplas, J.、Passos, A.、Cournapeau, D.、Brucher, M.、Perrot, M. 和 Duchesnay, É.（2011）。Scikit-learn：Python 中的机器学习。机器学习研究期刊，*12*，2825–2830。Penttilä, N.、Korpijaakko-Huuhka, A.-M. 和 Bóna, J.（2022）。典型与非典型芬兰成人言语中的不流畅簇：一项先导研究。临床语言学 &

语音学，*36*(1)，1–16。https://doi.org/10.1080/02699206.2021.1924861 Pittman, J. M.、Medina-Santos, Y.、Phillips, A., Jr. 和 B. C. Stark（2025）。用于失语症话语分析的实用机器学习（arXiv:2511.17553）。arXiv。https://arxiv.org/abs/2511.17553 Riccardi, N.、Nelakuditi, S.、D. B. den Ouden、Rorden, C.、Fridriksson, J. 和 Desai, R. H.（2024）。使用机器学习进行基于话语和病灶的失语商估计。神经影像：临床，*42*，103602。https://doi.org/10.1016/j.nicl.2024.103602 Saeb, S.、Lonini, L.、Jayaraman, A.、Mohr, D. C. 和 Kording, K. P.（2017）。临床机器学习需要近似实际使用场景。*GigaScience*，*6*(5)，gix019。https://doi.org/10.1093/gigascience/gix019 Stark, B. C.（2019）。失语症患者与年龄匹配成人的三种话语引出方法比较：对语言评估和结果的启示。《美国言语-语言病理学杂志》，*28*(3)，1067–1083。https://doi.org/10.1044/2019_AJSLP-18-0265 Stark, B. C.、Bryant、Themistocleous、den Ouden, D.-B. 和 Roberts, A. C.（2023）。失语症和神经源性沟通障碍中口语话语报告的最佳实践指南。《失语症学》，*37*(5)，761–784。https://doi.org/10.1080/02687038.2022.2039372 Stark, B. C. 和 Dalton, S. G.（2024）。关于失语症话语分析无转写实践的范围综述及其对未来研究的启示。国际语言与沟通障碍杂志，*59*(5)，1734 –

1762. https://doi.org/10.1111/1460-6984.13028 Stark, B. C., Dutta, M., Murray, L. L., Bryant, L., Fromm, D., MacWhinney, B., Ramage, A. E., Roberts, A., den Ouden, D. B., Brock, K., McKinney-Bock, K., E. J. Paek, T. G. Harmon, S. O. Yoon, C. Themistocleous, H. Yoo, K. Aveni, S. Gutierrez, & S. Sharma.（2021）。失语症口语话语评估标准化：一个设有交付成果的工作组。《美国言语-语言病理学杂志》，*30*(1S), 491–502. https://doi.org/10.1044/2020_AJSLP-19-00093

Stark, B. C., Meinert, K., Fromm, D., & MacWhinney, B.（2026年7月1–3日）。隐匿性失语症中的话语流利度中断：跨诱发情境的混合效应分析 [海报报告]。国际失语症康复会议，希腊雅典。Stark, B. C., & Fukuyama, J.（2021）。利用大数据理解有无失语症的说话者在独白式口语话语中任务与语言的交互作用。《语言、认知与神经科学》，*36*(5), 562–585. https://doi.org/10.1080/23273798.2020.1862258 Stark, B. C., & Themistocleous, C.（2026）。临床语言常模数据与任务特异性效应。《脑与语言》，*283*, 105850. https://doi.org/10.1016/j.bandl.2026.105850 Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., va n der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., ... SciPy 1.0 贡献者.（2020）。SciPy 1.0：Python 科学计算的基础算法。《自然·方法》，*17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2 Walters, K., Nakamura, M.S., Cosgrove, A. L., Zhang, H., Troutman, S. B. W., Gertel, V. H., & Diaz, M. T.（2026）。年龄和话语任务会独立影响非流利现象产出。《言语、语言与听力研究杂志》，*69*(10), 4758–4776. https://doi.org/10.1044/2026_JSLHR-25-00799 Wilcoxon, F.（1945）。采用秩次法进行个体比较。《生物统计学通报》，*1*(6), 80–83. https://doi.org/10.2307/3001968