谈谈语料库的语样选取问题
ON SAMPLE SELECTION OF THE CORPUS

曹 剑 芬
Cao, Jianfen

中国社会科学院语言研究所
Institute of Linguistics
Chinese Academy of Social Sciences

摘 要

自然语言处理的各个部门，包括分析－合成、自动识别、语言理解、机器翻译以及人工智能等等，都力图通过不同范围、不同规模的语料库来自动生成自然语言的各项规则，以供具体处理系统进行检索、训练或评估等方面应用。因此，怎样自觉地、有目的地运用语言学的原则，来建立既尽可能切近自然语言实际、又经济典型的语料库，就不是一个无足轻重的问题了。本文试图从语言学的角度，以汉语普通话语音库的选样原则为例，来谈谈语料库建库中必然涉及的语样选取问题。

ABSTRACT

It has become a common subject that people try to generate the rules of natural language automaticlly through different type and diffrent size of corpora, so that to meet the demands on the indexing, traning and evaluation raised in almost all of the area in natural language processing, including speech analysis－synthesis, automatic recognition, language understanding, machine－translation, artificial intelligence and so on. Consequently, how to consciously apply the linguistic principle in building of a corpus is not an insignificant issue in order to form an idea corpus which is not only as close as possible to the nature of real speech, but also typically and economically samled. This paper tries to discuss some aspects on samle selection from linguistic viewpoint by describing the main principle involved in the Speech Corpus of Standard Chinese.

0.前言

随着计算语言学和言语工程学的不断深入和迅速发展，无论是人工智能、自然语言理解、机器翻译，还是言语的人工合成和自动识别，都涉及对于人类自然言语的了解和仿生问题，包括知识的获取和表示、话语的生成和理解以及记和忆的机理及其相关的模型，等等。但是，由于这些问题无不涉及大脑黑箱的秘密，对此，人们多半还处于知其然而不知其所以然的状态。为了尽快解决自然语言处理过程中所遇到的种种棘手的问题，人们不得不从言语过程的另一端入手，即求助于既存的自然语言资料－－文字的和/或语音的，希望通过它们来自动

地生成规则，以供具体系统掌握和利用。至于一些研究部门，尤其是从事基础理论研究的部门，更是需要通过对于自然语料的分析和归纳，揭示自然言语运作的客观规律，从中提取言语链活动的各个环节上的典范和规则，以期从理论上提高对于这个人类所特有的智能现象的认识，为计算机的人工仿生提供理论基础。然而，自然语料是个汪洋大海，任何研究部门或应用部门都不可能使用这个大海里的全部材料，而只可能采用它的局部样本作为研究的典型或应用的模板。于是，各种各样的语料库便应运而生。由此可见，最初的语料库是在具体研究和应用自然语言的过程中自发形成的。

如今，随着各种各样言语处理或仿生系统如雨后春笋般地诞生，人们希望能够建立相对统一的训练和评估标准，因而对于语料库的要求也越来越高了。如果说最初的一些语料库还只是作为某些特定言语处理系统的副产品、在朦胧中自发诞生的话，那么，现在的人们是强调更加有意识、有目的地建库，从一定意义上来说，这样的语料库已不再是某个特定系统的附属品，而应该是一个独立自主、在相关领域里适用于不同课题的、标准化的系统。实际上，有关语料库的建立、管理及其研制业已发展为一个独立的学科分支。例如，1992 年在加拿大召开的第二届国际言语处理学术会议上，语料库问题已赫然作为一个专门分支学科进行学术交流。国际语音学会前主席、美国洛杉矶加州大学教授彼德·拉第福格特在大会报告中还专门论述了语料库和数据库的地位及作用[1]。由此可见，怎样建设好语料库已成为当今言语科学研究领域不可忽视的重要课题。本文试图从语言学的角度就语料库的建设以及语样选取的原则谈一点看法。

1. 语料库和自然语言规则

1.1. 通过语料库生成规则和根据语言学原则建库的关系

如上所述，语料库在当今言语科学研究中的地位是显而易见的。但是，要想建设好语料库，很重要的一条就是遵循自然语言的客观规律，努力按照 语言学的原则来建库。因为语言学是研究语言客观规律的科学，尽管任何一部分自然语料都能体现一定的语言规则，但是，只有根据语言学的原则有目的地选用语料，才能保证这个语料库能够有效地体现自然语言的客观规则。这种关系是由自然语言的本质特征决定的。

