




广东外语外贸大学
 GUANGDONG UNIVERSITY OF FOREIGN STUDIES

外国语言学及应用语言学研究中心
 Centre for Linguistics & Applied Linguistics

Guangwai-Lancaster Chinese Learner Corpus: A Profile

Hai Xu¹, Vaclav Brezina²

¹ Guangdong University of Foreign Studies ² Lancaster University

Acknowledgements

- ▶ This study is supported by the British Academy International Partnership & Mobility 2013 Scheme.
- ▶ In memory of our great colleague and friend Dr. Richard Xiao (1966–2006), who is the founder of the project.
- ▶ Grateful to our team members and student assistants.

Background

- ▶ Why bother with a learner corpus of Chinese?
 - Most of learner corpora that have been constructed so far are confined to English learner corpora (such as ICLE, ICNAE, CLEC, SWECCCL) (Pravec 2002).
 - With the rise of China as a global power, Chinese as a major world language has become an increasingly popular foreign language.
 - A learner corpus plays an instrumental role in teaching and learning Chinese. However, few Chinese learner corpora are publicly available, and they have the limitations:
 - Small in size
 - Limited data types
 - Biased towards Korean, Japanese and Southeast Asian learners

Background (cont'd)

- ▶ Publically available Chinese learner corpora
 - **HSK动态作文语料库** (Monitoring Corpus of HSK Essays)
 - By Beijing Language and Culture University (北京语言大学)
 - **留学生书面语料库、口语语料库** (The Overseas Students Written and Spoken Chinese Corpora)
 - By College of Chinese Language and Culture, Jinan University in Guangzhou (暨南大学华文学院)



HSK 动态作文语料库 ver.1.1

退出 | 注册

首页 字符串检索 模糊检索 错词检索 全文检索 统计 反馈 帮助

字符串检索

按“字”查询 按“词”查询 按“字符串”查询

所有 错误

条件一 ==请选择属性==

条件二 ==请选择属性==

条件三 ==请选择属性==

条件四 ==请选择属性==

条件五 ==请选择属性==

开始搜索 属性设置 重置

HSK动态作文语料库

- ▶ Size: 4.24 million Chinese characters
- ▶ Data type: only one specific type, i.e., written essays by HSK test takers
- ▶ Data produced by L2 Chinese learners from 101 countries between 1992 and 2005
 - Altogether 11,569 pieces of essays
 - Skewed towards Korean (4,171), Japanese (3,211), Singapore (843), Indonesian (739), Malaysian (422), Thai (374), Vietnamese (227), Burmese (202) speakers
- ▶ Corpus data have been annotated on the levels of Chinese characters, phrases, sentence patterns, etc., but the annotation scheme is quite complicated
- ▶ Chinese characters not segmented yet
- ▶ The online version offers 'concordance' function, but no wordlist, *n*-grams, keyness, etc.

[illegible][illegible]The logo for the Guangwai Lancaster Chinese Learner Corpus (CLC) is a stylized open book. The left page is red and contains the text "Guangwai Lancaster Chinese Learner Corpus" in white. The right page is blue and contains the text "CLC" in white.

Guangwai-Lancaster Chinese Learner Corpus (CLC): Goals

- ▶ To develop a one-million-token balanced corpus
- ▶ To meet the learner corpus design criteria discussed in Granger (2013)
 - Pertaining to the learner
 - Age
 - Gender: males vs. females
 - Mother tongue background
 - Pertaining to the task
 - Medium: written vs. spoken
 - Topic: heterogeneous
 - Timing: cross-sectional vs. longitudinal
 - From fully natural to semi-natural data
- ▶ Tokenization

Annotation: metadata and error annotation

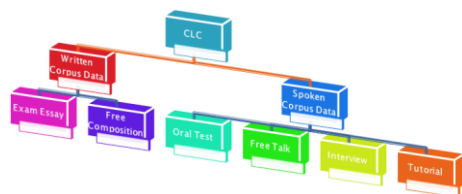
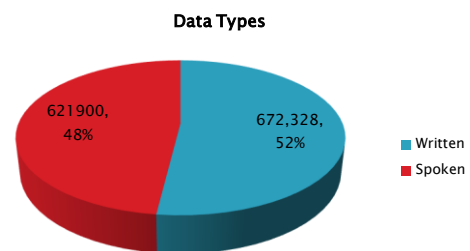
CLC: Key features

