

## 开发缘由:

- 我想拥有一个自己的拼音输入法，开源的，可以高度定制，没有强制弹窗、没有强制升级。
- 我想拥有一个自己的拼音输入法，拥有超高的词语输入准确率，摆脱对过度商业化输入法的依赖。
- 我想拥有一个自己的拼音输入法，可以众人参与改进，吸收最新的语言学成果。
- 我想拥有一个自己的拼音输入法，如同知名媒体人王晓峰的博文《[一个输入法的死掉](#)》描述的黑马神拼那样，可以在输入法中对古诗词信手拈来。
- Rime输入法有一个超大词库，[【SuperRime拓展词库】for 朗月拼音 & Win10拼音 \(700万词\)](#)，质量并不高，绝大部分都是未完全分词的错误词汇。[朗月拼音码表](#)中存在不少拼音错误。目前Rime输入法缺乏一个接近工业级质量的拼音库码表，如果词库必须要自己去养，现代汉语词典就有大约6万个词语，不吸收利用现有的语言频率成果，养词汇要等到猴年马月。
- [刘邵博综合多本词典整合的一个大词典](#)，词典共有词汇3669216个词汇。该词典未对词语进行有效筛选，虽然来源样本较大，是270G新闻语料，但是不具备典型代表性，不能囊括其他语料库的词语。同时这个词典没有拼音标注。

## 开发理念:

不以盈利为目的，本着开源共享的精神，使用网络上可以公开获得的数据，打造一个高准确率的拼音输入法，免除弹窗、捆绑安装、强制升级的烦恼。

商业化的输入法有经济利润的驱动，投入大量的人力，拥有较高的词库质量。当商业化倾向过于严重时，会影响用户体验。闭源的数据和代码，让一般民众无法参与到产品核心功能的改进，无法吸纳群体的智慧。

开源的和免费的输入法属于兴趣驱动，人力投入匮乏，良莠不齐，缺乏高质量的词库和功能体验。

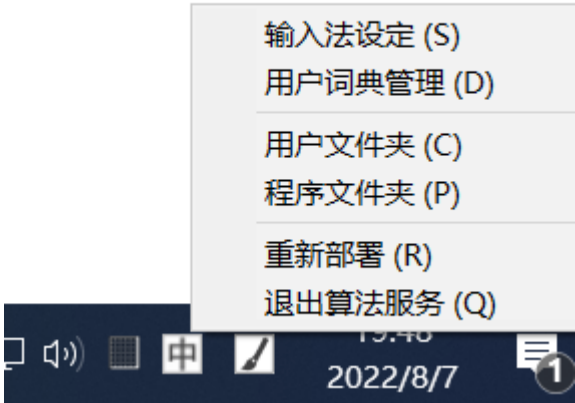
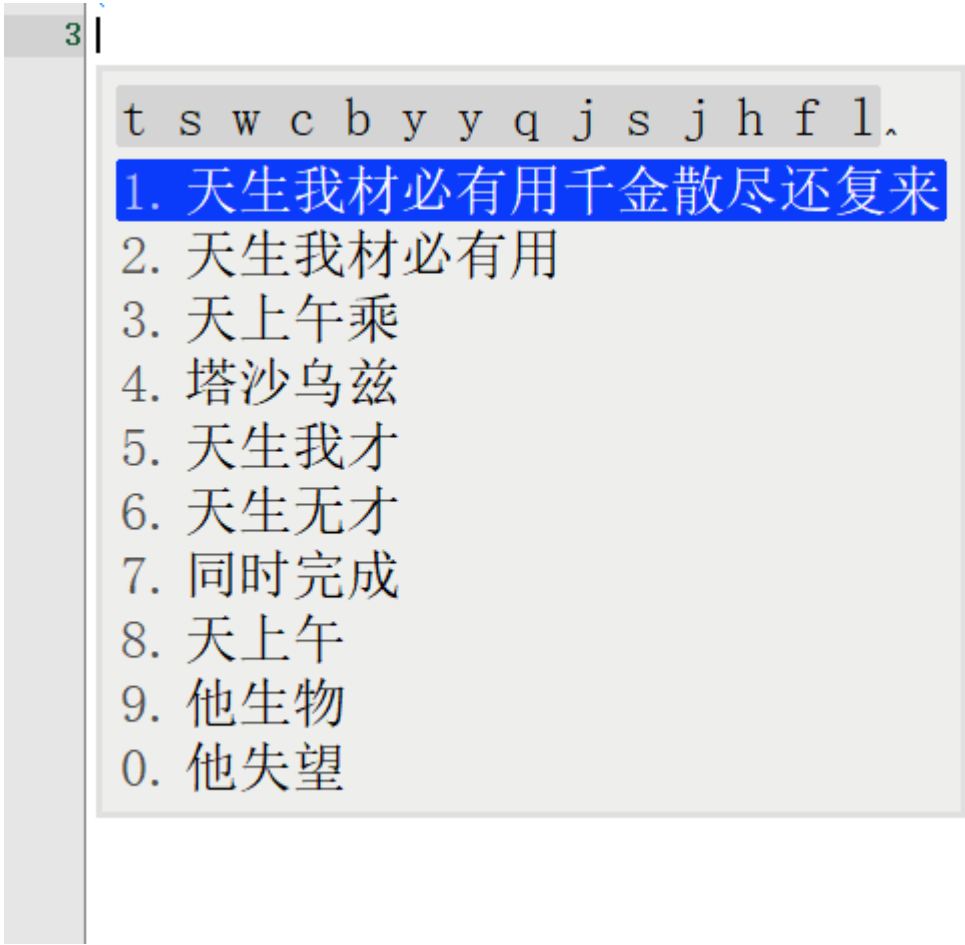
大学研究人员对于汉语词频、拼音、分词的学术性研究，拥有科研基金的支持，有高水平人才的参与，学术成果拥有较高的质量，但研究者没有将学术成果转化为实用性较强的拼音输入法倾向。

