A Method for Extracting Formulaic Sequences from a Student Corpus

Malcolm Prentice

Use of appropriate formulaic sequences can add fluency, accuracy and appropriacy to written English, and one important place these sequences occur are as sentence starters such as “Needless to say” or “At the same time”. This article describes and provides open source code for a tool created using Python and the Natural Language Toolkit which can help identify formulaic sentence-starters in an untagged corpus of student writing, for use in progress measurement and course design. Example results from two corpora are presented and discussed.

Keywords: formulaic sequence, Natural Language Processing, student corpora, writing, EFL

Introduction

This project was originally suggested by a colleague who wanted to investigate the way students identified as “good writers” were starting their sentences in comparison with other students in the same context. One possible difference was in the usage of formulaic sequences. Since hand-compiling a list of the first words in thousands of sentences is impractical, and since no known tool could help, a new one was created using the programming language Python, supported by modules from the Natural Language Toolkit (NLTK). This article describes that tool, which can be downloaded from http://code.google.com/p/first-five-words/

While several tools for finding word clusters exist, none have the flexibility needed to perform this specific task. For example, the "cluster" function on the WordList program (one of the Wordsmith tools available from http://www.lexically.net/) is aware of sentence breaks but cannot deliberately target sentence starters. As a closed source program, it and similar programs available cannot be adapted to the purpose. By comparison, a small open-source script such as the one described here can be quickly written to target a specific task, and can later be modified to handle any non-standard requirements, such as unusual text encodings, XML genre tags and filter lists. Everything described in this article is free to download, and the license under which the code has been released allows users to freely adapt and distribute any new version.

The practical application of this tool is rapidly identifying useful formulaic sequences in specific sentence positions in a target corpus and/or a collection of student written work. This allows the creation of lists of target language, allows comparison of target and current usage (as part of a needs analysis); and allows comparison of current and former usage (as a measure of instructional success or progress). The remainder of this article first looks at the two example corpora that will be used and reviews some literature on formulaic sequences, then outlines the procedure for using the tool, and finally discusses results from running the script on the two example corpora.

The corpora used

Research on formulaic sequences in large, part-of-speech (POS) tagged corpora of native English speaker writing already exists. For example, Liu (2011) identifies a list of 228 "multi-word constructions" in the academic writing sub-corpora of the Corpus of Contemporary American English and the British National Corpus, while Byrd and Coxhead (2010) lists 21 useful "lexical bundles" found in the academic corpus used to produce the Academic Word List (Coxhead, 2000).

However, while some authors consider native speaker corpora as providing a valid target against which non-native speakers can be measured (e.g. Shirato & Stapleton, 2007), others question how appropriate language taken from a native speaker corpus can be for EFL learners' needs, level and sociocultural context (Huang, 2011). Tan (2005) for example argues that what seem to be overuse, underuse or misuse mistakes as defined by frequency norms based on native English-speaker corpora can actually be valid ways of adapting the language to express concepts specific to the local culture. Therefore, the best writing corpus for a group of language learners may be a corpus of writing by slightly more advanced learners, if possible by learners from the same context. Two such corpora are analysed in this article.

The first corpus used, of writing by "more advanced learners", is the Uppsala Student English (USE) corpus compiled between 1999-2001 by Margareta Westergren Axelsson and Ylva Berglund at Uppsala University, which contains 1,497 essays on various topics in various genres averaging around 800 words written by 440 mostly first year Swedish university students of English. The essays were collected to assist needs analysis—"diagnosing the language difficulties our students experience at different levels" (Axelsson, 2000). Files are available from the Oxford Text Archive (http://ota.ahds.ac.uk/headers/2457.xml).

The second corpus used, of writing by "more advanced learners in the same context", is what will from here on be called the "Near Peer Student Competition" (NPSC) corpus. It is a growing collection of prize-winning essays from a Japanese university competition—currently 90 essays from 3 years—ultimately intended to help lower-level, younger students from the same university to write essays in the same genre. This is the corpus mentioned in the introduction as being the impetus, via a colleague, for creating the tool described below. Work on this project is on-going and results will be published in due course—the emphasis in this article is on testing the strengths and weaknesses of the tool itself.

