一、#2标注原则：按照句子结构划分#2。

1、第一步：判断字数

整句或分句超过七个字的，进行第二步划分。不超过七个字的，则不用标注#2。

2、第二步：划分连词

如果连词加上后面的内容超过七个字，连词后面要标注#2。一个字的连词，初始标注时在连词后标#2；如果最后需要合并，那就合并。

连词举例：

并列关系连词：和、跟、与、既、同、及、而、况、况且、何况、乃至等。

承接关系连词：则、乃、就、而、便、于是、然后、至于、说到、此外、像、如、一般、比方、接着等。

转折关系连词：却、虽然、但是、然而、而、偏偏、只是、不过、至于、致、不料、岂知等。

顺承关系连词：那么、原来、因为、由于、以便、因此、所以、是故、以致等。

选择关系连词：或、或者、还是、亦、非…即、不是…就是等。

假设关系连词：若、如果、若是、假如、只要、除非、假使、倘若、即使、假若、要是、譬如等。

比较关系连词：像、好比、如同、似乎、等于；不如、不及；与其…不如、若…则、虽然…可是等。

让步关系连词：虽然、固然、尽管、纵然、即使等。

递进关系连词：不但、不仅、而且、何况、并、且等。

条件关系连词：不管、只要、除非等。

目的关系连词：以、以便、以免、为了等。

如：虽然我没有那么好。——虽然=我没有那么好。

又慢慢徘徊了几步——又=慢慢徘徊了几步。

从=二零一六年的爆发=到=二零一七年降温，再到=二零一八年的冰点，北京土地市场=变化很大。（从……到，算不算连词）

3、第三步：划分结构

如果第二步结束，句子或分句，仍然超过七个字，那么优先把主语划分出来。主语划分完毕后，仍然超过7个字，接下来划分宾语，宾语划分后，仍

然超过7个字，按照句子结构，从大到小进行细分（当句子缺少主语，就直接划分宾语，宾语也没有，就直接细分其他结构）

例句1：就等=丈夫李天宇回来了

a、句子没有主语，所以跳过，直接划分谓语“丈夫李天宇回来了”

b、丈夫李天宇=回来了

主语

丈夫李天宇回来了，超过7个字，在该结构内进行划分，先划分出主语“丈夫李天宇”