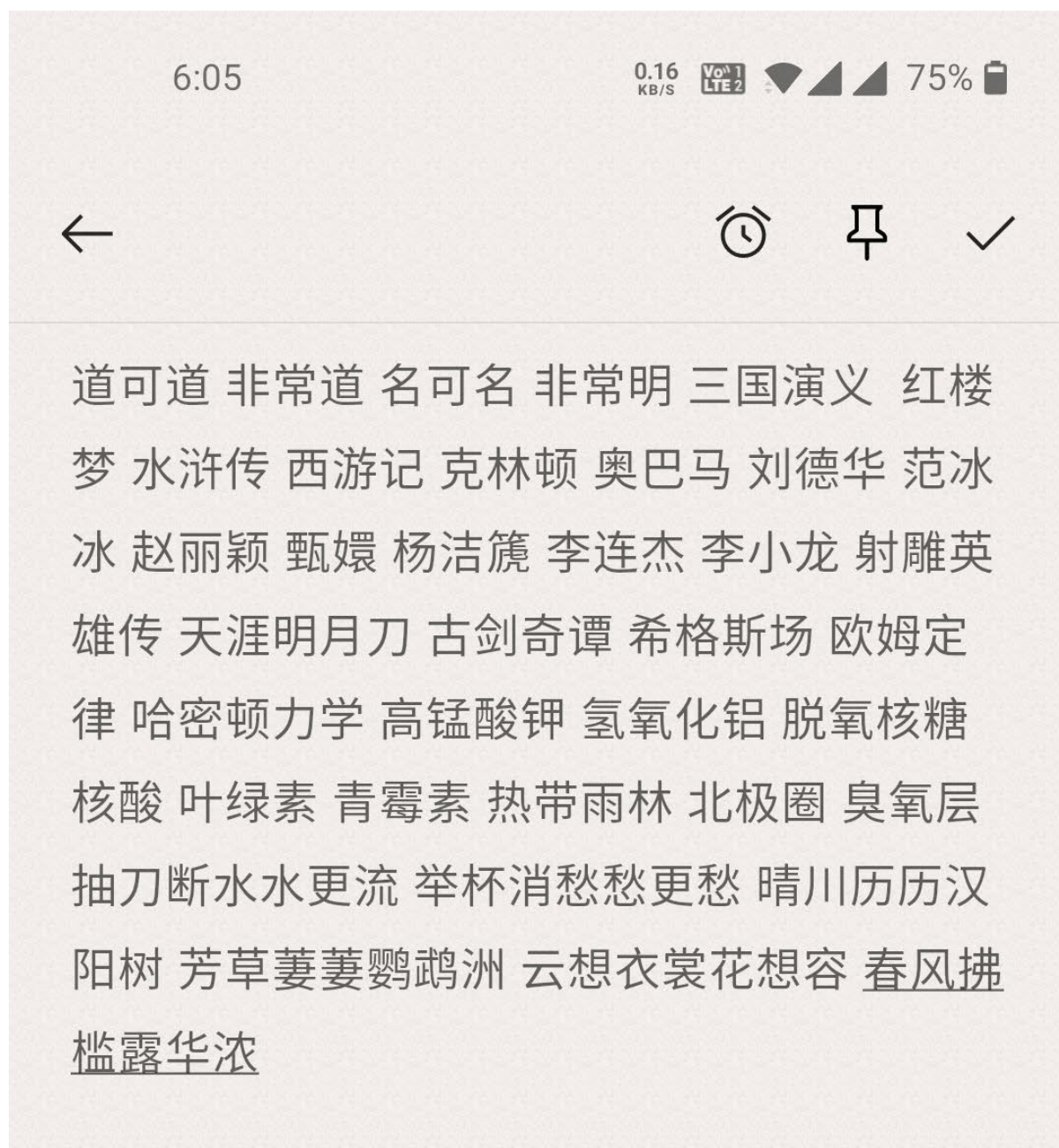
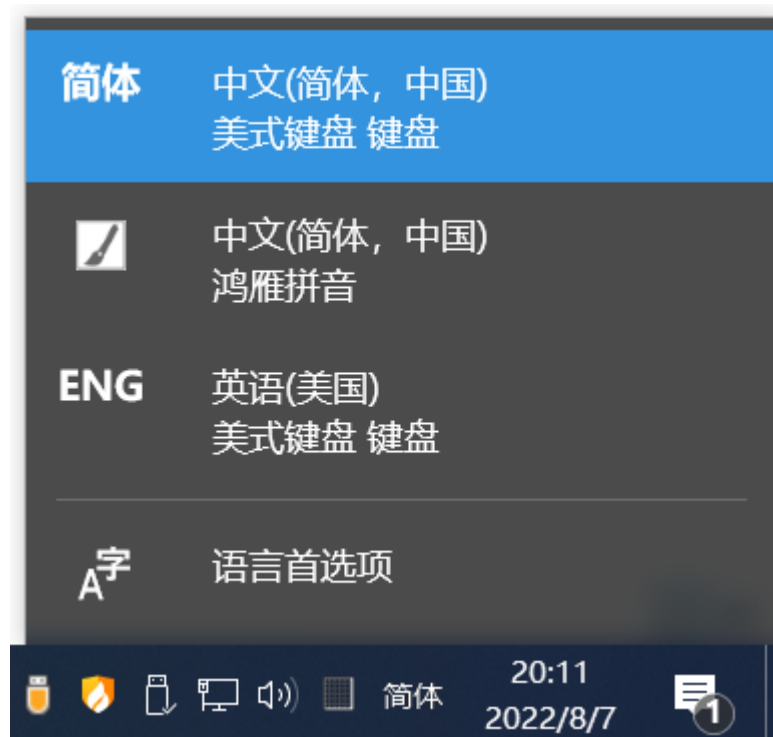
鱼与熊掌不可兼得，综合吸纳了商业化、开源化、学术化的产品三方优点，鸿雁拼音输入法诞生了，同时拥有windows版和安卓版。