自然语言本身是由一定的语言社会约定俗成的、自然的规则体系，它的运作过程遵循着一定的语音、语法和语义的结构规则。这些规则当然不可能期望完全由一个包罗万象的语料库来体现，更不是现存的某一个语料库所能全面囊括的，而只能通过一个个具体的、能够比较全面地体现自然语言某一方面或某些方面规则系统的语料库来实现。我们在建立每一个这样体现自然语言的局部规则系统的语料库的时候，就可以根据我们已经掌握的语言学原则，避免盲目性，使它有效地生成那些我们尚未掌握、但却必定存在的语言规则。

1.2. 言语处理系统的质量与语料库质量的关系

任何言语处理系统，无论是识别、合成，还是机器翻译或自然语言理解系统，都是在不同程度上、从不同的角度对于自然语言规则体系的模仿。从根本上说，要想提高这些系统的质量，就应该首先加深对于自然言语过程的认识。然而，自然言语是个谜，迄今为止，且不说

人们尚不能完全掌握它运作的客观规律及其相应的规则体系，就是对于已经发现和掌握的规则，也还由于知识表示或规则描写方面的种种局限而未能充分付诸应用，因而导致这些系统对于语料库的不可避免的依赖性。因此，从这个角度来说，一个语料库的质量将直接影响到一个语言处理系统的质量，因为它实际上负有训练和检验这个系统的全部责任。举例说，你想做一个为数字串的识别用的语料库，究竟应该存储什么样的和多大规模的语料？存一个个单念的数字显然不行，因为数字与其它单字一样，单念时是一个音节，一旦串起来连读构成了音节流，由于协同发音的作用[2, 3]，它们就表现为不同于单念状态的变体，而且，这些变体是随环境而定，具体存在于一个个数目字串之中。所以，为了识别需要，你就得存储这样的数目字串。那么，究竟存储多少合适？存储少了不够用，万一识别时碰到某种变体恰恰是这个语料库漏收的，机器就可能误识或拒识；存多了又不经济，还影响运行速度，这同样关系到你的系统的质量。当然，最好的办法是存储数目字在字串中变化的规则，使之能够自动生成所有可能的变体。关于这一点将在下文讨论。

2. 语样选取的基本原则

2.1. 随机取样与定向编辑的灵活运用

随机取样是最常用的一种选样方法，就是通过随机采用某一个或某几个具体学科领域里大量的现成语料(如广播、报刊、书籍文献上的大段话语)，来达到自然地覆盖这些领域里常用词语和专门术语语言样本或某种结构模式的目的。这种方法的好处在于，第一，语料来源广泛，自然度好，比较接近自然语言的统计特性；第二，由于不加人工干预，选样时就不必详细了解或费心考虑使用特定的语言结构规则。然而，这种选样方法在某种意义上说具有一定的盲目性，一方面不可避免地产生相同语样的大量重复，造成不必要的大存储量；另一方面，很可能所收的语样不够全面典型，由于随机取样所得的自然语料本身的局限性，有些在实际的自然语言里存在着的结构模式或规则难免被漏选。这是这种选样方法的主要弱点，而定向编辑正好可以克服这方面的不足。

定向编辑并不是简单地直接使用现成的自然语料，而是根据某种语言特定的结构规则，从自然语料中有计划地选用或有目的地编辑所需的语样，因而不但可以保证所选语样的全面典型，而且简明经济。当然，这种定向编辑的语料，在某些方面可能不如随机取样所得的语料来得自然。譬如说，为了保证全面典型，有时不得不收编少数比较生僻的词语。

随机取样与定向编辑的优、缺点是相对而言的，究竟应该采取哪种方式，还是应该根据具体语料库的建库目的而定。譬如说，某个部门正在建立一个特大规模的(约 50,000,000 字)综合性现代汉语语料库，其目的是适用于所有学科领域，所收的语料要基本上反映现代汉语各方面的自然平衡。因此，他们采用的是随机取样的方法，组织大批人力，尽可能多地收集各相关领域的典型文献资料，然后用随机抽样的方法抽取其中的部分文献中的部分语段，避免人工干预。这个库是个文字资料库，自然语言处理的有关部门，特别是涉及语义和语法研究的方面，都可望从中调用相关的语料。例如，你想了解汉语里十个数目字的使用频度，或许通过调用该库的相关语料进行分析，就能获得满意的统计结果。然而，假如你需要研究语音的特性，例如要了解连续话语里数目字串的协同发音规则的话，那末，即使是这样大规模的语料库也未必能完全概括与此相关的语音库所需要的语样。在这种情况下，最好还