划分结果：就等=丈夫李天宇=回来了

例句2：两颗硕大的泪珠=从眼角泌了出来

主语

主语为“两颗硕大的泪珠”，所以在主语后面划分#2。划分完后，“两颗硕大的泪珠”与“从眼角泌了出来”，都为7个字，所以不再进行划分。

划分结果：两颗硕大的泪珠=从眼角泌了出来

例句3：王硕这有理有据的论述从哲学观点阐发了爱的真谛。

a、王硕这有理有据的论述=从哲学观点阐发了爱的真谛。

主语

主语“王硕这有理有据的论述”划分完后，可以看到“王硕这有理有据的论述”与“从哲学观点阐发了爱的真谛”均超过7个字，所以要进一步划分

出宾语

b、王硕这有理有据的论述=从哲学观点阐发了=爱的真谛。

宾语

划分完宾语“爱的真谛”，我们发现“王硕这有理有据的论述”、“从哲学观点阐发了”仍然超过7个字，接下来划分按照句子结构再细分

c、王硕这有理有据的论述=从哲学观点=阐发了=爱的真谛。

“从哲学观点阐发了”划分完后，剩下的主语“王硕这有理有据的论述”超过7个字，再进行细分

d、王硕=这有理有据的=论述=从哲学观点=阐发了=爱的真谛。

1、主语 2、宾语

“王硕这有理有据的论述”，根据步骤，先划分主语“王硕”，超过7个字。再划分宾语“论述”。直到句子中没有超过7个字存在

划分结果：王硕=这有理有据的=论述=从哲学观点=阐发了=爱的真谛。

4、第四步：合并

当句子划分完成后，句中及句末出现单节奏时（1-2个字的词，以及3个字且无法再划分的词：1、为实词+虚词，2、为专有名词，如：感觉上，漂亮

的，矜持点，预测了，社科院，李小红 等等），视与前后部分语义的紧密程度和节奏平衡原则进行合并，当紧密程度相当时，考虑节奏平衡原则，

优先向节奏较少的合并（可以超过7个字）。

注意：

当单节奏出现在句首时，不作合并；例句：决心=不计一切代价，要得到力量。

当句首连续出现两个单节奏时，考虑节奏平衡原则，两个单节奏合并。例句：如果这时候=苏糖还没明白=自己穿越到了=星际时代的话

单节奏举例

想要吃：双节奏，想要+吃（两个实词）

有气势：双节奏，有+气势（两个实词）

你父亲：双节奏，你+父亲（两个实词）

去影响：双节奏，去+影响（两个实词）

云南的：单节奏，云南+的（实词+虚词）

世界上：单节奏，世界+上（实词+虚词）

每一天：单节奏，每+一天（虚词+实词）

第八届：双节奏，第+数词+量词

第八：单节奏，第+数词

例句1：就等=丈夫李天宇=回来了

单节奏 单节奏

合并后：就等=丈夫李天宇回来了

分析：句中单节奏为“就等”与“回来了”，“回来了”在句末，需要向前合并，“就等”在句首，我们可以不做处理。