- ▶ Over 1.2 million words
- ▶ 886 L2 Chinese speakers from 80 countries
- ▶ Three proficiency levels
- ▶ Fully error tagged
- ▶ XML format with rich metadata

CLC: Written vs. Spoken Data

- ▶ Ratio of written vs. spoken data
 - Original: 7 : 3
 - Adjusted: 5.2 : 4.8
- ▶ Why?
 - More spoken data available
 - Reflecting the predominance of speech in daily communication

CLC: Written vs. Spoken



CLC: Task Types

- ▶ Cover a variety of tasks types
- ▶ Written Data
 - Essays under test conditions vs. Free compositions
 - Descriptive essays
 - Argumentative essays
 - Letters
 - Business writing
- ▶ Spoken Data
 - Utterances in oral tests vs. Free conversations
 - Oral Tests
 - One/two/three test taker(s) vs. an examiner
 - Monologues by an L2 examinee
 - Free Talks (Monologues): "My Hometown" & "A Memorable Trip"
 - Tutorials: NNS vs. NS
 - Interviews: NNS vs. NS

Topics for Written Test Essays

- ▶ a business figure that I admire
- ▶ a custom in your country that is different from China
- ▶ childhood
- ▶ China and me
- ▶ Chinese learning and my personal development
- ▶ describe a person
- ▶ introduce a favourite book or movie
- ▶ love in my heart
- ▶ my hobby
- ▶ my idea of
- ▶ my living
- ▶ my story
- ▶ naming
- ▶ ten years later
- ▶ travel
- ▶ letter

Topics for Written Test Essays (cont'd)

- ▶ picture writing—blame shouldn't go too far
- ▶ writing using keywords
- ▶ retell a text that impresses you most
- ▶ advice
- ▶ interview report
- ▶ continuation writing task
- ▶ business writing
- ▶ my ideal career
- ▶ my wealth view
- ▶ Russians don't believe in love
- ▶ view on Chinese dialects
- ▶ what is the most important thing in your life
- ▶ why do I like this restaurant
- ▶ miscellaneous

Topics for Free Compositions

- ▶ a memorable city
- ▶ a memorable event
- ▶ a memorable trip
- ▶ a movie or book that moves me
- ▶ a story about my Chinese learning
- ▶ a story that impresses me
- ▶ Carol's story
- ▶ China and me
- ▶ country-specific features
- ▶ describe a person
- ▶ GDUPS in my view
- ▶ introduce a party
- ▶ It's too late to
- ▶ Katayama Yumi first arrived in China
- ▶ Laura learning to drive a car
- ▶ my Chinese friend and me
- ▶ my dream
- ▶ my experience of
- ▶ my favourite thing
- ▶ my hobby
- ▶ my hometown
- ▶ my living
- ▶ my monthly expenses
- ▶ plan after graduation
- ▶ room cleaning
- ▶ the place where I live in Guangzhou

Topics for Free Compositions (cont'd)

- ▶ letter
- ▶ shorten a text
- ▶ continuation writing task
- ▶ advice
- ▶ announcement
- ▶ art review
- ▶ film review
- ▶ how to keep fit
- ▶ if the world came to an end
- ▶ my ideal career
- ▶ my writing in Chinese
- ▶ the benefits of having a boyfriend or girlfriend
- ▶ what is the most important thing in your life
- ▶ your view of euthanasia
- ▶ business letter
- ▶ miscellaneous

<初级>
<乌兹别克斯坦>
<乌兹别克语>
<男>
<考试作文>
<score: 80>

Sample writing
(descriptive)

我的爱好

每个人<oms>都</oms>有爱好.有的人爱<orth>看</orth>电影,有的人爱散步,我的爱好是<orth>看</orth>电影.
<orth>看</orth>书和书法,书法很难,但是有意思,周末<orth>我</orth>一般练习<orth>书</orth>法.我也喜欢<orth>看电影</orth>影.我一<orth>般</orth>看英语电影.<orth>我</orth>要看汉语<orth>电</orth>影,但是我<oms>看</oms>不明白.他们说得很很快.