是采用定向编辑的方法建立一个专门的数字串语料库。这是因为，汉语里与数目字相关的音节虽然不多，充其量也就 20 来个。但是，一旦构成数目字串，情况就不同了。可以想象，实际上可能存在的长、短不等的数目字串结构的数量必将大得惊人，不是任何通过随机取样的语料库所能全面覆盖的。如若根据一定的语音结构规则来编辑这部分语料，充其量用大约几百个两音节数字串就能全面覆盖所有可能的数目字间协同发音的规则了。

2.2. 定向编辑的基本原则及选样举例

通常，为某一特定领域或特定目的而建立的语料库都会不同程度地采用定向编缉的方法来选取语料。例如，美国奥力根言语理解中心建立的"姓名拚读电话言语语料库"的语料就是首先编制了一个特定的问题单，然后通过电话让许多说话人按所要求的方式拚读他们的姓名而选取的[4]。

尽管由于具体的建库目的不同，语料库的种类不同，规模也大小不一。但是有一点是共同的，即都是作为反映自然语言某一方面或某些方面特性的代表，成为体现这些方面规则系统之大成。因此，语料选取的一条基本原则应该是确保模式全面、语料典型，使之成为代表自然语言某种局部规则体系的缩影。兹以普通话两音节结构语音库的语样选取为例加以说明。

2.2.1 根据协同发音规则和汉语语音结构特点选样，确保模式全面

人们发现，在连续话语中，我们通常所熟悉的音节的语音模式发生了复杂的变化，形成了各种各样的随机变体。为了解决连续话语合成中语音的自然度问题和自动识别中词或音节的边界定位及切分问题，人们迫切希望掌握音节在连续话语中所有可能出现的变体模式。然而，这些变体随语境而定，存在于具体的自然语言的汪洋大海之中，要想掌握它们，的确不是一件容易的事。譬如，以"料"这个音节为例，它在两音节词"材料"和"料理"中分别体现为两个不同的变体。而到了"材料理论"这样的多音节结构或片语里，"料"这个音节又由于前后语音环境的不同而体现为另外一种变体。可以想见，它在别的语音环境中又会以新的面貌出现。普通话里一共有 1300 多个不同的音节，每个音节随时有可能同包括它自身在内的其它音节一起构成两音节或三音节结构，从而产生无数个不同的变体。显然，要想从自然语言的汪洋大海中把每个音节的一个个具体的变体都找出来那是不现实的。当然，这个问题也并非无法解决。在连续话语里，无论实际上出现的协同发音关系多末错综复杂，但真正涉及的不外乎是一连串的两个相邻音节之间的协同发音效应。因此，我们可以首先建立一个两音节结构语料库，从中提取所有可能存在的不同音节之间的协同发音模式，然后，每一个音节的具体变体就可利用相关的协同发音模式来生成了。

那么，就普通话而言，究竟用多少种两音节结构才能概括所有可能存在的音节间的协同发音模式呢？从普通话的基本语音结构来看，即使不计声调区别，也有 400 多种声韵组合，即 400 多个独立的音节。在自然语言里，这些音节都有可能处于彼此相邻的位置上，而且，不管是以两音节的词、词组的形式出现，或者只是作为词间或短语间毗邻的两个音节，它们之间都会产生协同发音的效应。因此，最简单、最保险的办法就是用数学上的排列组合来计算，那末即使不计音节的四声区别，这个语料库也起码得收十几万个两音节结构。显然，这样的

语料库肯定是模式全面，语料翔实，无论是编辑合成还是模式识别，都可从中获取典型样板。然而，在大多数情况下，这末大规模的语料库未必都是现实的和必要的，因而可以根据语言学的原则大大加以压缩。

2.2.2. 根据语音学上简约的原则选样，确保语料经济典型

所谓经济典型，就是力图用最少的典型语料来代表最多的、可能出现的语音组合模式。

根据协同发音的一般规律和汉语语音的结构特点，一个音节的不同变体主要体现在两个方面，第一，它的声母可能因前临音节韵母的不同而发生不同的变化；第二，它的韵母因后接音节声母的不同而发生不同的变化。但是，无论是声母的变化还是韵母的变化，归根结缔都体现为前一音节的韵母与后一音节的声母之间的音联关系[5]。根据普通话语音的音位配列规则和相关的结构特性，在选取两音节结构语料库的语样时应该遵循这样一个基本原则：作为两音节结构的前音节，可以不考虑其声母的异同，但必须确保其韵母的全面性；相反，作为这个结构的后音节，则必须确保其声母的全面性。这样，大约最多用一万多个两音节结构就能包括普通话里所有可能出现的音节间音联的模式了。这个库的特点是模式典型、全面，语料相对精炼，比较适用于规则合成或语音学的基础研究，可以据此进行声学分析，从中提取必要的特征参量和统计模式，建立相关的参数数据库，等等。

