What can language learners do

In the context of machine translation

May LI
Tongji University

20200903

内容提要

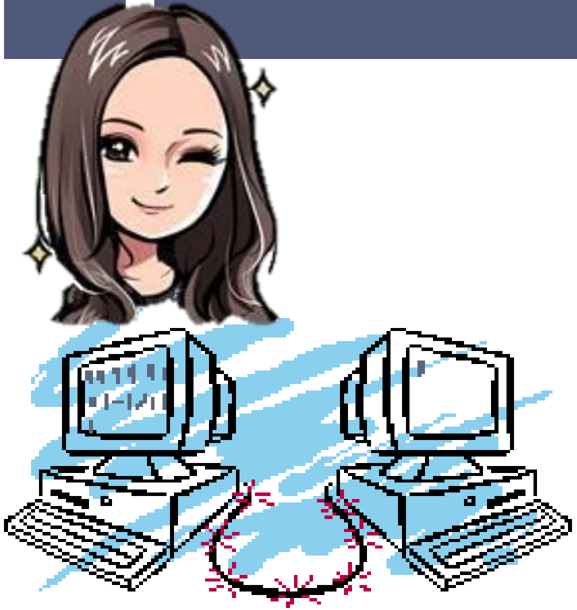
1 MT development

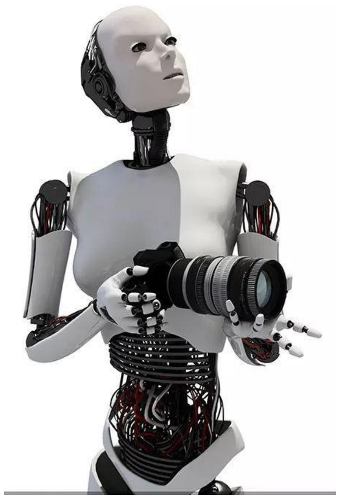
2 PE involvement

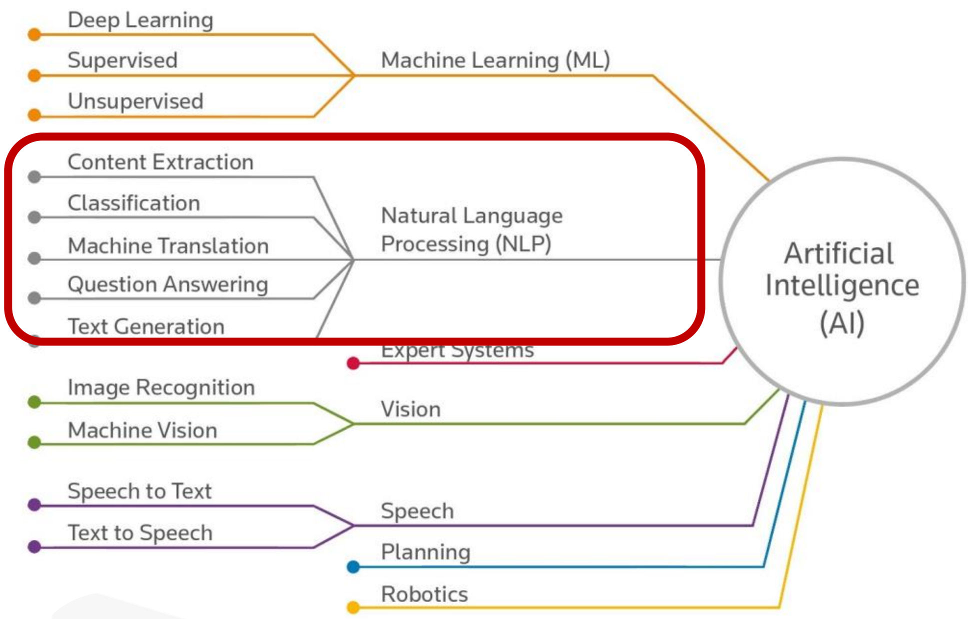
3 MTPE research

MT developmnet

1

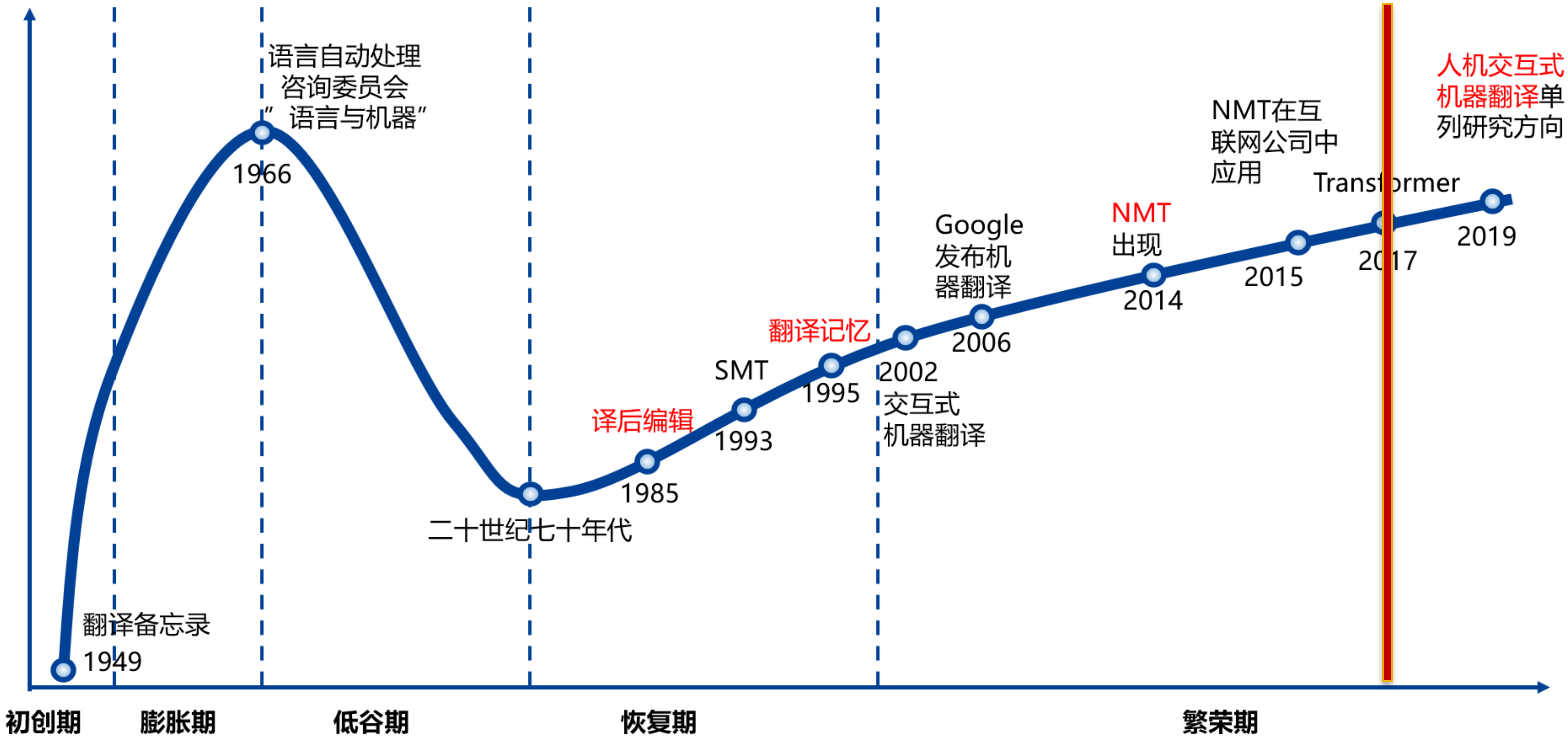




