# 浅析普通话韵律词的组构规则

熊子瑜

中国社会科学院语言研究所

**提要** 已有研究表明，纯粹基于声学语音学线索很难把韵律词边界（B1）和韵律词内部的音节边界（B0）准确地区分开来。本研究在对 ASCCD 朗读语料库进行手工分词和词性标注的基础上，对比分析了韵律词和文本词之间的区别，研究数据表明，绝大部分单字词不能独立构成韵律词，需要跟与其相邻的词语进行合并，这是导致韵律词和文本词不一致和很难对齐的重要原因。合并时，单字词是选择向左合并还是向右合并，在很大程度上受控于其自身的词类：单音节的量词、助词、方位词、名词和介词等倾向于向左合并；单音节的形容词、数词、副词、代词和动词倾向于向右合并；此外，单字词的合并方向还可能会受到相邻词语的词类和词长等因素的影响。通过规则把某些单字词跟其相邻词语进行合并之后，能够在一定程度上提高韵律词的识别正确率。

**关键词** 韵律词 文本词 合并 词性 词长

## 零 引言

随着语音识别和合成以及人工智能等技术的发展和要求，有关普通话的韵律结构和韵律单元的自动切分问题的研究越来越受到语音学界和言语工程学界的关注。近年来，学界围绕这一课题大致开展了以下三个方面的工作：

一、考察与普通话韵律结构和韵律单元（例如音节、韵律词、韵律短语、语调短语、语调组等）有关的声学语音学线索，包括停顿、音长、音高和音强等方面的韵律特征[林茂灿 2000；钱瑶 2001；熊子瑜 2003；王蓓等 2004]。已有的研究表明，在较大的韵律边界处通常会出现较长的语音停顿、边界前的音节韵母通常会有一定幅度的拉长、边界后的韵律单元会通常会出现一定幅度的音高重设等等。

二、考察普通话的音步和韵律词等韵律成分的组构规则，以及与此有关的音系、词汇和句法等方面的语言学线索[冯胜利 1996；王洪君 2000；曹建芬 2001，2003]。已有的研究表明，普通话的韵律词在词长上具有“二常规、三可容、一四受限”的特点，两个单音节的直接成分通常会紧密地结合在一起，单音节的粘着性成分（如介词、语助词、方位词等）通常会依附于相邻的主干成分，等等。

三、从识别或合成的角度采取一些统计模型来预测或检测普通话的韵律边界类型[应宏等 1999；牛正雨等 2001；胡伟湘等 2002；吴晓如等 2003]。已有的研究表明，采取一定的统计模型，利用某些韵律特征以及词类、词长等信息可以有效预测或检测出语流中的韵律边界类型。

对于语音合成来说，韵律边界的预测难点主要在于韵律短语边界；而对于语音识别来说，韵律边界的检测难点主要在于韵律词边界。如胡伟湘等（2002）采用时长、基频和音强三个方面的韵律特征来描述朗读话语的韵律间断模式，并采用决策树方法对韵律边界进行了自动检测和识别。其结果表明，尽管综合正确率比较满意，但对韵律词边界的正确识别率较低，只有 55%左右，很多韵律词边界被错误地识别为 0 级韵律边界（即韵律词内部的音节边界）。

韵律词边界的检测准确率较低，这表明仅依靠声学语音学线索很难把韵律词边界和韵律词内部的音节边界有效地区别开来。因此除了要利用声学语音学线索之外，还应进一步考虑其他信息（如词类、词长和相邻词项等）在判别韵律词边界时的作用。

## 一 语料及数据准备

本研究采用的语料取自ASCCD朗读语篇语料库[1]，

[1] 有关该语料库的介绍请参见 http://ling.cass.cn/yuyin/product/product_6.htm

该语料库由 18 篇小短文组成，合计约 8760 个音节，由 10 位（5 男 5 女）说标准普通话的成年发音人朗读。其前期的语音学标注主要包括音节、声韵母、韵律边界和重音指数等信息，完全由手工完成。