<高级>
<芬兰>
<芬兰语>
<男>
<考试作文>
<score: 95>

Sample writing
(continuation writing task)

<prompt>她是我们的女儿吗?
结婚两年以后,西美和丈夫有了第一个孩子,他们非常高兴.
现在,西美又抱起了孩子.她摸着孩子的头,孩子甜甜地对她微笑.看着孩子可爱的样子,西美也高兴地笑了.可是看了一会儿,西美也开始伤心起来——“尽管你很可爱,我和你爸爸也非常爱你,可你真的是我们的孩子吗?”
孩子出生以后,从医院回到家中,亲戚朋友见了都很吃惊,因为她长得既不像西美又不像西美的丈夫.西美和丈夫皮肤都很白,头发都是金色的卷发.夫妻俩都有一双大大的蓝眼睛,而且两个人都是双眼皮.可这孩子是咖啡色的皮肤,长着一头黑色的直发,一双小小的黑眼睛非常明亮,两只眼睛全是单眼皮.
孩子九个月左右的时候,西美和丈夫又去了那家医院,“这是我们的孩子吗?”西美问,“她当然是你们的孩子.在我们医院,每个孩子出生以后,都有一个号码,你的孩子是6号,所以肯定错不了.”医生解释说.
西美和丈夫又去找那家医院的护士了解情况.一个护士告诉他们,当时有另一个孩子也和父母长得都不一样,父母皮肤颜色很深,可孩子皮肤颜色很浅;父母都是黑色的直发,但孩子是金色的卷发;父母是黑色的小眼睛,可孩子是蓝色的大眼睛.在睡觉,那个护士还告诉了他们对夫妻的地址.
西美带着孩子找到了那对夫妻的家.——</prompt>

来开门的妇女看起来像西美抱的孩子<punct>.</punct>一模一样.她一看到西美<punct>.</punct>就明白了是怎么回事.<ind>而</ind>高兴她叫了:“我的孩子!”西美把孩子交给她<ind>了</ind>,但是始料未及,她心里还是有点不舒服.妇女<ind>也</ind>把金发的孩子交给西美时,孩子哭起来了.
回到家了,孩子还在哭,而哭得<enf>死去活来</enf>.这样,一连几天,孩子一直在哭,哭得让西美和丈夫也受不了.这几天,<meaning>西美还日日夜夜担心是不是那个咖啡色的孩子没归家</meaning>.
过了一个星期,西美和丈夫明白了,原来那个咖啡色的孩子跟他们<oms>在</oms>一起已经差不多一年了,他也不知道他们的孩子了.西美和丈夫就带着金发的孩子再次往那<ind>另</ind>一对夫妇家去<ind>了</ind>.西美一定要彻底解决这个问题.

<初级>
<西班牙>
<西班牙语>
<男>
<考试作文>
<score: 71>

Sample writing (letter)

<prompt>你有一个中国朋友在中国的其他城市（北京或者上海等），请你给他（她）写封信，告诉他（她）你在广州的情况，可以介绍你的学校，同学，广州的天气，食品等，也可以介绍你去旅行的情况。</prompt>

<name><punct>: </punct>
你好吗，你的中文怎么样？<meaning>从十月我们没见面
</meaning> 所以我写这封信。
我在广州一边<erff>工做</erff>一边在广外大学学习中文。我们每天学习三个小时，有语法，口语，听力还有YUE DU课。我在<erff>初极</erff>四班。我的同学们都<oms>说</oms>中文说得很厉害。我的中文也进BU了，你呢？在北京你的课怎么样？
<meaning>广州的风si不太美丽但是还可以</meaning>。有花园，瓜园，动物园，BO物GUAN什么的。有的地方好看，有的地方有点儿zang。
这里的天气比JIAO RE,我不XI GUAN。现在雨下得<oms>很
</oms>厉害。
广州的小吃很好吃，有CHOU豆腐，有水果，有<meaning>肉寺寺
</meaning>。