语言属于公共领域的财产，广大人民群众贡献了整个语言体系的走向趋势。人民群众的语言是开源非加密的，商业拼音输入法在获取成千上万人的开源的语言后，分析其中的规律，推出更符合语言规律的拼音输入法，形式却是闭源的、加密的，而且是私人领域的财产。这在法律和道德上是不对等的。成熟的商业拼音输入法应当适当程度公开其获得的语言规律，也采用开源的形式。这叫取之于民，还之于民。算法可以理解为商业机密，词条数据认为完全属于私人财产是不合适的。算法的创造者是软件公司，而词条的贡献者并不是软件公司，而是来自成千上万的人民群众贡献的语料库，这属于公共领域的财产衍生品，同样属于公共领域的财产。词条数据的归属权大部分属于共用领域，少部分属于私人领域。

一些包含弹窗、捆绑安装、强制升级的商业化输入法，以前因为其强大的拼音词库你不得不用，从此可以对它们说再见了。

软件截图：





cffjln^

1.春风拂槛露华浓



充分发掘 | 充分发 | 测方法 | 常丰富 | 处罚法 |

!	@	#	\$	%	^	&	*	()	隱藏
Q	W	E	R	T	Y	U	I	O	P
全選	~	-	+	\	␣	{ }	:	;	
A	S	D	F	G	H	J	K	L	
Shift	`	剪下	複製	貼上	時間	"	'	退格	
方案	數字	<	鴻雁拼音			>	?	編碼	
中文	符號	,				.	/	Enter	



檻露华浓 很高兴再次见到你

hen gao x zai ci jian dao ni

1.很高兴再次见到你

很高兴 | 很搞笑 | 很高 | 很搞 | 行奥 | ▶



### 使用注意事项:

中文词语上屏，使用标点符号、回车键、或者空格按两次。单个空格可以用于整句输入的分词。

输入字母的半途，使用英文上屏，按下shift键。

初次安装输入法会生成词库索引，可能会占用较高的资源。尤其是鸿雁拼音手机输入法，需要等待1-4分钟，LevelDB数据库需要处理337万的数据。这个时候软件会出现无响应的状态，请耐心等待。

可以使用拼音的简拼输入词语，如键入“tswcbyy”，候选词列表一个是“天生我材必有用”。没有模糊音选项，一个汉字要么输入声母或者首字母，要么输入全拼。



请使用规范汉语拼音。比如“嗯”字，新华词典这个字的拼音有“ń ńg ń ńg ń ńg”，目前市面上的输入法可以使用“en”打出“嗯”字。这里不破坏拼音标准，拼音输出单字“嗯”，请输入“ng”或者“nnnn”。至于为什么不用“n”而用“nnnn”，下面作出解释。

在拼音输入词语的时候可以用每个字的拼音第一个字母组合起来作为简拼输入。在汉字中有一些汉字的拼音只有一个字符，比如“诶 è”、“阿 à ā ē”、“诶 ń ńg”、“姆 m ń”，单个的拼音转化为英文字母，“a o e”这些字母作为简拼输入不会出现词语竞争，“n m”这两个字母单独输入，每一次输入时都会出现单独的“诶 姆”这些单音字列表，干扰使用体验，故把完整拼音是“n”或者“m”的汉字输入拼音分别改为“nnnn”、“mmmm”。这里采用类似转移字符的方式绕过这个小狼毫输入法缺陷。

本输入法未对词语的马尔可夫链概率链进行统计，就是前后两个词语的相关概率并未统计。鉴于需求的轻重缓急，本方案比较简单粗暴，只追求单个汉字和单个词语极致的拼音库质量和数量。

因为词库数据量较大，安卓平台的版本按照同文输入法的默认步骤安装会无法安装成功，会出现无响应、进入不了输入法界面的状态。需要按照“鸿雁拼音手机输入法安装步骤.pdf”文档，按照特定的步骤才能成功安装。