Mention is also made below of the *British Academic Written English* corpus (BAWE), in order to contrast small corpora of EFL student work with a large corpus of native-English speaking students' academic essays. The BAWE corpus consists of 2761 undergraduate and postgraduate assignments written between 2004-2007 by 1039 students at the universities of Warwick, Reading and Oxford Brooke. The essays cover 35 disciplines and 13 genres of academic work, meaning the texts could be anything from a psychology literature review to an engineering specification. Files are available online from the Oxford Text Archive (http://ota.ahds.ac.uk/headers/2539.xml).

Defining and identifying formulaic sequences

Wray (2002, p.9) identified 61 terms that are synonymous or similar in meaning to “formulaic sequence”, the most common of which are “chunk” and “multi-word unit”. Another common term not on the list is “Lexical bundle”. A related term is “collocation”, but while those are usually operationalized very clearly (in terms of dispersion, frequency, statistical significance and words to be excluded) the definition of “formulaic sequence” tends be less precise. As a result, estimates of how much language can be considered formulaic varies from 4% to 80% (Wray, 2002, p.28). Read and Nation (2004, p.24) note that the concept of “a word” is difficult enough to define, even before the words are combined into sequences. Schmitt and Carter (2004) choose to avoid a definition and instead offer “useful characteristics which are typical of formulaic sequences” but which not every formulaic sequence need possess, such as semantic prosody and holistic storage. Wray (2002, p.34) lists other aspects that could form part of a definition, including structure, compositionality, fixedness, phonological form, fluency, and stress.

Unfortunately, the characteristics listed above do not translate easily into a method for extracting formulaic language from a corpus, but at the same time only so much data can be processed by hand. Most authors recommend triangulation between qualitative (human) and quantitative (computer) methods (Bird & Coxhead, 2010, Read & Nation, 2004, Wray, 2002).

However, that initial computer search is not straightforward, as frequency and dispersion are not an especially good indicator of formulaicity (Wray, 2002; Read & Nation, 2004; Liu, 2011). As Dörnyei (2009, p.297) puts it, “not every frequently co-occurring string of words forms a chunked whole on the one hand, and not every formulaic sequence is all that frequent on the other”. Formulaic sequences can have “slots” which allow substitutions, for example *“x percent of (determiner + noun)*” (Liu, 2011), and even in POS-tagged corpora it is only really possible to search for non-contiguous sequences if the target is already known (Read & Nation, 2004, p.32). The small teacher-collected corpora that the tool described below is intended to process are unlikely to be POS-tagged, so “slot” searches like this would be impossible. In any case, there is no consensus over the point at which such changes become so extensive that the sequence is no longer formulaic. Read and Nation (2004, p.25) suggest that expressions which allow insertions, inflection, substitution, deletion *and* transformation are too flexible to be formulaic, but are unclear as to where it is on the line between flexible and fixed that “formulaic” begins. The current script can only identify fixed expressions.

Additionally, normal statistical methods for identifying significant combinations are not usable, as they work by comparing the sequence components’ co-occurrence with the components’ separate occurrence in a full text. By focusing on sentence starters we select a smaller sample of a smaller population *within* a text, and the size of that population is difficult to determine. If a 5-gram sentence starter contains a 4-gram formulaic expression plus a comma, is its significance to be measured against a population of unigrams, 4-grams or 5-grams? Is it necessary to allow for the fact that one side of the n-gram is locked in place at the beginning of a sentence? A new definition of statistical significance is needed.

For now, this article assumes that sentence starters such as “In this essay I will” are potentially useful chunks of target language, that a frequency search is a necessary first step in extracting the language from a corpus, and that while there are unresolved issues this approach is still preferable to sorting hundreds of sentence fragments by hand.

The benefits of formulaic language

There are a number of reasons why formulaic sequences deserve attention. Firstly, “grammar on its own will overgenerate acceptable strings” (Wray, 2002, p.15)—there are a large number of grammatically correct alternatives for each formulaic sequence described in this article but most of them would sound strange to a native speaker. While native-like speech may not be every learner’s aim, most would rather not put effort into creating a novel way of expressing a meaning that is more commonly represented with a formulaic sequence, especially if using that sequence reduces the chance of error (Boers, Eyckmans, Kappel, Stengers, & Demecheleer, 2006). Secondly, it is “more efficient and effective to retrieve a prefabricated string than create a novel one” (Wray, 2002, p.18). Formulaic sequences are retrieved as ready made chunks from declarative memory, without the need to process individual words into utterances and as such can increase fluency (Wray, 2002, p.189; Segalowitz, 2010, pp.33-34) as their use “allows the speaker to attend to other aspects of communication and to plan larger pieces of discourse” (Dörnyei, 2009, p.294). While time pressure is less of an issue in writing, there are still situations in which fluency is important and formulaic sequences may help, such as exams with time limits, looming deadlines, large volumes of email, and synchronous written communication tools such as Skype, Yahoo Messenger or Google Talk. Wray (2002, p.84) also suggests that formulaic sequences can help the reader comprehend by marking discourse structure. Finally, formulaic language can also indicate membership of a community, such as the use of academic English by academics (Segalowitz, 2010).