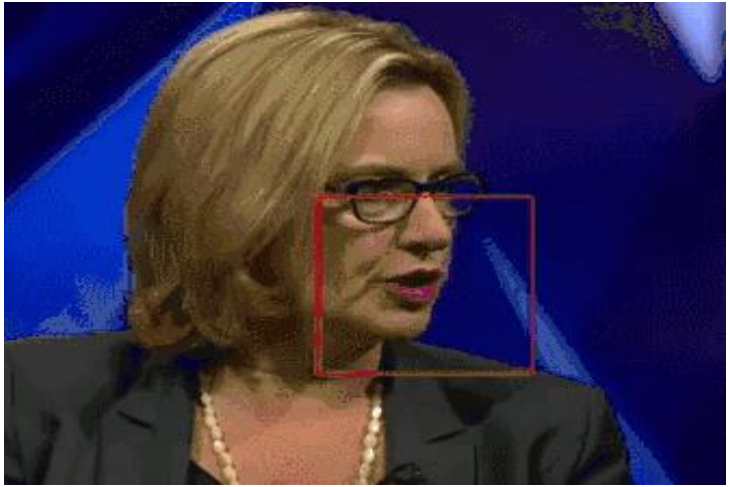


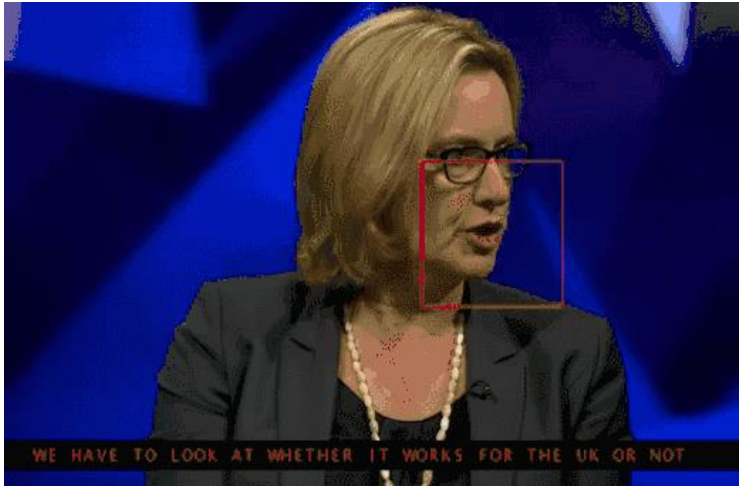
qq音乐如何听歌识曲

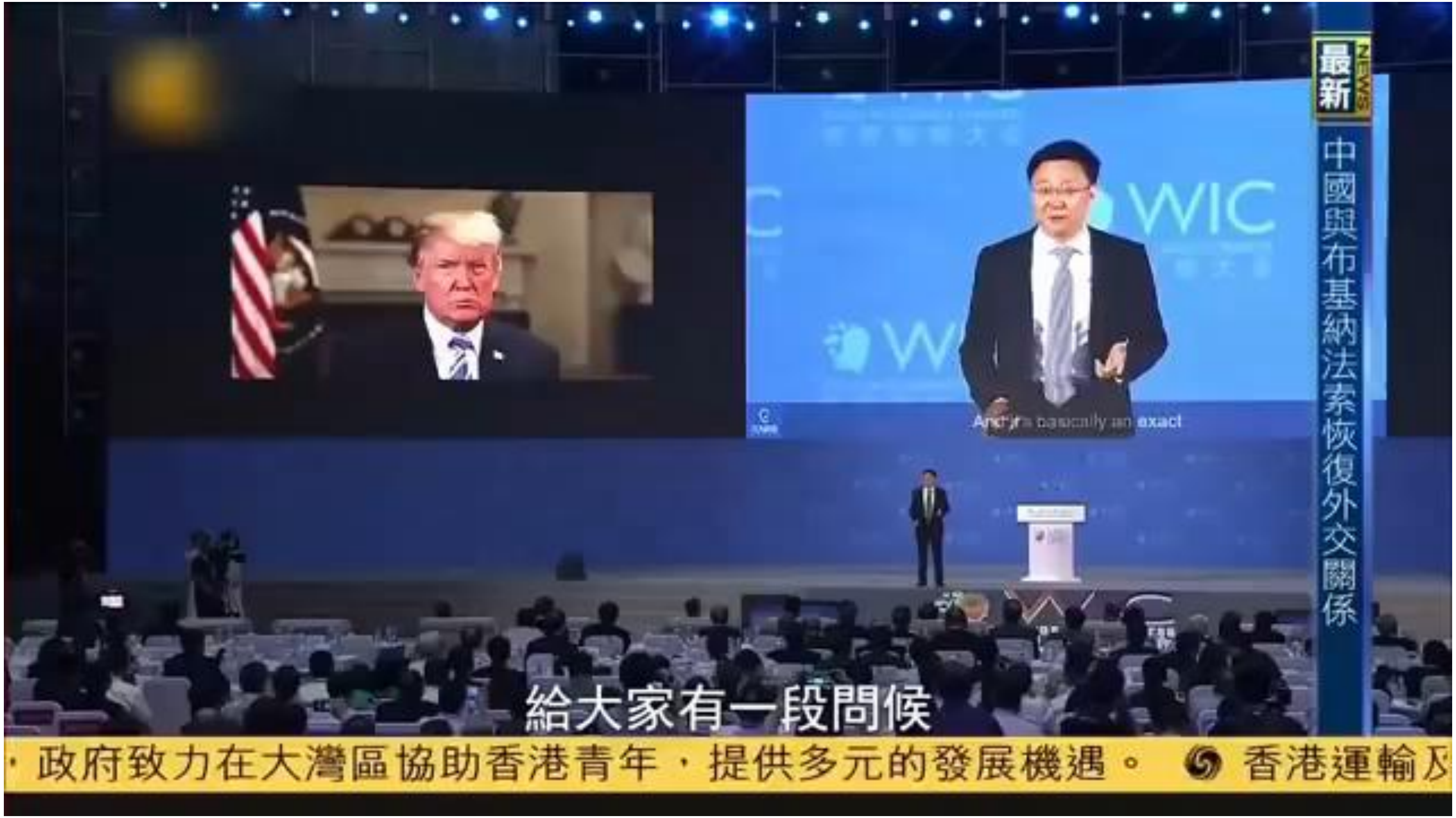
Source：Andy McDonald，tcworld 2019, Stuttgart