鸿雁拼音输入法windows版只能在Windows 7及以上平台使用，不支持windows xp平台。软件在windows 7 32bit、windows 7 64bit、windows 10 64bit上测试通过。

鸿雁拼音输入法安卓版在一加手机上Android 10测试通过。

## 输入法采用的技术框架：

【小狼毫】输入法 <https://github.com/rime/weasel>

同文 Android 输入法 <https://github.com/osfans/trime>

## 使用协议：

约定GNU通用公共许可证为本输入法的使用协议，协议链接(<https://www.gnu.org/licenses/gpl-3.0.html>)。

不限于任何商业、个人使用。如果引用本输入法的数据，务必标注数据来源并且对修改的部分开源。

鸿雁拼音输入法采用【小狼毫】输入法源代码修改编译，其中具体业务代码并未改动，做了如下修改：软件界面繁体转为简体，更改软件名称为鸿雁拼音，去除rime官方网站链接，去除其他输入方案，集成鸿雁拼音方案，更改软件图标为源代码中另一套更好看的图标方案。

鸿雁拼音手机输入法采用同文 Android 输入法框架，具体业务代码并未改动，做了如下修改：修正默认数字键盘的输入错误，中文输入键盘上字母调整为大写，调整输入框提示符为空，去除软件中的QQ群、讨论论坛、捐助相关的文字和链接。

## 数据来源：

## 高权参考词库

现代汉语词典第7版

百度百科与维基百科的词条标题的交集(约50万条)

唐诗三百首、宋词三百首、老子道德经、论语、诗经的整句

李白诗句全集

世界各个国家国名全称、简称

中华人民共和国行政区划省级、地级、县级

以上词库中的词语，除了生僻字，大部分得以保留。

## 简繁转换

Open Chinese Convert <https://github.com/BYVoid/OpenCC>

## 词频、字频、新词数据来源语料库

百度百科约560万个词条(14.5GB,约59亿字)

维基百科约400万个词条(10.1GB,约40亿字)

微博语料(7.4GB,约30亿字)

微信公众号语料(2.9GB,约12亿字)

新闻语料(12.6GB,约51亿字)

1946年-2003年人民日报全部数据纯文本(3.1GB,约11.6亿字)

联合国平行语料库中文部分(1.4GB,约5.5亿字)

殆知阁古代文献txt大全集(4.8GB,约17亿字)

语料库分析的时候分割成600MB大小的区块，共计114个区块。如果一个词语在三个及以上不同的区块出现，这个词语就成功入选。在两个字的词语、三个字的词语、四个字的词语中，选择排名靠前的词语180万条，再合并高权参考词库后，共计230万词条。

词语分析并未采用结巴分词这样的分词软件，直接采用简单粗暴的机械分词，假如两个汉字紧挨着，把这两个汉字视为两个字的词语。三个字、四个字的词语判断以此类推。“你吃饭了吗”，这段话被拆分成“你吃”、“吃饭”、“饭了”、“了吗”、“你吃饭”、“吃饭了”、“饭了吗”、“你吃饭了”、“吃饭了吗”，并进行统计。这种机械分词方式保证统计所有出现的词语排列组合。

结巴分词的c++版本分析速度最快，单线程约为11MB/s。其他的分词软件的速度一般低于11MB/s，大部分在1MB/s左右。简单粗暴的机械分词速度远远高于这个速度。使用golang实现的机械分词软件，对645MB的文本统计两个字的词语出现次数，并对出现的130万个中文词语频率由高到低排序，耗时仅为37.9s。分词、统计、排序三个任务加在一起，速度仍然高达17MB/s。机械分词对于大型的语料分析非常有利。本软件使用语料库约为58GB大小，分词统计排序共耗时4.3小时。如果按照一般中文分词软件的速度1MB/s，仅仅分词就需要16.5小时，时间成本大大增加。现有的分词软件分词功能并不完备，最为流行的结巴分词，也会出现错误的分词。语言千变万化，就算当今人工智能发展到比较发达的程度，还是无法处理语言精确分析的所有问题，因为语言存在不少隐含的变量，与现实世界的事实存在对应，这些是语料库的本身的信息无法提供的，需要人工核实。机械分词虽然存在冗余的组合词汇和不当位置的切割获得的词汇，这些缺点并不影响输入法输入过程中的体验，甚至会更好的辅助输入。机械分词生成的概率统计模型更接近真实的词语分布。只要语料样本足够大、足够全，那么日常工作生活中，朋友聊天、商业交流，普通工作中的文字录入、学术人员和传媒人员写作等等所需的词语可以全面覆盖。如果一个人活120岁，每天认识一个新词，一生认识的新词有4.38万个。230万的词语已经足够覆盖日常生活所需的词语。