In summary, a student who learns and uses the sequence “In this essay I will” has saved time and energy that can be used elsewhere, has put at least 5 correct tokens on paper, has made it easier for the reader to understand the first sentence, and has self-identified as someone capable of following signposting conventions in a “first person opinion essay” genre.

Teachability of formulaic language

There is limited research on the teachability of formulaic language—Dörnyei (2009, p.297) suggests this is because of the difficulty of operationalizing the concept. Boers et al. (2006) found that formulaic language can be taught through awareness raising—learners who underlined chunks while reading were later able to use those chunks in speech, and as a result were judged to be more fluent. Taguchi (2007) confirms that formulaic sequences can also be taught directly, although the instruction in that study was limited to small Japanese "grammatical units". One issue is that while it is possible to give a clear threshold for which *collocations* are frequent enough to be worth attention (Shin & Nation, 2008), formulaic sequences may be difficult to count in the same way.

Wray (2002, p.280) warns that acquiring formulaic sequences is a complex process, and teaching them could in some cases be detrimental. At the very least, when formulaic sequences are made the subject of instruction or awareness raising, care should be taken that students do not confuse "frequent and possibly useful" with "compulsory". In terms of technique, Byrd and Coxhead (2010) suggest offering "multiple focused encounters in context and in classroom" to supplement incidental exposure, the use of notebooks, regular revision, and ensuring the students understand why formulaic sequences are valuable.

Method

Preparation

Teachers wishing to use this tool must first collect an electronic copy of student written work. If teachers require students to word-process their essays, then even if a printed copy is used for marking, students can be asked to send an email attachment of their final draft. This will probably be a Microsoft Word document, which must be converted into a plain text file since formatted documents contain a lot of invisible information that only the original program can interpret. This need not be a labour intensive process of opening and "Saving As" files one by one. Mac users can use Automator, and Windows users a Word Macro, to batch convert an entire folder of Microsoft Word documents to text files. Once the essays to be analysed are in a folder in plain text (txt) format, the process is as follows:

Procedure

1) If using Windows, download and install Python from http://python.org. Python comes pre-installed on Mac and Linux.
2) Download and install the Natural Language Toolkit from http://www.nltk.org. Follow the instructions to download all supporting data files, including sentence tokenization models.
3) Download the most recent version of the script ("first-five-words.py") from http://code.google.com/p/first-five-words/
4) Open a terminal window, change directory (*cd*) to where the script is and type "python first-five-words. py"
5) After a few seconds, a dialogue box will appear to allow you to choose which folder the files are in.

The script processes around 100 short essays per second, producing three filtered CSV files for 3- 4- & 5-gram sentence starters (where "gram" is defined as any token including punctuation). The CSV files are formatted for Microsoft Excel. If being used on extremely large corpora (tens of thousands of files), it might be necessary to apply a minimum occurrence threshold. The threshold is currently set at 0, but this can be changed by opening the script with a text editor.

Results: Example analysis of two learner corpora

The top twenty 3- 4- & 5-gram sentence starters from the USE and NPSC corpora are given in the appendix. The aim is not to suggest language for teaching, but rather to illustrate the characteristics of the data produced. Teachers are recommended to run the script themselves on an appropriate corpus for their students, selecting the useful formulaic language and discarding the high frequency non-formulaic chunks and sentence tokenization errors.

Summary statistics on corpora contents

Below is some information on the corpora—how many sentences there are, how many different sentence starter "types" were found, and what percentage of the sentences in each corpus are started by a sequence from the top twenty results.