在语音学信息标注的基础上，本研究采用自动分词和词性标注工具对ASCCD朗读语篇语料库的录音文本进行切分和标注，并聘请了两位具有语言学专业背景的研究人员对自动分词结果及词性进行了手工校正[2]，然后采用自动对齐工具将文本分词结果及词性标注信息写入语音标注文件之中，使文本文件中的每个汉字与语音标注文件中的相应音节对齐。但由于某些发音人在朗读时出现了个别多字、少字和错字，以及标注时出现了个别拼音录入出错，难免会造成部分数据不能自动对齐，对于这部分不能自动对齐的数据，我们进行了手工修改和补充。为了便于述说和区别，本研究把文本分词的结果称为“文本词”，把语音分词的结果称为“语音词”或“韵律词”。文本词基本上等同于词典词，语音词指语流中紧密相连的语音成分，由标音员根据语感和听感进行切分，它通常由一个或多个文本词组构而成。

校正好标注数据之后，我们从标注文件中提取出本研究所需要的数据，主要包括以下两个方面：（1）文本词数据，主要包括当前词的词性、当前词的词长，前接词的词性、后接词的词性、前接词和后接词的词长[3]，当前词的左边界和右边界类型[4]，当前词的末音节韵母时长、当前词的右端停顿时长、当前词内部各音节的韵母时长组合；（2）韵律词数据，主要包括韵律词的字数、所含文本词的数目，所含文本词的词长组合（如 1/1/、1/2/、2/1/和 2/2/等），所含文本词的词性组合（如动词+名词、形容词+名词、副词+动词、数词+量词+名词等），韵律词末音节的韵母时长、韵律词的右端停顿时长，韵律词内部各音节的韵母时长组合。

## 二 数据分析

### 2.1 韵律词与文本词的对比

有研究表明，汉语韵律词在长度上存在“二常见、三可容、一四受限”的特点。单字词很难单独构成一个韵律词，它通常需要跟与其相邻的词语一起组构成韵律词。这一特点在 ASCCD 朗读语料库中表现的也很明显，韵律词长度的数据如下：

表一：ASCCD 语料库中韵律词长度的数据

| 韵律词长度[5] | 数目 | 比率 |
|---|---|---|
| 单音节 | 2864 | 8.9% |
| 两音节 | 14733 | 45.8% |
| 三音节 | 8009 | 24.9% |
| 四音节 | 4612 | 14.3% |
| 其他 | 1980 | 6.1% |
| 合计 | 32198 | 100.0% |

数据表明，单音节韵律词仅占韵律词总数的 8.9%。那么，在 ASCCD 朗读语料库中的单字词有多少呢？会不会是由于单字词本身就少，从而导致单音节韵律词的比重偏低呢？下表给出了 ASCCD 朗读语料库中文本词长度的统计数据：

表二：ASCCD 语料库中文本词长度的数据

| 文本词长度[6] | 数目 | 比率 |
|---|---|---|
| 单字词 | 29193 | 51.7% |
| 多字词 | 27260 | 48.3% |
| 合计 | 56453 | 100.0% |

数据表明，在 ASCCD 语料库中，单字词的数目超过了词语总数的一半。但由于汉语韵律词在长度上存在“二常见、三可容、一四受限”的特点，有

[2] 本文在分词时主要参考《现代汉语词典》，除了熟语和成语外，尽可能切分到词典词。当遇到某些词典里未收录的专有名词（如人名、地名、物名和时间词等），本研究将其处理为一个词，如“太阳宫、火腿肠、孙庆福（人名）”等。

[3] 本文把前接词和后接词的词长分为无词、单字词和多字词三种类型。胡伟湘等（2002）研究表明，基于声学语音学线索能够有效地检测出韵律短语的边界，所以本研究在分析词语的上下文语境时以韵律短语为基本单元。如果某文本词处于韵律短语的起首位置，则设定其前接词的词长类型为“无词”。如果某文本词处于韵律短语的末尾位置，则设定其后接词的词长类型为“无词”。

[4] 本文把当前词的左（右）边界类型分为有韵律边界或无韵律边界两类。如果某文本词的左（右）端韵律边界类型为 0，则设定其左（右）边界为“无韵律边界”，简写为“无”；否则设定其左（右）边界为“有韵律边界”，简写为“有”。

[5] 本文把韵律词长度分为五种类型：单音节、两音节、三音节、四音节和其他，凡是音节数目等于或大于 5 个的韵律词都归入其他类。

[6] 本研究将文本词长度分为两种类型：单字词和多字词，凡是含有两个或两个以上汉字的词都归入多字词一类。

将近 90%的单字词不能直接构成韵律词，需要与其他相邻词语合并成一个韵律词，这导致文本词和韵律词之间存在严重不一致和不能对齐的现象：一个韵律词可以由一个或多个文本词构成；一个文本词有时也可能被切分成多个韵律词。下面按含有文本词的个数对韵律词进行归类，并统计出 ASCCD 朗读语料库中各类韵律词的数目，结果如下：