最终结果：就等=丈夫李天宇回来了

例句2：两颗硕大的泪珠=从眼角泌了出来

分析：可以明显的看到句中没有单节奏，所以不需要合并

最终结果：两颗硕大的泪珠=从眼角泌了出来

例句3：王硕=这有理有据的=论述，从哲学观点=阐发了=爱的真谛。

单节奏 单节奏 单节奏

合并后：王硕=这有理有据的论述，从哲学观点阐发了=爱的真谛。

分析：因为“王硕”在句首，所以可以不作合并，“论述”和“阐发了”，是在句中的单节奏，故合并。

最终结果：王硕=这有理有据的论述，从哲学观点阐发了=爱的真谛。

例句分析

她喜欢彪德西的钢琴曲。

分主语：她=喜欢彪德西的钢琴曲。

分宾语：她=喜欢=彪德西的钢琴曲。

合并：她喜欢=彪德西的钢琴曲。 分析：“喜欢”为句中单节奏，在合并时，与前后的关系都十分紧密，这时，我们优先向节奏较少的“

她”合并

南松与贵州恒丰俱乐部之间已经就其买断事宜有了基本共识。

分主语：南松与贵州恒丰俱乐部之间=已经就其买断事宜有了基本共识。

分宾语：南松与贵州恒丰俱乐部之间=已经就其买断事宜有了=基本共识。

细分1：南松=与=贵州恒丰俱乐部=之间=已经就其买断事宜有了=基本共识。

细分2：南松=与=贵州恒丰俱乐部=之间=已经=就其买断事宜=有了=基本共识。

合并：南松=与贵州恒丰俱乐部之间=已经就其买断事宜=有了基本共识。

据美常驻联合国代表黑莉的话称。

没有主语，分宾语：据=美常驻联合国代表黑莉的话=称。

细分1：据=美常驻联合国代表黑莉的=话=称。

细分2：据=美=常驻联合国代表=黑莉的=话=称。

合并：据=美常驻联合国代表=黑莉的话称。

通过一个将死之人的嘴说出来，似乎更显得沧桑厚重，更有说服力。

分连词：通过一个将死之人的嘴说出来，似乎=更显得沧桑厚重，更有说服力。

分主语：通过=一个将死之人的嘴=说出来，似乎=更显得沧桑厚重，更有说服力。

没有宾语，细分：通过=一个将死之人的=嘴=说出来，似乎=更显得沧桑厚重，更有说服力。

合并：通过=一个将死之人的嘴=说出来，似乎更显得沧桑厚重，更有说服力。 分析：“嘴”、“通过”、“似乎”为三个单节奏，因为“

嘴”、“似乎”不在句首，所以需要合并。

计白是被民政局的工作人员叫醒的。

分主语：计白=是被民政局的工作人员叫醒的。

分宾语：计白=是被=民政局的工作人员=叫醒的。

细分：计白=是被=民政局的=工作人员=叫醒的。

合并：计白=是被民政局的=工作人员叫醒的。 分析：“是被”在语义上与“民政局的工作人员”更近，所以往后合并。之前会议讨论结果