<高级>
<伊朗>
<波斯语>
<男>
<自由作文>
<score: NA>

Sample business writing

主题： 询价<erff>即</erff> 产品<orth>网</orth>络 代理
尊敬的 A 女士：
您好！
在<orth>网</orth>站上看到贵公司的信息<punct>，</punct>即贵公司生产<oms>的</oms>各种USB小产品，比如USB暖手套、USB小电扇、USB小冰箱。。。本人对贵公司的产品十分感兴趣<punct>，</punct>并看好贵公司的信誉。因此想询问一下贵公司的价格，包装，质量，付款方式<erff>即</erff>交货<erff>日期</erff>。希望能与贵公司合作并得到双赢。
我的联系方式如下：
EM 邮件
电话
2014年 9月 11日
张

<中级>
<韩国>
<韩语>
<女>
<口语考试>
<score: 90>

Oral Test: S vs. E

S: 我是中三班<name>。
E: 嗯。
S: <reading>。
E: 好，读到这里，不错。下面呢是第二题，请你抽一张纸，第二题。
S: <l pause>第一。
E: 啊，第一，好，请你说一说在中国遇到的新奇的事情，吧。
S: 对我来说，刚来广外的时候，我很受煎，因为在校园的，在校园的楼<laugh>大多是宿舍。
E: 嗯。
S: 还有我看见了很好看的衣服挂在房间的外面，可是<laugh>我们国家不是这样。拿我们学校来说，嗯，在校园的楼大半儿是教学楼，嗯，大概三所楼就是宿舍。
E: 嗯。
S: 并且，并且在我们国家，我们国家的房子都有阳台，所以洗完以后，把洗好的衣服，在那儿的晒台。完了。
E: 嗯，好，这个是你遇到的新奇的事，对吗？
S: 对。
E: 好，说得不错。今天考试结束。呃，你帮我问一下<name>，韩国的同学，谢谢。
S: 谢谢老师。
E: 嗯。

<S1 初级;S2 初级>
<S1 墨西哥;S2 柬埔寨>
<S1 西班牙语;S2 柬埔寨语>
<S1 女;S2 男>
<S1 口语考试;S2 口语考试>
<S1 96;S1 95>

Oral Test: S1, S2, (S3) vs. E

E: 呃，第几个？一、二、三、四、五、六。
S1: 第<s_pause>，第四个。
E: 谈论考试。
S1: 嗯。
E: OK. 好。
S2: 嗨，<name>，最近怎么样？
S1: 我？不错啊？你呢？
S2: 听说你要考，你快要考HSK了，是吧？
S1: 是啊，咳，别提啦。我一听到考HSK的事我就头疼。
S2: 头什么疼啊？你不是说很容易吗？
S1: 容易是容易，不过，没那么多时间了。期末考试到了，怎么办？
S2: 糟糕！期末考试到了，我还没复习呢！
S1: 没关系啦，你不是说你的手机什么都可以叫<laugh>？
S2: 是啊，<laugh>考试的时候用手机比较方便嘛。
S1: 我知道，可是，靠手机不如靠自己的努力。
S2: 没问题啊，我也可以靠自己的努力。
S1: 如果我们想考好的话，那就一起加油吧！
S2: 嗯，来嘛，我懒得要命。
S1: <laugh>可以慢慢来了，一定要努力。
S2: 好了，我们俩都去试试<laugh>。
E: <laugh>挺好的，准备时间还挺自然的，准备了好长时间？