表三：韵律词内部所含文本词的个数分类

| 韵律词内部所含文本词的个数 | 数目 | 比率 |
|---|---|---|
| 1 | 15522 | 48.2% |
| 2 | 11015 | 34.2% |
| 3 | 3258 | 10.1% |
| 4 | 423 | 1.3% |
| 其他[7] | 1980 | 6.1% |
| 合计 | 32198 | 100% |

数据表明，由单个文本词直接构成的韵律词仅占韵律词总数的 48%，另有将近一半的韵律词是由两个或两个以上的文本词合并而成的。本文把这种由两个或两个以上文本词合并而成的韵律词称为“复杂韵律词”，把由单个文本词直接构成的韵律词称为“简单韵律词”。由于语流中存在大量的复杂韵律词，所以基于文本词的边界很难直接有效地预测出韵律词的边界。面对这种情况，只有充分掌握复杂韵律词的构词规则，了解各类文本词在选择韵律边界类型时的倾向性和制约条件，才有可能去正确预测哪些类型的文本词通常会合并在一起形成复杂韵律词。下面按其所含文本词的词长组合类型对韵律词进行归类，并统计出 ASCCD 朗读语料库中各类韵律词的数目，结果如下：

表四：韵律词的词长组合类型数据

| 韵律词类型 | | | | 比率[8] | 比率和 |
|---|---|---|---|---|---|
| 简单韵律词 | | | 1/ | 8.9% | 48.3% |
| | | | 2/ | 34.1% | |
| | | | 3/ | 4.4% | |
| | | | 4/ | 0.9% | |
| 复杂 | 是否 | 否 | 2/2/ | 5.7% | 5.70% |
| | | 是 | 1/1/ | 11.7% | 39.9% |
| 韵律词 | 含有单字词 | | 2/1/ | 10.9% | |
| | | | 1/2/ | 4.9% | |
| | | | 3/1/ | 0.7% | |
| | | | 1/3/ | 0.3% | |
| | | | 1/1/1/ | 4.7% | |
| | | | 1/1/2/ | 2.3% | |
| | | | 2/1/1/ | 2.1% | |
| | | | 1/2/1/ | 1.0% | |
| | | | 1/1/1/1/ | 1.3% | |
| 其他 | | | | 6.1% | 6.1% |

数据表明，含有一个甚至多个单字词的复杂韵律词占韵律词总数的 40%，占复杂韵律词总数的 88%。鉴于复杂韵律词大都含有一个甚至多个单字词，本文认为，单字词是构成复杂韵律词的重要组成部分，是导致文本词和韵律词之间存在严重不一致和不能对齐的主要原因，掌握单字词与其相邻词语的合并规则，是有效预测复杂韵律词的基本前提。所以本文接下来将以单字词为主要考察对象，分析单字词的词性、以及单字词相邻词语的词性和词长等因素跟单字词的韵律边界类型之间的关系，以求掌握单字词的合并方向和条件。

### 2.2 文本词的韵律边界类型

按照文本词在韵律词中所处的位置，本研究将文本词的韵律边界分为以下四种基本类型：（1）1-X-1，即单个文本词独立组成韵律词，其左右两端都有韵律边界；（2）1-X-0，即文本词处于韵律词首，其左端具有韵律边界，右端没有韵律边界；（3）0-X-1，即文本词处于韵律词末，其左端没有韵律边界，右端具有韵律边界；（4）0-X-0，即文本词处于韵律词内部，其左右两端都没有韵律边界。为了便于分析，本研究有时会把文本词的韵律边界分为左边界和右边界，如果文本词处于韵律词首，则设定其左边界为 “有”，否则设定为“无”；如果文本词处于韵律词末，则设定其右边界为“有”，否则设定为“无”。