为：计白=是被民政局的=工作人员=叫醒的。按照现在规则，“叫醒的”需要合并。20190617：没问题，合并。

班中只有稀稀拉拉的几个孩子没人接。

分主语：班中只有=稀稀拉拉的几个孩子=没人接。

细分：班中只有=稀稀拉拉的=几个孩子=没人接。 分析：由于划分结束后，没有1-2个字的词以及三个字单节奏，所以不需要进行合并。之

前会议讨论结果为：班中=只有稀稀拉拉的=几个孩子=没人接。

20190617：“班中”如果是状语，可以这样划分：班中只有=稀稀拉拉的=几个孩子=没人接。

二、#3标注规则(#3用$表示）

第一步：

按照字数，当一个句子小于20个字，则不需要标注#3

第二步：

当句子大于或等于20个字，则需要根据以下情况，把句子中的某些#2，标注成#3（一般每20个字，可以标注一个#3，#3倾向少标，模糊地

带不标注，#3两边的字数尽量匀称。）

1、在句子中插入逗号、冒号，顿号等标点符号而不影响句意的地方，可以标注#3

如：研究院院长杨现领=也认为=租赁市场=是未来的=发展方向

改成逗号不影响：研究院院长杨现领=也认为，租赁市场=是未来的=发展方向

标注结果：研究院院长杨现领=也认为$租赁市场=是未来的=发展方向

一群距离咱们=上万公里的外国人=居然从=近两千年历史的=重重迷雾里=准确把握住了=刘备这个人物的精髓，

改成逗号不影响:一群距离咱们=上万公里的外国人，居然从=近两千年历史的=重重迷雾里，准确把握住了=刘备这个人物的精髓，

标注结果：一群距离咱们=上万公里的外国人$居然从=近两千年历史的=重重迷雾里$准确把握住了=刘备这个人物的精髓，

一年四季=随时随地=都会有人=动脑筋=要弄死你的狗=偷掉你的狗！

插入顿号不影响：一年四季=随时随地=都会有人=动脑筋=要弄死你的狗$偷掉你的狗！

插入逗号不影响：一年四季=随时随地=都会有人=动脑筋$要弄死你的狗$偷掉你的狗！

标注结果：一年四季=随时随地=都会有人=动脑筋$要弄死你的狗$偷掉你的狗！

2、当句子中存在两个可升级为#3的#2时，先看成分，如果成分划分导致一部分小于七个字或者严重不均衡，则按照字数均衡标注

如：美国媒体警告=美对华贸易政策=将严重影响地方经济。》美国媒体警告美对华贸易政策$将严重影响地方经济。

朴成华曾说=韩国球队逢中国队必胜的日子=一去不复返了。》朴成华曾说韩国球队逢中国队必胜的日子$一去不复返了。

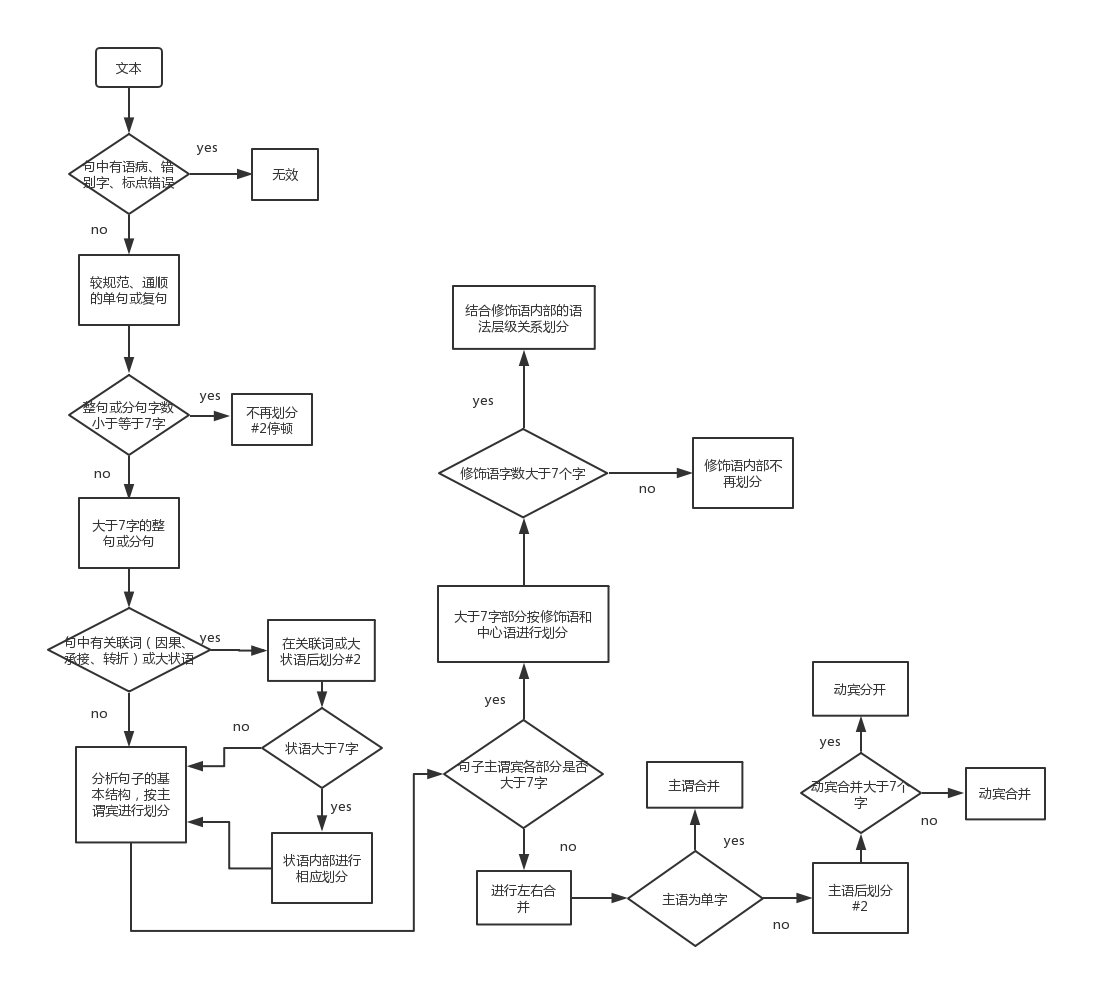
注：句子中出现用小括号表示的解释说明文字，数据筛选时去掉此类数据，后期针对此情况讨论具体的标注方法