NPSC (3598 sentences)

3 grams	***2823 types***
4 grams	***3365 types***
5 grams	***3528 types***

The top 20 NPSC 3-grams start ***7.84%*** *of all sentences in the corpus.*
The top 20 NPSC 4-grams start ***2.92%*** *of all sentences in the corpus.*
The top 20 NPSC 5-grams start ***0.64%*** *of all sentences in the corpus.*

USE (59605 sentences)

3 grams	***37743 types***
4 grams	***50287 types***
5 grams	***55718 types***

The top 20 Uppsala 3-grams start ***3.73%*** *of all sentences in the corpus.*
The top 20 Uppsala 4-grams start ***1.95%*** *of all sentences in the corpus.*
The top 20 Uppsala 5-grams start ***0.85%*** *of all sentences in the corpus.*

Discussion

Description of the data extracted from the Uppsala and NPSC corpora

Although two or three of the Uppsala n-gram sentence starters are sentence tokenization errors, they have been left in to illustrate the kind of data that teachers will be handling when they process their own choice of corpora. With or without these errors, the top results cover a reasonable percentage of the sentences in the corpus. This suggests that, in the context of teaching students how to start a sentence, a number of them would be well worth deliberate attention in class.

However, there is some noise in the USE corpus results, and significant problems with the BAWE corpus. The "Punkt" sentence tokenization method (Kiss & Strunk, 2006) used in the script is the best available method, and it handles basic non-sentences such as abbreviations well. However, the method requires a trained sentence model, and the model currently used is trained on newspaper rather than academic English. It seems to have problems with citation fragments such as ", 2005)." and with numbered lists. This is not a problem in the NPSC corpus, which contains only fully formed sentences with no references, and it is only a minor problem in the USE corpus. However, so many of the top 20 BAWE results were tokenization errors that the results were not worth presenting in the appendix. Tokenisation errors skew the coverage percentages and hide the target language, and so if the tool is to be used for academic writing corpora rather than composition essays, it would be worth training a tokenizer for the purpose.

The data show why 3-gram, 4-gram *and* 5-gram searches are run: each captures different formulaic phrases. The chunk "*For example ,*" is the top 3-gram results but variation in the 4th and 5th tokens dilute its count in the other lists, while "*On the other hand (,)*" only appears on the NPSC 4 and 5 lists.

Hand cleaning is necessary, as sentence tokenization errors and high frequency non-formulaic or "structurally/semantically incomplete" (Liu, 2010) sequences are present. As mentioned above, the amount of work that needs done may depend on the suitability of the sentence tokenizer for the corpus. Corpora may also miss or overrepresent certain formulaic expressions (Wray, 2002, p.27), and this becomes worse the smaller the corpus. The solution is choosing an appropriate corpus or sub-corpus for students' needs—for example, as the USE corpus clearly contains a lot of first person opinion essays ("*I think that*"), a sub-corpus of the BAWE might be better for teachers wanting to help students with academic dissertations.

Planned improvements

Firstly, XML tags. These are labels put either side of a token or sentence (e.g. "*<title> The Fifth Child </title>*") to carry information on the text between them. For the purpose of this article—handling small untagged collections of student work-tags were thought to be unnecessary. However, the USE corpus and BAWE corpus both contain XML metadata tags (currently ignored by the script) that could enable improving the results with a measure of dispersion (see below). In larger corpora (such as the British National Corpus) and in the other versions of the BAWE, tags also surround words and sentences, carrying information on part of speech and other linguistic information. Making the script XML-aware would allow it to use these corpora, and possibly even to run searches for formulaic sequences with "slots". Both Python and NLTK have tools for handling XML tags.

Secondly, large corpora. Text files containing too many lines are not loaded by standard spreadsheet software-Excel for example truncates files at the 65536th line. This is not a problem for corpora up to USE size, but to fully analyse BAWE size corpora or larger would require more of the analysis to be built into the script itself.

Thirdly, dispersion. A single essay using anaphora (in the rhetorical sense) has pushed the phrase "*Let freedom ring from (the)*" into the top results. While human checking can easily spot this, it would also be possible to add a count of how many essays each phrase appears in and only allow phrases that have counts over a certain dispersion threshold. Similarly, multiple essays in the USE corpus on Doris Lessing's novel "*The Fifth Child*" has forced the phrase "*The Lovatts were a happy family*" into the top rank. A simple essay count would not detect this, but using an XML-aware script could allow a dispersion threshold to be applied using genre metadata tags.

Finally, filter lists. In order to avoid repeatedly hand checking the same language, a filter list could be added to the script to allow teachers to collect a list of "not formulaic" and "already collected" sequences to be excluded from subsequent analysis.