来源：黄国平，腾讯Transmart 交互技术与功能特性 20200603




WE HAVE TO LOOK AT WHETHER IT WORKS FOR THE UK OR NOT



最新 NEWS

中國與布基納法索恢復外交關係

給大家有一段問候

·政府致力在大灣區協助香港青年，提供多元的發展機遇。 香港運輸及

20160928 Google: NMT

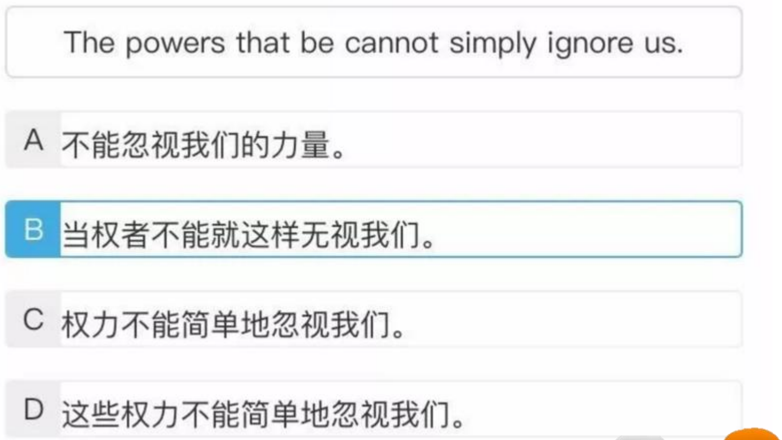
NMT

原文（标题）	How Much Risk Is Being Added by Insurers' Banking Acquisitions?
译文 （GT1-2016.03）	多大的风险正在由保险公司的银行收购增加？
译文 （GT2-2016.10）	保险公司的银行收购增加了多少风险？
人工翻译	保险商收购银行（股权）会增加多少风险？

引自微信公众号“译言千金”

人机翻译大PK

人工翻译VS机器翻译，鹿死谁手，你说了算



The powers that be cannot simply ignore us.