身为大学同学的罗昊生，康雪妍（男女主角）等六人

《白鹿原》里有两个女人最为悲惨

字数均衡和语法之间需要衡量

合成文本韵律度#2规则 （旧）



**标注总则：**

（1）按照大语义结构拆分；

（2）语义结构内部如果大于7个字，继续拆分成小语义结构；

（3）大语义结构内部，相邻两个小语义结构均为单个节奏（韵律词），合并该两个小语义结构。？？？

举例：我们从来不生产任何具有转基因成分的西红柿品种。

（1）按大语义结构拆分：我们 / 从来不生产 / 任何具有转基因成分的西红柿品种。

（2）拆分大于7个字的大语义结构：我们 / 从来不生产 / 任何 / 具有 / 转基因成分的 / 西红柿品种。

（3）合并小语义结构：我们 / 从来不生产 / 任何具有 / 转基因成分的 / 西红柿品种。？？？

**一、标注目的及对#2的定义**

1、为了使合成效果更自然，合成音的停顿更符合自然朗读规律、更具有多样性，需要对文本中可能出现的、合理的换气停顿处进行标注。

2、#2指的是，文本中可能出现的、合理的换气停顿处。

**二、标注规范**

1、在文本中#2出现的地方标记“/”。

a.大致为2~3个节奏（#1）标注一个#2，一般为车3-7个字。

b.仅在子句子中进行#2标注，子句子边界和句群边界不进行标注；

c.顿号视为语调短语边界，默认该边界有长停顿；

d.其他句子内部的符号（书名号、引号、括号、破折号等）视具体情况进行标注。

例：

* 是否真的要做恶人，才会更直接的维护到自己的利益？

是否真的/要做恶人，才会/更直接的/维护到/自己的利益？

* 我是鼻子扁平、鼻孔朝天，因此下雨时喷嚏不断。

我是/鼻子扁平、鼻孔朝天，因此/下雨时/喷嚏不断。

* 也许正是站在这样的高度上来剖析问题的严重性，影片的典型感才显得那样的灼灼夺目。

也许/正是站在/这样的高度上/来剖析/问题的严重性，影片的典型感/才显得/那样的/灼灼夺目。

* 一七七六年七月四日，在费城正式通过《独立宣言》，宣告美国诞生。

一七七六年/七月四日，在费城/正式通过/《独立宣言》，宣告美国诞生。

2、在同一个句子中，仅标注一种合理的停顿方式。

例：

* 你大女儿叫苟丽萍，二儿子叫苟中建，三儿子叫苟中成。

你大女儿叫/苟丽萍，二儿子叫/苟中建，三儿子叫/苟中成。

* 你大女儿/叫苟丽萍，二儿子/叫苟中建，三儿子/叫苟中成。 （以上两种标注二选一即可）

你大女儿/叫/苟丽萍，二儿子/叫/苟中建，三儿子/叫/苟中成。 （不可用的错误标注）

3、对有问题的文本进行丢弃或修改。

1. 严重句意不通和内容不完整的文本，不再进行#2的标注，文本后标记“ #”；
2. 文本中出现错字和不能忽略的标点错误时，对文本进行尽可能少的改动，然后进行标注；
3. 文本中出现网络语和繁体字时，进行相应改写，不改变原文本的句意、句式、标点等，然后进行标注；
4. 文本结尾缺少标点时，根据句意相应补上标点，然后进行标注。
5. 忽略“的地得”的错误使用。

例：

* 误差理论与测量平差基础是由武汉大学出版社的。 #
* 式的大学学分将成为这种教育模式的发展方向。 #
* 如果你想做一个骇客，去读 #
* 兼于他是国王后场里相对不挑队友的那个，也许他真能活到最后。

鉴于/他是国王后场里/相对不挑队友的那个，也许/他真能活到最后。

* 有多少人就此累了，倦了如今的我们甚至没走几步，也就散了。

有多少人/就此累了，倦了，如今的我们/甚至没走几步，也就散了。

* 我木有照片，因为我是一个好人，不愿意吓唬别人。

我没有照片，因为我/是一个好人，不愿意/吓唬别人。

* 然後將讀音數據剪切到原文本之後，再點排序。

然后/将读音数据/剪切到/原文本之后，再点排序。

**三、一般规律**

1、如果主谓短语中的主语和动宾短语中的宾语的音节较多，在主语之后、宾语之前标注#2。

* 负担的问题/仍如水缸里的葫芦，摁一摁/就下去，手一松/又上来。（“负担的问题”作主语）
* 台球皇帝亨德利/退役二十七载，他的光辉生涯/载入体坛周报。（“台球皇帝亨德利”和“他的光辉生涯”分别作子句的主语）
* 玩命冲抢着/每一个/前场进攻篮板，争取二次进攻。（“每一个前场进攻篮板”作宾语）
* 落容桂跟我说/排号排到/三百多号哦！（“排号排到三百多号”作宾语）

2、介词短语、方位短语坐状语时，在这两种短语的后面可以标注#2。

* 荧屏上/满是宫闱床第、太监妃子、天子奴才，一片/打打杀杀之声。（方位短语“荧屏上”作状语）
* 山东省工商局/在法庭上/提供了/案发时/当事人摁手印的笔录。（介词短语“在法庭上”作状语）

3、对一些较长的名词性短语的处理。

1. 地址通常由这些部分组成：国家、省份、市、区（县）、社区名（乡、镇）、街道名、路名、门牌号、

具体地名。标注时优先保证将#2标注在层级单位后。

* 中国江苏省/南京市/玄武区/石象路七号/中山陵风景名胜区
* 北京东城区/东华门街道/多福巷胡同/甲二十二号
* 法国/曼恩卢瓦尔省/昂热

1. 公司名通常由这些部分组成：行政区划、字号、行业特点、组织形式或经济性质。
2. 行业特点通常包括产品名称、行业名称和产业名称，例如“贸易、信息科技、广告、企业管理”等。
3. 按组织形式一般分为股份有限公司（简称股份公司）、有限责任公司（简称有限公司）和个人独资公司。