Conclusion

The current script is a good alternative to compiling thousands of sentence fragments by hand, and produces a list of candidate n-gram sentence starters from which useful formulaic sequences can be chosen. However, some work is still needed before it can be used effectively to analyse certain types of corpora, or to find non-fixed formulaic sequences. The next step, in collaboration with a colleague, is to attempt the project which prompted the creation of this script—to actually use the formulaic language extracted from the NPSC corpus with a group of lower level students in the same context and discover whether they can acquire the sequences and whether the quality of their writing improves as a result.

References

Axelsson, M. (2000). USE-The Uppsala Student English Corpus: An instrument for needs analysis. *ICAME Journal*, *24*, 155-157.

Boers, F., Eyckmans, J., Kappel, J., Stengers, H., & Demecheleer, M. (2006). Formulaic sequences and perceived oral proficiency: putting a Lexical Approach to the test. *Language Teaching Research*, *10*, 245-261.

Byrd, P., & Coxhead, A. (2010). On the other hand: Lexical bundles in academic writing and in the teaching of EAP. *University of Sydney Papers in TESOL*, *5*, 31-64.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*, 213-238.

Dörnyei, Z. (2009). *The Psychology of Second Language Acquisition*. Oxford: OUP.

Kiss, T., & Strunk, J. (2006). Unsupervised Multilingual Sentence Boundary Detection. *Computational Linguistics*, *32*, 485-525.

Read, J., & Nation, P. (2004). Measurement of formulaic sequences. In N. Schmitt (Ed). *Formulaic Sequences (pp.23-35)*. Amsterdam: John Benjamins.

Segalowitz, N. (2010). *Cognitive Bases of Second Language Fluency*. New York: Routledge.

Shirato, J., & Stapleton, P. (2007). Comparing English vocabulary in a spoken learner corpus with a native speaker corpus: Pedagogical implications arising from an empirical study in Japan. *Language Teaching Research*, *11*, 393-412.

Schmitt, N., & Carter, R. (2004). Formulaic sequences in action: An introduction. In N. Schmitt (Ed). *Formulaic Sequences (pp.1-22)*. Amsterdam: John Benjamins.

Taguchi, M. (2007). Chunk learning and the development of spoken discourse in a Japanese as a foreign language classroom. *Language Teaching Research*, *11*, 433-457.

Tan, M. (2005). Authentic language or language errors? Lessons from a learner corpus. *ELT Journal*, *59*, 126-134.

Wray, A. (2002). *Formulaic Language and the Lexicon*. Cambridge: CUP.

Appendix: Data on Uppsala and NPSC corpora

Table 1 Top twenty 3-gram sentence starters from the USE Corpus, with count, occurrence as % of 3-gram sentence starters in corpus, and cumulative coverage of sentence starters.

USE 3-gram	Count	% Occurrence	% Cumulative
I think that	269	0.45	
“ (p	167	0.28	0.73
I do not	152	0.26	0.99
.	148	0.25	1.23
I believe that	148	0.25	1.48
When it comes	139	0.23	1.72
It is not	136	0.23	1.94
In this essay	129	0.22	2.16
On the other	127	0.21	2.37
It is a	116	0.19	2.57
One of the	114	0.19	2.76
The fact that	110	0.18	2.94
This is a	90	0.15	3.10
I think it	88	0.15	3.24
I don’t think	81	0.14	3.38
In my opinion	79	0.13	3.51
Of course ,	76	0.13	3.64
It is also	75	0.13	3.76
First of all	74	0.12	3.89
Harriet and David	72	0.12	4.01

Table 2 Top twenty 3-gram sentence starters from the USE corpus, with count, occurrence as % of 4-gram sentence starters in corpus, and cumulative coverage of sentence starters.

USE 4-gram	Count	% Occurrence	% Cumulative
.	148	0.25	
" (p .	137	0.23	0.48
When it comes to	137	0.23	0.71
On the other hand	125	0.21	0.92
In this essay I	103	0.17	1.09
I think it is	65	0.11	1.20
I do not think	52	0.09	1.29
At the same time	43	0.07	1.36
I think that the	42	0.07	1.43
I would like to	41	0.07	1.50
To sum up ,	41	0.07	1.57
I would say that	37	0.06	1.63
In other words ,	37	0.06	1.69
First of all ,	36	0.06	1.75
In the beginning of	34	0.06	1.81
I will try to	33	0.06	1.86
It was what they	33	0.06	1.92
As a matter of	32	0.05	1.97
On the contrary ,	29	0.05	2.02
I believe that the	28	0.05	2.07

Table 3 Top twenty 5-gram sentence starters from the USE Corpus, with count, occurrence as % of 5-gram sentence starters in corpus, and cumulative coverage of sentence starters.