A 不能忽视我们的力量。

B 当权者不能就这样无视我们。

C 权力不能简单地忽视我们。

D 这些权力不能简单地忽视我们。

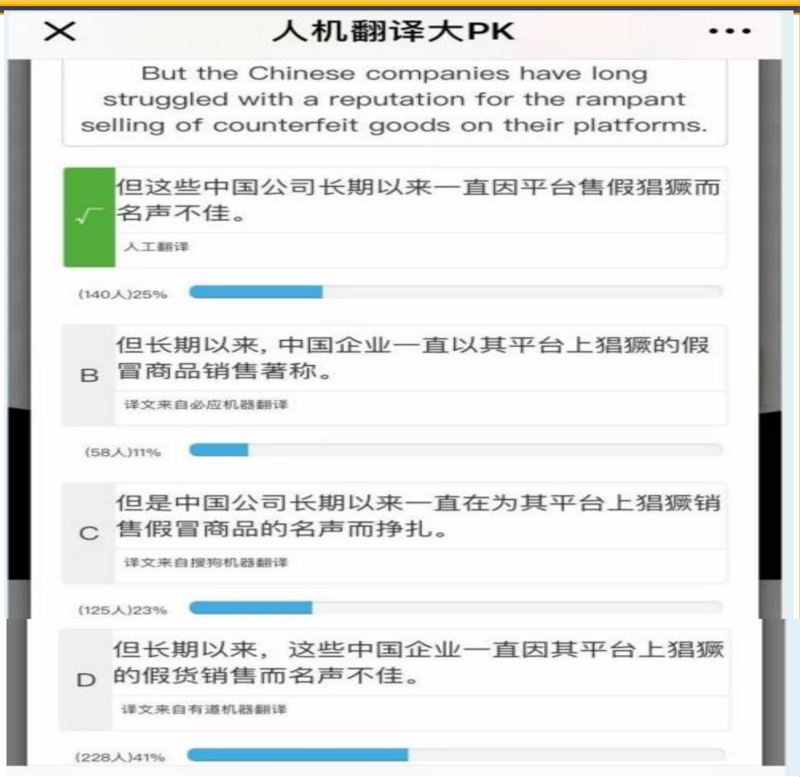
正确率：53%

A：25%

B：11%

C：23%

D：41%



人机翻译大PK

But the Chinese companies have long struggled with a reputation for the rampant selling of counterfeit goods on their platforms.

√ 但这些中国公司长期以来一直因平台售假猖獗而名声不佳。

人工翻译

(140人)25%

B 但长期以来，中国企业一直以其平台上猖獗的假冒商品销售著称。

译文来自必应机器翻译

(58人)11%

C 但是中国公司长期以来一直在为其平台上猖獗销售假冒商品的名声而挣扎。

译文来自搜狗机器翻译

(125人)23%

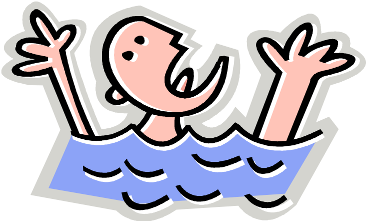
D 但长期以来，这些中国企业一直因其平台上猖獗的假货销售而名声不佳。

译文来自有道机器翻译

(228人)41%

试译宝 20181127

MT can speak like humans



Way out?

- 90% translation replaced by MT;
- MT: semi-products.

How to improve productivity with MT?

2

PE involvement

- **quality of final product**
- **quality requirements**
- **levels of post-editing**

M
T
P
E

GOALS OF MT

1. To understand →

- MT without PE
- Light PE

GOALS OF MT

2. To communicate →

- Minimal PE:
- Full PE:

	Light Post-editing	Full Post-editing
Grammar	minor issues are acceptable	must be correct
Punctuation	variations/errors are acceptable	must be correct and consistent
Spelling	minor issues and/or variations are acceptable	consistent including hyphenation
Terminology	understandable and usable	accurate and consistent
Style and tone	not offensive	appropriate for content
Style	variations are acceptable	Consistent: headers, list items
Formatting	not important	consistent with the source text

Productivity

Productivity of a post-editor:

- qualification to which extend a post-editor is productive in transforming a raw output into a text that fulfills its intended function;

- measured by looking at the speed of a post-editor by taking into account his accuracy and proficiency.