按经济性质通常分为：私营企业，集体企业和外资企业（中外合资企业、外商独资企业和中外合作企业）。

**附1：部分表示行业性的词语**

商业：经贸、商贸、工贸、贸易

建设：建设工程、装饰装潢、建材、房地产开发、道路桥梁、园林绿化

服务：家政、咨询、中介、停车管理、搬家、餐饮

休闲娱乐：旅游、山庄、渡假村、疗养、夜总会、歌舞厅、酒吧

教育、文化、宣传：广告、传媒、科技、院校

医药、健身：医院、药品、医疗器械、健身器材、高尔夫

附2：部分特殊行业或小规模生产加工经营业的名称格式

厂、矿、场、中心、院、部、会所、超市、俱乐部、宾馆、饭店、美容院、连锁业、阁、园、楼、斋、轩……

这类标志性词语出现时，可以视具体情况跟前。

例：

* 厦门明都/房地产开发/有限公司
* 北京铭万智达/科技有限公司
* 广东华南顺通/大宗商品/交易中心
* 建筑科学研究院/新技术开发中心
* 北京国安/足球俱乐部

1. 定中短语前加上定语形成多层定语时，视具体情况进行#2标注。

* 我们学校的/八十年代的/两位/有三十年教龄的/优秀语文老师/也当选为代表。
* 他的一件/刚买的/新羊皮夹克/也拿来了。
* 朋友送的/很小的/两个花瓶。

**4、对含数字文本的处理**

1. **具体数值**
2. 对具体数值进行#2标注时，#2优先标记在数值权重后，通常为3~4个权重标记一个#2；
3. 如果数值的最小权重后存在孤立的数字，孤立的数字需向前合并，权重前的数值视具体情况重新标注#2；
4. 如果数字前出现序数词标志“第”或者“百分之、千分之”等，数字后出现三位名称或量词，这类字词需要与数字合并；

* 一万一千/零三十二点三
* 一千二百三十四亿/五千零六十七万/八千零九点一
* 一千二百三十点/四五六七
* 一万一千二百/三十点四五六
* 葡萄酒质量/百分之七十/来自于葡萄。
* 这批货物的总价值/约为/五千两百万美元。

1. **非具体数值**
2. 通讯号码、订单号等均属于非具体数值，超过6位的数字，通常3~5个数字划分一个#2。
3. 数字后可能带有“号、组、行、列”等表示单位的词，这类词通常需要向前合并。

c.根据文本中实际数字位数，参照下面的规则进行#2标注：

1~5位：内部不再划分#2

6位：3+3

7位：3+4

8位：4+4

9位：4+5

10位：5+5

11位：3+4+4

12位：4+4+4

例：

* 请一三二五四号/到第十二号柜台/办理业务。
* 我现用/幺八二/二五零六/八七幺九/这个号码。

**5、明显带有导航性质的句子：从/地点A/到/地点B**

* 从/健康服务中心/到/鑫鑫商店，怎么走最快？
* 搜索/从/蓝城一号/到/厦门盟泰兴/贸易公司/最好走的路线。
* 从/王州综合经营部/到/盛奎百货商场，怎么走最便捷？

**6、时间名词和表示具体时间的短语作状语时，后面可以标注#2。**

较长的表示具体时间的短语需要进一步划分，在“年、月、日、时（点）、分、秒”这类权重后标注#2。

* 今天/头一次感觉/宝贝真的/变成男子汉了，长大了。
* 二零一三年/十二月三十一日/十二时三十四分/五十六秒

**7、由两个及以上较长词语构成的联合短语，通常有“和、及、与、还是、并、并且、或、或者、而、而且**

* ”等标志性的连词，视具体情况可以在连词前标注#2。
* 让我们享受/一点点/拾级而上的生活/和为此/付出努力的过程。
* “一点点拾级而上的生活”和“为此付出努力的过程”都是音节较多的短语，所以连词“和”的前后可以都标注#2。
* 支持和怂恿/加朗集团/扩大战火，殃及百姓。
* 联合短语“支持和怂恿”音节较少，该短语内部不再标注#2。