其次，视具体应用目的的不同，这个库还可以大大压缩。譬如说，假如不考虑两音结构中后音节韵母的不同对于音节间协同发音可能产生的影响的话，那末，这个库起码可以压缩到两千多个词条，这样可以大大节省存储量，而且，如果主要作为规则归纳用的话，这个库的规模还可以进一步压缩。因为从语音学的角度看，与两个音节之间协同发音直接相关的是前音节的尾音音段与后音节的首音音段之间的关系，某些相关研究表明[6,7]，普通话里音段之间的相互关系主要表现为后音段对前音段的影响；在音节之间，涉及的主要也是后音节的首音音段对于前音节的尾音音段逆向的协同发音作用，而同一类尾音与首音之间的音联模式是相似的。因此，可以根据语音学上简约的原则，对这些尾音与首音加以分类归纳，然后分别选取代表语样，大约用一、二百个词条就足以概括典型的音联模式了。

2.2.3 统筹考虑语音、语法和语义之间的相互制约关系，避免语料的片面性

自然语言是语音、语法和语义的统一体。语义是语言信息的底层内涵，语音是它的表层表达，语法则是语言信息由底层向表层转换的中介。因此，语音的结合和变化必然伴随着语法和语义的制约。兹以普通话里变调和轻声有关的现象为例略加说明。

在汉语普通话里最复杂的变调现象要数"上上相连"了，这是普通话里最重要的变调规则。它的基本规则是两个上声（即第三声）音节相连，第一个上声会变读为阳平。但是，这里有一个条件，即这两个上声音节必须都是读正常重音的，否则，第一个上声就可能不是变成阳平，而是变成半上或保持不变。譬如，"椅·子"、"老·子"和"耳·朵"虽然也都是上上相连，但是，其中的第一个上声并没有变为阳平。当然，"老子"中的第一个字也可以变读阳平调，但这时，它后面的"子"字必须重读，而且，词义也变了，所以跟"老·子"不是一回事。其次，由一个上声音节通过重叠法组成的两音节结构，它们的变调形式也有两种，具体取决于这种结构的词性和轻重音格式。以上这些都是语音、语义和语法相互制约的具体表现。因为

在普通话里，轻声的存在不仅是个重要的韵律特征，有时还有区别语义的作用，是语法上的构词手段之一。这种轻声词的语音形式不但是相对稳定的，而且还会影响连读变调。因此，在编辑反映"上上相连"变调规则的语料时，就不能单纯注意音节本身的声调，而必须考虑可能存在的语法或语义方面 的影响，以避免所选语料的片面性。

在选用语句的语料时，就更得注意语义和语法的影响了。因为，即使由同样音节系列构成的句子，由于底层语义内涵的不同，会通过一定的语法制约影响到表层的语音特性的变化。因此，同样的音节系列并不一定能生成同样的语音变化模式。当然，这里涉及语句中的重音和时长分布与 节奏和韵律的关系问题，虽然不属于本文讨论的范围，但却同样表明，语料库的语样选取必须注意语音、语义和语法的相互制约关系。

3. 小结：

综上所述，建立语料库是与自然语言打交道，因此，在选取语料时就或多或少、不可避免地要运用语言学的原则，还要了解具体语言的结构特性和规则。只有自觉地了解和运用这些特性及规则，才能建立起比较理想的、合乎自然语言特性的语料库。

参 考 文 献

[1] P. Ladefoged(1992). Knowing enough to analyze spoken language, The Proceedings of ICSLP'92, Vol.1:1－4.

[2] P. Keating(1988). Coarticulation and timing, UCLA Working Papers in Phonetics, Vol. 69:1－2.

[3] 曹剑芬(1990).《现代语音基础知识》，人民教育出版社，北京，1990，130－132 页。

[4] R. Cole, K. Roginski, M. Fanty(1992). A telephone speech database of spelled and spoken names, Proceedings of ICSLP'92, Vol.2:891－893.

[5] 许毅(1986)."普通话音联的声学语音学特性",《中国语文》，1986 年第五期。

[6] 曹剑芬、杨顺安(1984)."北京话复合元音的实验研究",《中国语文》，1984 年第六期。

[7] 杨顺安(1989)."普通话音节间的协同调音及其合成模拟",《语言研究所语音研究报告》，1989 年。