Productivity Testing

- provides information on the difference in speed between MT post-editing and translation from scratch;

- an indirect way of evaluating translations;

- scores not assigned directly to translated segments;

- but effort measured in terms of time and edit-distance.

https://www.taus.net/knowledgebase/index.php?title=Productivity

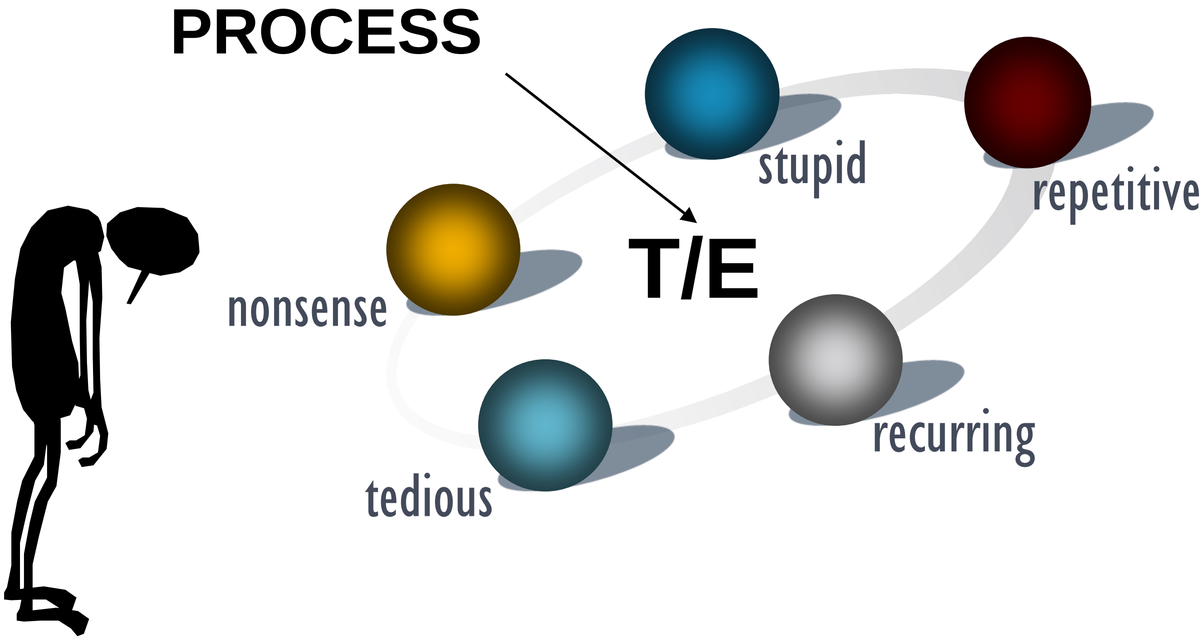
3

MTPE research

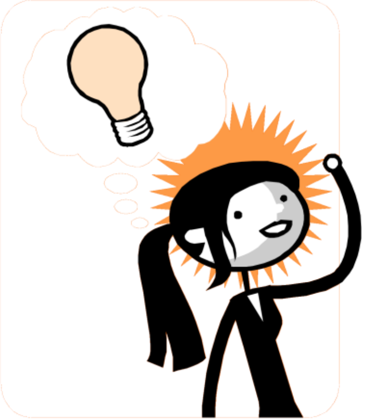
PE process

Car

User manuals