USE 5-gram	Count	% Occurrence	% Cumulative
.	148	0.25	
In this essay I will	71	0.12	0.37
On the other hand ,	58	0.10	0.46
It was what they had	33	0.06	0.52
As a matter of fact	31	0.05	0.57
When it comes to reading	28	0.05	0.62
2 .	26	0.04	0.66
I do not think that	25	0.04	0.70
In the beginning of the	24	0.04	0.74
There are a lot of	24	0.04	0.79
A happy family .	20	0.03	0.82
3 .	19	0.03	0.85
On the other hand I	18	0.03	0.88
The Lovatts were a happy	18	0.03	0.91
When it comes to the	18	0.03	0.94
I think that it is	16	0.03	0.97
1993 .	15	0.03	0.99
4 .	15	0.03	1.02
But on the other hand	15	0.03	1.04
This is one of the	15	0.03	1.07

Table 4 Top twenty 3-gram sentence starters from the NPSC corpus, with count, occurrence as % of 3-gram sentence starters in corpus, and cumulative coverage of sentence starters.

NPSC 3-gram	Count	% Occurrence	% Cumulative
For example ,	38	1.06	
In addition ,	26	0.72	1.78
However, I	22	0.61	2.39
Of course ,	19	0.53	2.92
When I was	18	0.50	3.42
There are many	15	0.42	3.84
I want to	14	0.39	4.22
Do you know	13	0.36	4.59
I think that	12	0.33	4.92
I think the	12	0.33	5.25
In fact ,	11	0.31	5.56
As a result	10	0.28	5.84
Have you ever	10	0.28	6.11
I think it	10	0.28	6.39
Also, I	9	0.25	6.64
However, the	9	0.25	6.89
In Japan ,	9	0.25	7.14
One day ,	9	0.25	7.39
First of all	8	0.22	7.62
I was so	8	0.22	7.84

Table 5 Top twenty 4-gram sentence starters from the NPSC corpus with count, occurrence as % of 4-gram sentence starters in corpus, and cumulative coverage of sentence starters.

NPSC 4-gram	Count	% Occurrence	% Cumulative
As a result ,	10	0.28	
First of all ,	8	0.22	0.50
I think it is	8	0.22	0.72
Let freedom ring from	8	0.22	0.94
On the other hand	8	0.22	1.17
When I was in	7	0.19	1.36
In my opinion ,	5	0.14	1.50
At that time ,	4	0.11	1.61
At the same time	4	0.11	1.72
Do you know the	4	0.11	1.83
For example, when	4	0.11	1.95
In addition, there	4	0.11	2.06
In other words ,	4	0.11	2.17
It is said that	4	0.11	2.28
One day, I	4	0.11	2.39
The most important thing	4	0.11	2.50
There are a lot	4	0.11	2.61
This summer, I	4	0.11	2.72
When I was a	4	0.11	2.83
“ When I heard	3	0.08	2.92

Table 6 Top twenty 5-gram sentence starters from the NPSC corpus with count, occurrence as % of 5-gram sentence starters in corpus, and cumulative coverage of sentence starters.

NPSC 5-gram	Count	% Occurrence	% Cumulative
On the other hand ,	8	0.22	
First of all, I	4	0.11	0.33
Let freedom ring from the	4	0.11	0.44
There are a lot of	4	0.11	0.56
This summer, I went	4	0.11	0.67
As you can see ,	3	0.08	0.75
At the same time ,	3	0.08	0.83
In addition, there are	3	0.08	0.92
The most important thing is	3	0.08	1.00
What if you have to	3	0.08	1.08
“ What do you think	2	0.06	1.14
A lot of animals lost	2	0.06	1.20
As a result, the	2	0.06	1.25
Ashley said, “ I	2	0.06	1.31
At that time, my	2	0.06	1.36
But not only that .	2	0.06	1.42
For example, a lot	2	0.06	1.47
For example, they pick	2	0.06	1.53
I can also find a	2	0.06	1.58
I didn’t know what I	2	0.06	1.64