<中级>
<泰国>
<泰语>
<女>
<口语考试>
<score: 92>

Oral Test: Monologue

S: 啊，我叫<name>。呃，读句子，<reading>。第二，解，解释“习惯成自然”，就是，每天做，做得很长时间就习惯了。第三，我要介绍那个，如果我的朋友来广州的话，我要带他去，呃，五羊石像，因为五羊石像是广州的标志，呃，从<unclear>变成了那个<clear_throat>石头。呃，在环境的，在环境优美的<clear_throat>在环境优美的公园里，五羊石像，呃，五颜六色，栩栩如生，很漂亮，<clear_throat>我觉得我的朋友应该去。然后我要带他去白云山。白云山是广州的，呃，风景胜地，有丰富自然景观<lex3>。然后如果他想要吃什么，我要带他去吃肠粉<clear_throat>，因为肠粉是广东的，呃，一种特色菜，呃，可以放上多种多样的酱料，比如说，呃，呃，比如，鱼皮，呃，猪肉，牛肉，什么的。点菜很容易，呃，等了不久也可以吃了。如果我的朋友来广州的话，我要，这些事我要介绍给他。

<高级>
<芬兰>
<芬兰语>
<男>
<口语课堂练习>
<score: >

Free Talks: Monologue
My Hometown

S: 我的家乡叫<foreign>，呃，中文名字好像叫库奥皮奥<laugh>。呃，它位于芬兰东部，呃，在<s_pause>一个叫<foreign>的省。呃，那个地区也叫芬兰的湖泊区，因为那边<s_pause>湖比<sigh>陆多，就是湖泊比陆地多。呃，所以<foreign>也位于，呃，一个大湖泊中间的半岛上，所以，四周都有湖。呃，它是十六世纪初由瑞典国王建立的，呃，目的就是要把那个，当时的瑞典的东部稳定下来。呃，但是，因为每次瑞典和俄罗斯之间发生战争，它就会引向到这个地区，所以一直到今天那边的人比较少。呃，<foreign>的，呃，人口只有大概十一万。啊，虽然<s_pause><foreign>这座城市比较小，但是那里也住上了比较多芬兰的名人，比如，几个，呃，比如几个芬兰的总统。呃，除了，除了<laugh>总统以外好像<s_pause>没有什么太<s_pause>有名的人物，一些歌手，一些演员这样的。呃，还有一位，呃，六十年代七十年代的喜剧演员，呃，他叫<foreign>，他在芬兰非常<s_pause>有名。他已经去世了，但是我父母那一代特别喜欢他。呃，<foreign>的风俗，<sigh>，好像，我也不太认识，因为<foreign>已经被那种芬兰的主流文化，就是那种芬兰的，芬兰的普通化<s_pause>文化。呃，<l_pause>占领了。呃，其实本地风俗已经差不多没了，可以说芬兰东部也包括<foreign>的风俗，呃，只是一种生活习惯或者生活方式，就是，呃，大家<s_pause>，传统来说大家都会在市中心的<s_pause>一个最大的市场集合，然后那里很活泼。还有<foreign>人不太喜欢工作<laugh>，不太喜欢劳动，有点懒。所以，呃，可以说<foreign>人很悠闲。呃，当地的美食，呃，大部分都是海鲜。但是实；实际上不是海的，因为是从湖泊得到的<s_pause>鱼和贝类这种东西。

<中级>
<斯洛伐克>
<斯洛伐克语>
<女>
<口语课堂练习>
<score:>

Free Talks: Monologue
A Memorable Trip

S: 我很喜欢旅行，在欧洲已经去过一些地方，比如说，奥地利，捷克，匈牙利，意大利，希腊，英国什么的。最近我的很大的梦想去中国旅行，变成了现实。呃，两个星期以前，我有机会去去有些很有意思的地方，是因为我的斯洛伐克的朋友们来中国参加一个武术的比赛。呃，这个比赛结束了以后，我，我们决定一起去旅行。先去了华山，又去，武当山，然后去北京。我最喜欢，呃，华山和武当山的风景。呃，它的高山。虽然，呃，我们斯洛伐克也有这样的高山，可是并没有，并没有寺庙，所以对我来说，这些地方，都很有意思。呃，我们想去，呃，看一看这些寺庙的时候，呃，先要，上很长的楼梯。呃，这个让我，很累，可是，一定很值得上去。呃，北京，北京很大。我听说，呃，那儿什么，什么都能生活，可以找到自己的房子。我来北京的时候发现，就是这样的，那儿每个地方都有来历。有很多地方我们要参加，可是没有时间，所以，我们去的地方只是故宫、长城、天安门。还有，因为在北京我斯洛伐克同学学习，所以我跟她见面了以后，一起去很有名的老舍茶馆，那儿，茶很贵<laugh>。可是因为我们太喜欢这个地方，决定玩得开心。呃，因为，我现在很想念我的家乡，我的朋友们，也因为这些地方都很有意思，很漂亮，所以，这次旅行对我来说很难忘。还有，我希望未来再有机会去武当山，因为在那，呃，安静的，呃，什么问题让我忘记。在那我觉得很开心，呃，也有机会，呃，学习武术，我对武术很感兴趣，所以，一定要回去。