**8、连谓短语中的谓词性词语之间视具体情况标记#2。**

* 怀着一丝希望/去找朋友/打听消息
* “怀着一丝希望”、“去”、“找朋友”和“打听消息”这四个谓词性词语共同构成一个连谓短语。整个短语的音节数较多，可以
* 适当划分。划分时为了使“去”字不被孤立，把“去”字向后合并。
* 这话/我听了很高兴。
* “听”和“很高兴”共同构成一个连谓短语，“听了很高兴”音节数较少，短语内部便不再划分。同时，划分时为了使“我”字不
* 被孤立，把“我”字向后合并。

**9、关联词语后通常可以标注#2。**

常见关联词语举例：

同时、同样、另外、而、而是、也、又、便、就、再、于是、然后、后来、接着、继而、终于、或者、或是、或、还是、还不如、倒

不如、即、就是说、尚且、何况、甚至于、况且、要不然、以免、免得、只是……

**四、详细规则**

以下规则不分优先级。

1、有词缀的词内不停延

现代汉语典型词缀包括：

前缀：阿、初、第、老、小、打

后缀：子、家、头、儿、们、边、巴、尔、乎、然、于

中缀：里、不

2、动词（v）+趋向补语结构内不停延

趋向补语包括：上、下、进、出、回、开、过、起、来、去、上来、下来、进来、出来、回来、开来、过来、起来、上去、下去、进

去、出去、回去、过去

3、中心语（v/a）+（得）+程度补语结构内不停延

中心语（v/a）+得+程度补语：很、透、要死、要命、不行、多、慌、厉害、不得了、不成、了不得、去了、够呛、够受的、邪乎、不

堪、吓人、够瞧的、凶、邪行、可怜、可以、够劲儿、惊人、出奇、过分、离谱、什么似的；

中心语（v/a）+程度补语+（了）：极、死、透、多、坏、慌、疯、傻、蒙、不堪、不过、够劲儿、万分、绝顶、无比、非常、异常、

出奇、过分、绝伦、透顶、之至、之极、至极；

4、动词+（得）+结果补语结构内不停延

结果补语通常由形容词充当，也有少数动词和动词性短语，数量太多，不一一列举。

5、动词+数量结构补语结构内不停延

数量太多，不一一列举。

6、动介结构内不停延

能跟在动词后面作补语的介词有：自、到、在、向、于、至、给

7、中心语+部分助词结构之间不停延

中心语+后附助词：的、地、得、着、了、过、看、般、一般、一样、似的、来着

8、句末（包括分句）语气词前不停延

常见语气词：的、了、吧、呢、啊、哇、呀、哟、哦、哈、嘛、呗、罢了、而已、也罢、也好、啦、嘞、喽、咯、着呢、吗、么

9、方位短语内不停延

名词或动词性词语+方位词构成方位短语，方位词包括：

单音节：上、中、下、前、后、左、右、东、南、西、北、中、里、内、外、间

之＋单纯方位词：之上、之中、之下、之前、之后、之左、之右、之东、之南、之西、之北、之中、之里、之内、之外、之间

以＋单纯方位词：以上、以下、以前、以后、以左、以右、以东、以南、以西、以北、以里、以内、以外

单纯方位词＋头：上头、下头、前头、后头、东头、南头、西头、北头、里头、外头

单纯方位词＋边（儿）：上边、下边、前边、后边、左边、右边、东边、南边、西边、北边、里边、外边、旁边

单纯方位词＋面（儿）：上面、下面、前面、后面、左面、右面、东面、南面、西面、北面、里面、外面

其他：上下、前后、左右、里外、内外、底下、当中、中间

注意：此规则有例外，如：人群#2里三层，外三层。

10、动词（v）+代名词结构之间不停延

代名词按意义可分为人称代名词、疑问代名词及指示代名词，常见的代名词有：

人称代名词：我、我们、咱们、你、您、你们、他、她、它、他们、她们、它们、自己、自个儿、别人、人家、大家、大伙、彼此

疑问代名词：谁、什么、哪、哪儿、哪里

指示代名词：这、那、这会儿、那会儿、这儿、这里、那儿、那里

注意：此规则有例外，疑问代名词“多会儿”的用法较为特殊，如：你是#2 多会儿来的？