下面根据韵律边界类型的归类结果，统计了文本词的词长类型与其韵律边界类型之间的关系，结果如下：

[7] 本研究将长度等于或大于 5 个音节的韵律词统统归入其他类，其内部不再分析词语组构情况。

[8] 限于表格宽度，各类韵律词的“数目”一栏被略去，其值等于各类韵律词的“比率”*韵律词总个数（32198）。

表五：文本词的词长类型与韵律边界类型

| 韵律边界类型 | 总词数 | 单字词所占比率 | 多字词所占比率 |
|---|---|---|---|
| 1-X-1 | 15815 | 18.10% | 81.90% |
| 0-X-0 | 6858 | 81.50% | 18.50% |
| 1-X-0 | 16893 | 53.70% | 46.30% |
| 0-X-1 | 16887 | 69.10% | 30.90% |
| 合计 | 56453 | 51.70% | 48.30% |

数据表明，尽管单字词和多字词在 ASCCD 语料库中的所占份额相差不大，但它们在选择韵律边界类型时却有着显著不同的倾向性：在韵律边界类型为“1-X-1”型的文本词之中，82%的属于多字词，只有 18%的属于单字词；而在韵律边界类型为“0-X-0”型的文本词之中，82%的属于单字词，只有 18%的属于多字词；在韵律边界类型为“0-X-1”型的文本词之中，单字词占到了 70%左右，相对于多字词而言，单字词更倾向于向前附着于其相邻的词项。

### 2.3 影响文本词的韵律边界类型的因素分析

除了文本词的词长类型会影响其韵律边界类型之外，文本词的功能类别也会对其选择韵律边界类型产生影响。从功能和意义上可以把词语分成两大类：实词和虚词，实词通常能够单用或单说，虚词通常不能单用或单说。相对而言，能够单用或单说的词语往往是比较自由的，而那些不能单用和单说的词语往往是不自由的。

为了分析文本词的功能类型与其韵律边界类型之间的关系，本文对 ASCCD 朗读语料库中文本词的词类及其词长数据进行了统计，结果如下：

表六：文本词的词类和词长分布信息

| 词类 | 总词数 | 单字词所占比率 | 多字词所占比率 |
|---|---|---|---|
| V（动词） | 14610 | 49.00% | 51.00% |
| N（名词） | 14297 | 22.00% | 78.00% |
| U（助词） | 5548 | 99.20% | 0.80% |
| D（副词） | 4654 | 64.20% | 35.80% |
| A（形容词） | 3409 | 25.20% | 74.80% |
| R（代词） | 2967 | 66.10% | 33.90% |
| M（数词） | 2242 | 73.70% | 26.30% |
| Q（量词） | 2109 | 92.40% | 7.60% |
| P（介词） | 1991 | 93.60% | 6.40% |
| C（连词） | 1881 | 49.00% | 51.00% |
| F（方位词） | 1420 | 68.20% | 31.80% |
| T（时间词） | 1126 | 1.80% | 98.20% |
| Y（语气词） | 169 | 94.10% | 5.90% |
| O（象声词） | 30 | 100% | |
| 合计 | 56453 | 51.70% | 48.30% |

数据表明，尽管单字词和多字词在 ASCCD 语料库中的所占份额相差不大，但对于不同类型的文本词来说，单字词和多字词所占的比率差异很大，文本词的词类与词长之间密切相关：（1）在助词、量词、介词、语气词、象声词和数词中，单字词占绝对优势；（2）在方位词、代词和副词中，单字词占微弱优势；（3）在时间词、形容词和名词中，多字词占优势；（4）在动词和连词之中，单字词和多字词的所占份额几乎相当。

根据表六数据，象声词、语气词和时间词中的单字词数目较少，所以在接下来的统计分析中被剔除了，其他 11 类单字词的词类与其左边界类型之间的关系如下所示：

表七：单字词的词语类型与其左边界类型

| 词类 | 总词数[9] | 左边界 | |
|---|---|---|---|
| | | 无 | 有 |
| A | 681 | 62.4% | 37.6% |
| C | 232 | 28.9% | 71.1% |
| D | 1321 | 47.7% | 52.3% |
| F | 945 | 94.4% | 5.6% |
| M | 1351 | 64.9% | 35.1% |
| N | 2693 | 88.0% | 12.0% |
| P | 1002 | 76.9% | 23.1% |
| Q | 1948 | 98.9% | 1.1% |
| R | 756 | 59.4% | 40.6% |
| U | 5449 | 98.2% | 1.8% |
| V | 4811 | 68.8% | 31.2% |
| 合计 | 21189 | 80.6% | 19.4% |