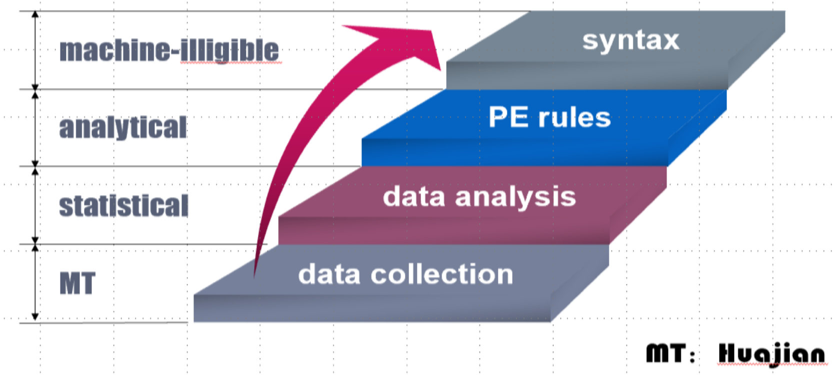
MOE project: MTPE



- **Goal**
  - PE automation
- **Design**
  - MT error patterns

- **PE automation**

MOE funded project （2007）



李梅 2013：英汉机译错误分类及其数据统计分析

Taxonomy

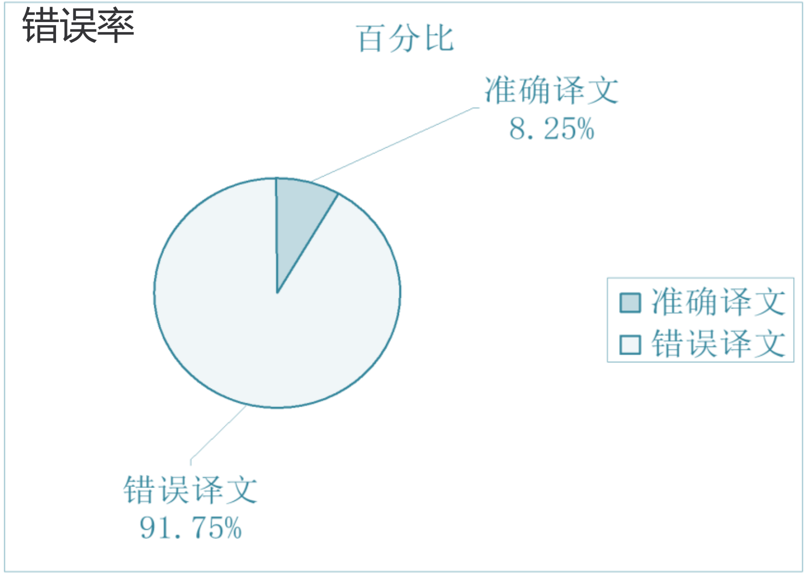
Lexical	Syntactic	Typographical + ZT	
1) Terminology	1) order	1) Symbol	1) Identical
2) POS	2) NP	2) Punctuation	2) Almost
3) Abbr	3) VP	3) Bracket	3) Zero T
4) Verbs …	4) PP	4) Unit	
5) Missed	5) Passive	5) Number	
6) Overdone	6) Infinitive		
	7) Gerund		