对单个汉字、单个词语的频率使用220亿字的语料库得出精确的词频数据，用于鸿雁拼音输入法的输

入过程中的候选列表排序。高权参考词库的词语保证了标准词语的全面性，基于典型的大规模语料库分析得到的新词保证了常见中文输入过程中的几乎所有可能组合。就词库的质量和数量而言，本方案拥有超高的准确率。在两个字到四个字的词语中，存在没有完全分词的词语。这是可以接受的。比如“了吗”这个词语是由两个独立的标准词语组合而成，虽然在词典书上不算一个词条，在我们日常聊天语句这个词语却比较常见。五笔输入法是以拆字作为输入单元。鸿雁拼音输入法包含几乎我们日常生活中的所有词语组合，那么鸿雁拼音输入法进入拆词输入的时代。如果你打字的时候没有出现这个词，请你不要奇怪，要么就是你用的词语不属于词典上的标准词语，要么这个词语是你自造的词语别人几乎不用，要么这个词语是一个罕见的词语。输入法默认禁用用户词典，如果用户有开启用户词典的需求，可以查看Rime输入法文档修改配置。词库已经足够全面，一般情况下不需要补充新词。用户输入的词库会打乱原有的候选词语排序，干扰盲打的进程。每次输入同样的按键候选词语都是一致的，可以轻松实现盲打。只要记住候选词语的顺序，闭着眼睛都能打字。智能ABC曾经是中国大陆使用人数最多的输入法软件，原因就是相对宽松的拼音输入方式，并有词语拼音的首字母简化输入，拼音输入更符合一般人的思考习惯。鸿雁拼音输入法，立志成为新时代的智能ABC。

## 拼音数据来源

Unicode 14的字符，使用最新版perl正则引擎“`\p{han}`”作为识别汉字字符的标准，去除没有拼音的部分，剩下的字符入选到输入法可输入汉字列表中。无论是微博语料库，还是百度百科语料库、人民日报语料库，常用的汉字大约5000个，而汉典收录了93898个汉字，异体字字典收录106330个字，绝大部分汉字躺在书中睡觉，一般我们很少接触到它们。

找到一个大而全并且准确、可用的拼音库，存在不少的难度。公开的拼音数据库大部分存在不少错误，权威的拼音数据库，比如现代汉语词典、汉语大字典没有可靠的官方文本数据。办法总是有的，可以在多个拼音库的基础上，按照权威性、准确性采用分级投票的方式获得可靠性高、准确率高、涵盖汉字数量多的拼音库。

以新华字典、通用规范汉字字典、异体字字典为准，作为第一阶梯数据。使用汉典网、百度汉语、字统网的数据作为补充，作为第二阶梯数据。unicode 13标准中的汉字拼音、字海（叶典）网的拼音数据存在不少错误，辞源、古汉语常用字字典第5版、汉语大词典、汉语大字典、现代汉语词典第7版的拼音数据因为数据来源是通过OCR获得的，也存在不少错误。这些数据仅用于第三梯队，不直接采纳数据，仅仅对第一、第二阶梯的拼音数据投票。

按照前述的拼音数据合并方案输出的汉字-拼音数据库，涵盖汉字共计41442个，拼音的权威性、准确性、多音字的数据完整性得到较大改善。

## 下载链接：

<https://hong-yan.lanzouw.com/b00vkivc>

密码:1234