数据表明，不同词类的单字词对左边界类型的选择具有不同的倾向性：（1）98.9%的单音节量词、98.2%的单音节助词、94.4%的单音节方位词、88%的单音节名词和 77%的单音节介词没有左边界，倾向于跟前接词进行合并；（2）71.1%的单音节连词有左边界，独立于前接词；（3）单音节副词和代词

[9] 统计时剔除了那些处于韵律短语起首的单字词。

对于左边界类型的选择没有显著的倾向性。

下面统计分析这 11 类单字词的词类与其右边界类型之间的关系，数据如下：

表八：单字词的词语类型与其右边界类型

| 词类 | 总词数[10] | 右边界 | |
|---|---|---|---|
| | | 无 | 有 |
| A | 668 | 93.0% | 7.0% |
| C | 874 | 60.5% | 39.5% |
| D | 2911 | 79.9% | 20.1% |
| F | 311 | 63.7% | 36.3% |
| M | 1564 | 93.3% | 6.7% |
| N | 1629 | 66.8% | 33.2% |
| P | 1665 | 49.5% | 50.5% |
| Q | 1331 | 53.9% | 46.1% |
| R | 1812 | 76.2% | 23.8% |
| U | 3846 | 31.8% | 68.2% |
| V | 5956 | 71.9% | 28.1% |
| 合计 | 22567 | 64.9% | 35.1% |

数据表明，不同词类的单字词对其右边界类型的选择具有显著不同的倾向性：（1）93.3%的单音节数词、93%的单音节形容词、79.9%的单音节副词、76.2%的单音节代词和 72%的单音节动词没有右边界，倾向于跟后接词进行合并；（2）68.2%的单音节助词有右边界，独立于后接词；（3）单音节介词和量词对于右边界的选择没有显著的倾向性。

对于那些倾向性不够显著的词或者需要进一步细致区分的词，还可以进一步考虑其前后词语的词性和词长等线索来帮助确定其合并方向。

下表给出了单音节动词在前接不同类型的词语时，对其左边界类型的选择结果，以考察前接词类对单音节动词合并方向的影响，数据如下所示：

表九：单音节动词与前接词之间的边界类型

| 前接词类型 | 总次数 | 左边界 | |
|---|---|---|---|
| | | 无 | 有 |
| D | 1608 | 83.7% | 16.3% |
| V | 1116 | 74.5% | 25.5% |
| N | 756 | 46.2% | 53.8% |
| R | 458 | 74.7% | 25.3% |
| U | 262 | 42.4% | 57.6% |
| C | 201 | 67.7% | 32.3% |
| F | 108 | 59.3% | 40.7% |
| M | 74 | 70.3% | 29.7% |
| P | 70 | 44.3% | 55.7% |
| Q | 68 | 23.5% | 76.5% |
| A | 47 | 42.6% | 57.4% |
| T | 43 | 32.6% | 67.4% |
| 合计 | 4811 | 68.8% | 31.2% |

数据表明，当单音节动词前接副词、代词、动词和数词时，该单音节动词明显倾向于跟其前接词合并。

下表给出了单音节介词在前接不同类型的词语时，对其左边界类型的选择结果，以考察前接词类对单音节介词合并方向的影响，结果如下表十所示。其数据表明，单音节介词更倾向于跟前接的动词、副词、代词和连词合并。当单音节介词前接名词时，它倾向于跟后接词合并。

表十：单音节介词与前接词之间的边界类型

| 前接词类型 | 总次数[11] | 左边界 | |
|---|---|---|---|
| | | 无 | 有 |
| v/ | 449 | 86.9% | 13.1% |
| d/ | 207 | 85.5% | 14.5% |
| n/ | 127 | 38.6% | 61.4% |
| r/ | 99 | 86.9% | 13.1% |
| c/ | 61 | 82.0% | 18.0% |
| 合计 | 943 | 79.7% | 20.3% |

此外，本研究还考察了后接词的词类对单字词合并方向的影响，以及前后词长等因素对单字词合并方向的影响。限于篇幅，数据不再一一列出，下面仅给出前后词长因素对单字词合并方向的影响数据：