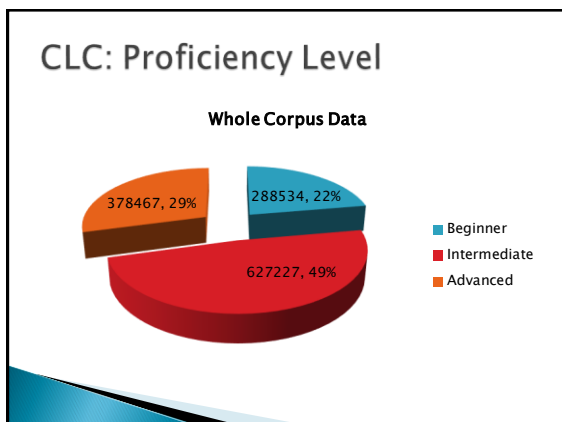
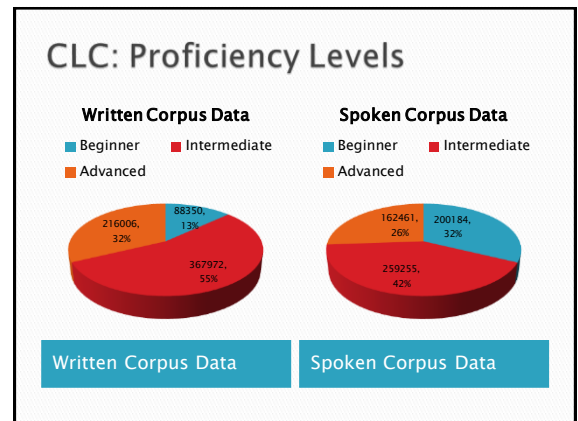
<中级>
<美国>
<英语>
<男>
<口语辅导>
<score:>

Free Conversations: 1 NS vs. 1 NNS

NNS: 像格-云顿。
NS: 对，对，对，向格-云顿。
NNS: Yeah...
NS: 那个广告，非常的有名。
NNS: 但是我从小时候看他的.....
NS: 他，对，他也是.....
NNS: 对，对，因为我，我很喜欢他，他是那个.....
NS: 他也是那个动作，action star。
NNS: 对，对，对。
NS: 对吗？
NNS: 对对对。就是因为他真的是我的偶像。
NS: 噢，是吗？
NNS: 然后我，我觉得<_pause>我很希望可以碰到他。
NS: 哦！就突然碰到他了。
NNS: 然后就，呃！他跟他的家人一起，就是没有很多人拍照，就是midnight十二点钟。
NS: 哦。
NNS: 没有没有人在，然后我注意看，他就<_pause>呃，使劲儿，right?
NS: <laugh>。
NNS: Oh my god! 这是因为，其他明星我都不管，就是哦，哦，这是明星，但是他，他就是.....
NS: 对，对，对，因为他就是，因为看他的原因，就是我可以那个动作还可以学跆拳道，什么，我来中国就是因为他。
NNS: 哦，哦，你就是因为他.....

CLC: Proficiency Level

- Original Plan
 - Beginners vs. Intermediate vs. Advanced
 - 3 : 4 : 3
- Adjusted
 - Written: 2 : 5 : 3
 - Spoken: 3 : 5 : 2
- Reasons
 - The written output by beginners is relatively small
 - Beginners can produce more spoken data than written data
 - Very few learners can reach the advanced level
 - In addition, the syllabus for the advanced learner is focused more on writing than on speaking
- Currently
 - Written: 13.14% : 54.73% : 32.13%
 - Spoken: 32.19% : 41.69% : 26.12%



CLC: Country Background

- Corpus data were produced by L2 learners from altogether **80** countries
 - Written subcorpus: from 76 countries
 - Spoken subcorpus: from 56 countries
- L2 learners from top 10 countries
 - Korea, Thailand, Russia, Kazakhstan, Vietnam, Japan, Venezuela, Uzbekistan, Ghana, Brazil
- Data not skewed towards East/Southeast Asian L2 learners

CLC: Gender

- ▶ Goal
 - Males vs. Females
 - 5 : 5
- ▶ Currently: 43.55% : 56.45%
 - Written: 33.57% : 66.43%
 - Spoken: 54.35% : 45.65%

Longitudinal Corpus Data

- ▶ The corpus contains over 200 learners' longitudinal corpus data (between 2012 and 2015), which we can "track the same learners over a particular period [i.e. at different proficiency levels]" (Granger 2013).

e.g.

- S_F_ANA_RU 初一-D 中一-B 中三B(2)
- S_M_CJR_TH 初二-A 中二(2)
- S_F_QJY_KR 中二 高级

Metadata (Headers) in Text

<中级>	Proficiency level
<美国>	Country background
<英语>	L1
<男>	Gender
<考试作文>	Task type
<score: 89>	Score

Transcription Conventions

- ▶ In the process of transcription and digitalization, we have retained some transcription conventions, so that some information (e.g. some features of L2 speakers' writing and speaking) will not be missing.

Transcription Conventions in Spoken Corpus Data (1)

- ▶ Speakers ID
 - E: OK. 好。
 - S2: 嗨 · <name>, 最近怎么样?
 - S1: 我? 不错啊? 你呢?
 - NNS: 然后我, 我觉得<l_pause>我很希望可以碰到他。
 - NS: 哦! 就竟然碰到他了。

Transcription Conventions in Spoken Corpus Data (2)

- ▶ Fillers
 - 呃 (23,083), 嗯 (24,820), 啊(6,087), 哦 (2,718), 噢 (1,791)
 - 我想告诉你, 刚, <s_pause>呃, <l_pause><laugh>呃, <s_pause>结, 呃, 呃, 婚礼的以前, 呃, 新郎, 呃, 要去他的爱人的家, 在那儿, 呃, 新娘的朋友, 呃, 交朋友嘛, 呃, 见面新郎。
- ▶ Pauses
 - <s_pause> pauses <3 seconds
 - <l_pause> pauses >3 seconds
- ▶ Meta-linguistic behaviour and comments
 - <clear_throat> <sniffle> <laugh> <sigh> <sneeze> <whisper>

- ▶ <reading>
 - E: 到这句，到这句。嗯，开始。
 - S: <reading>
 - E: 嗯，好，可以，行，读得不错。
- ▶ <foreign>: code switching
 - 呃，我的家乡，呃，叫<foreign>。呃，这个名字是新西兰的语言的，呃，毛利语的。
- ▶ Anonymisation
 - <name>
 - <number> (e.g. phone number)
- ▶ <unclear>

- Typo
 - <orth> a Chinese character written with an additional or missing stroke
-
- 她会拿出十<orth>块</orth>钱给人家·还会买
- <foreign>
 - Anonymisation
 - <name>
 - <number> (e.g. phone number)
 - <email>
 - <unclear>

- ▶ Surface Strategy Taxonomy
 - Omission
 - Incorrect inclusion
 - Incorrect form
 - Wrong order
- ▶ Error types based on linguistic categories
 - Orthography/Pronunciation
 - Punctuation
 - Lexis
 - Grammar
 - Collocation

- (Dulay *et al.* 1982; Ellis & Barkhuizen 2015)