表十一：前后词长类型与单字词的合并方向

| 前后词长组合 | 总次数 | 左边界 | | 右边界 | |
|---|---|---|---|---|---|
| | | 无 | 有 | 无 | 有 |
| 单-X-单 | 5483 | 73% | 27% | 82% | 18% |
| 单-X-多 | 3842 | 91% | 9% | 36% | 65% |
| 多-X-单 | 2807 | 56% | 44% | 81% | 19% |
| 多-X-多 | 2965 | 79% | 21% | 32% | 68% |

数据表明，当后接多字词时，单字词倾向于向左合并；当后接单字词时，单字词倾向于向右合并。

以上研究表明，单字词的词类信息、前接词和

[10] 统计时剔除了那些处于韵律短语末尾的单字词。

[11] 统计时除了剔除那些处于韵律短语起首的单音节介词，还剔除了总次数小于 20 的统计结果。

后接词的词类、词长信息都能够用来帮助判定其合并方向：或向左或向右合并。

## 三 结论

由于语流中存在大量难以自成韵律词的单字词，导致韵律词和文本词不一致和很难对齐，故而直接基于分词结果很难准确预测韵律词（特别是复杂韵律词）的边界。鉴于此，本研究在分词的基础上，根据单字词的词类、其前接词和后接词的词类、词长等信息，对单字词的合并方向进行了初步研究，以图把握复杂韵律词的构词规则。研究数据表明，单字词是选择向左合并还是向右合并，在很大程度上受控于其自身的词类：单音节的量词、助词、方位词、名词和介词等倾向于向左合并；单音节的形容词、数词、副词、代词和动词倾向于向右合并；此外，单字词的合并方向还可能会受到相邻词语的词类和词长等因素的影响。

本研究按照上述规则和统计结果把语料库中的单字词跟相邻词语进行合并，删除大量的词边界，根据合并后的边界信息，来判别韵律词边界和韵律词内部的音节边界，我们得到了将近 90%的综合正确率。

**致谢**：本研究得到国家社科项目“汉语语调模式的研究”（编号: 60475043）和社科院语言所 ASCCD 朗读语篇语料库的支持。

**主要参考文献**

曹剑芬 汉语韵律切分的语音学和语言学线索，载于《新世纪的现代语音学》，清华大学出版社，2001年。

曹剑芬 基于语法信息的汉语韵律结构预测，《中文信息学报》2003年03期；

冯胜利 论汉语的“韵律词”，《中国社会科学》1996年01期；

胡伟湘等 汉语韵律边界的声学实验研究，《中文信息学报》2002年01期；

林茂灿 普通话语句中间断和语句韵律短语，《当代语言学》2000年04期；

牛正雨等 基于边界点词性特征统计的韵律短语切分，中文信息学报，2001，15（5）：19-25；

钱瑶等 普通话韵律单元边界的声学分析，《新世纪的现代语音学》，清华大学出版社，2001年；

王蓓等 汉语韵律层级结构边界的声学分析，《声学学报》(中文版)2004年01期；

王洪君 汉语的韵律词与韵律短语，《中国语文》2000年第6期；

吴晓如等 基于韵律特征和语法信息的韵律边界检测模型，《中文信息学报》2003年05期；

熊子瑜 韵律单元边界特征的声学语音学研究,《语言文字应用》2003年02期；

应宏等 基于结构助词驱动韵律短语界定的研究，中文信息学报，1999，13（6）：41-46；

# On the Compositional Rules of prosodic words in Mandarin

Xiong Ziyu

Institute of linguistics, Chinese academy of social sciences

**Abstract:** Previous studies show that it is hard to distinguish precisely the prosodic word boundary (B1) from the syllable boundary within the prosodic word (B0) solely based on acoustic phonetic clues. On the basis of manual segmenting and POS tagging the ASCCD Database of read speech, the present study focuses on the differences between a prosodic word and a lexical word. Findings indicate that prosodic words fail to be aligned with lexical words to a great extent, because most of the lexical words have to merge with other neighboring words and phrases to form prosodic words. While they merge, the orientation is mainly determined by the parts of speech of the words themselves: left-side merging for the monosyllabic measure words, auxiliary words, localizer, nouns, and prepositions; right-side merging for the monosyllabic adjectives, numerals, adverbs, pronouns, and main verbs. Further, the orientation could also be influenced by other variables such as the parts of speech and the length of the neighboring words. Thereafter, a set of merging rules were set up to improve the recognition of prosodic words.

**Keywords**: prosodic word lexical word merging part of speech word length