Type	Level	Orthography	Punctuation	Lexico-Grammar	Collocation
Omission			<punct> [pronunciation] —marked only for omission ●星期一是和星期三有马和星期五的差别吗？<punct> </punct>	<omts> [character omission] ●中国是<oms>-制。<oms>魅力的国家。 ●在中国有很多好吃<oms>。 </omts> 完！ 一千八百<oms>-前<oms>了7考试	
Incorrect inclusion			<incl> [incorrect inclusion of a lexicogrammatical form] ●如果，还有三哥、你骑马，马<incl>> -></incl>-骑啊！ ●比较<incl>-错</incl>-多，不过又便宜。 ●我去年所写的词曲总结<incl>-?</incl>-第二十年。 ●我们从中学开始就认识O<incl>-个</incl>-年都是好朋友。		
Incorrect form		<orth> [orthography] - understood as incorrect strokes used ●真弄糊涂了一顿 ●章中<ortho>-</ortho>-后 		<erff> [erroneous form] ●会讲中文-erff-说- </erf>- ●有些奇怪的东西：特别黑<erff>-</erff>- ●我-erff-很厉害，能看-音乐书-然后上了大学之后， ●我-erff-很认真；<erff>-比如热爱-	<coll> [collocation error] ●你可以<coll>-充沛地-去游- </coll>-中国的自然美。 ●这样的搭配只有the col+Lause=爱7。
Wrong order			<wro> [wrong order] ●呢、都挺、挺、挺、<wro>-排列在 ●</wro>-的顺序上 ●对呀，那咱俩没话说了，了解您造化了嘛跟人家（wro）</wro>- ●他-<wro>-我们- </wro>-来着！		

- ▶ **<meaning>**给别人自己很漂亮 **</meaning>**
- ▶ **<meaning>**不是对手你，不是对手你已经认识他 **</meaning>**
- ▶ E: 嗯，嗯。好，那么你觉得我们照顾老人的时候最重要的是什么？
- ▶ S: 呃，嗯，我觉得重要，呃，**<meaning>**可以吃过，嗯，没，没吃了，嗯，做过，做过，做过的药 **</meaning>**。

- ▶ 呃，在那儿找到一个俄罗斯的姑娘，结婚她。
- ▶ 我真的很高兴学习在中国。
- ▶ 你包子比我好，包我饺子，你包饺子比我好
- ▶ 我请他降价一点。
- ▶ 他们，他们公司努力。
- ▶ 可是，比较年轻人，他们觉得那，嗯，做泡菜很难，麻烦，所以他们不做。
- ▶ 我们国家的人口一共有八百万万口人。
- ▶ 我的名字是<name>，我来自韩国的。
- ▶ E: 请问，学校附近有餐馆吗？S: 没有。

Interlanguage Errors (2)

- ▶ 以后，我呢，越来越发展，发展以后，我觉得踢足球也越来越好。
- ▶ 故宫非常漂亮，非常灿烂。
- ▶ 我，我们国家如果随便把，把垃圾扔在外面的话，呃，严格的罚款。
- ▶ 你可以享受韩国的古代史。
- ▶ 这次的旅游之后我们的关系<l_pause>变，变熟了。
- ▶ 我的家乡比较文化。
- ▶ 教育，呃，有，有血腥，血腥的折磨
- ▶ 有很，呃，承认的，承认的态度，sincere attitude
- ▶ 冬天的时候很冷。风景，很衣。

Interlanguage Errors (3)

- ▶ 如果，还有三步，你过马，马上路吗？
- ▶ 喷，呃，好吃的东西，呃，呃，都是，鲜，呃，鲜新的蔬菜。
- ▶ 比较很多，不过，又便宜。
- ▶ 我妈妈是一个有条有理的人。他是一个上的天堂，下的厨房的人。

Next Stages...

- ▶ Error tagging
- ▶ Data analysis
- ▶ Corpus depositing and releasing
- ▶ Developing an online query system

References

- ▶ Dulay, H., Burt, M. & Krashen, S. (1982). *Language Two*. Oxford: OUP.
- ▶ Ellis, R. & Barkhuizen, G. P. (2005). *Analysing Learner Language*. Oxford: OUP.
- ▶ Granger, S. (2013). "Learner corpora" In C.A. Chapelle *The Encyclopedia of Applied Linguistics*. London: Blackwell.
- ▶ Pravec, N.A. (2002). "Survey of learner corpora". *ICAME Journal* 26: 81–114.