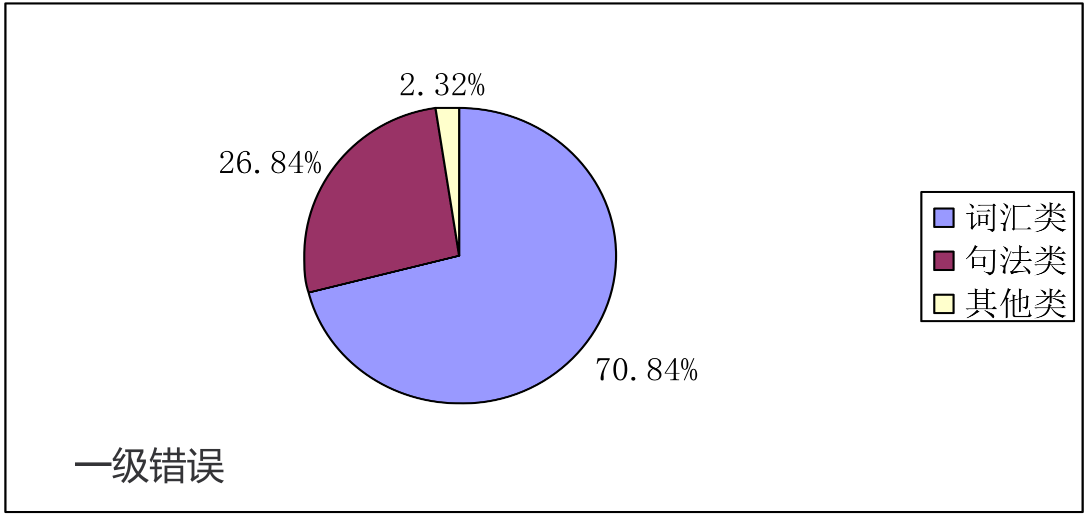
李梅 2013：英汉机译错误分类及其数据统计分析

Research data

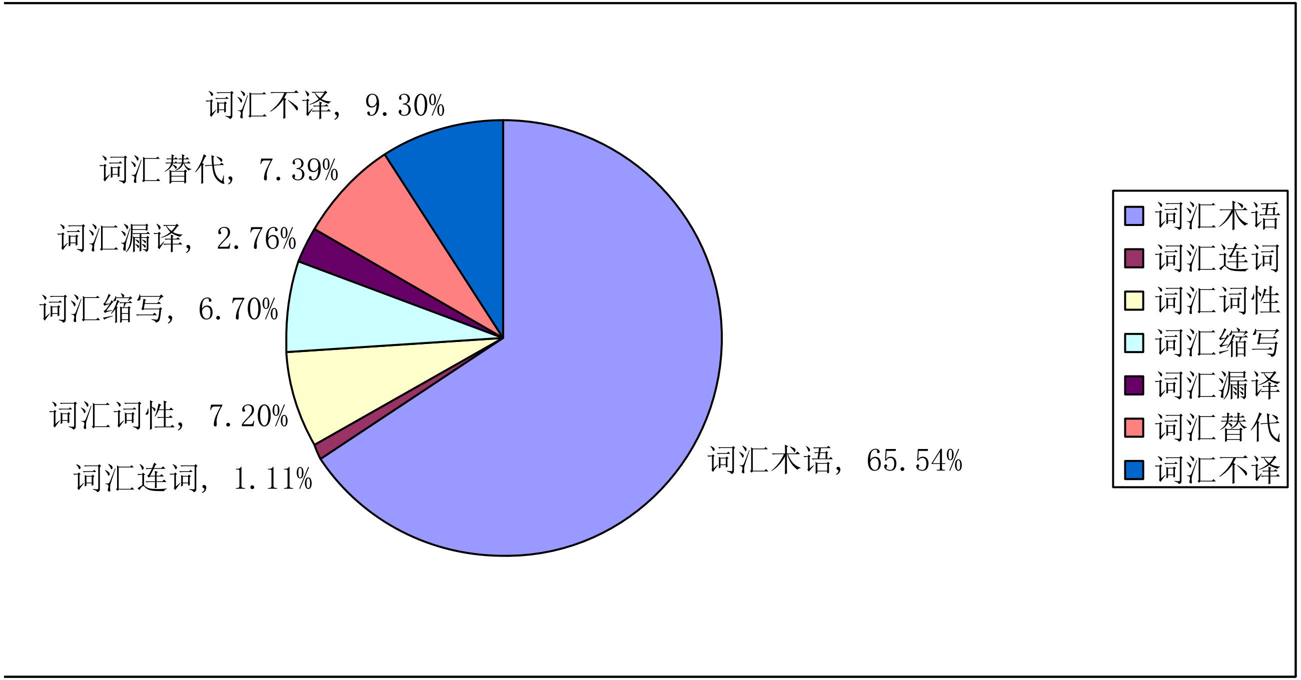
No.	Taxonomy	Number	Percentage
1	terminology	67428	46.43%
2	conjunctions	1138	0.78%
3	POS	7410	5.10%
4	Abbreviations	6892	4.75%
5	Missed	2844	1.96%
6	Replacement	7600	5.23%
7	No translation	9569	6.59%
8	Word order	13975	9.62%
9	Noun Phrases	3618	2.49%
10	Verb Phrases	10470	7.21%
11	Prep Phrases	5888	4.05%
12	Passives	3836	2.64%
13	Infinitives	317	0.22%
14	Gerunds	880	0.61%
15	Miscellaneous	3366	2.32%
total		145,231	100%

MTPE research results





Lexical errors



examples

lexical

•Terminology

-mute *哑的; 静噪
-circuit *巡回; 电路
-ground body
*身体地；车身搭铁
- Intelligent *聪明的；智能

• Preposition

-for *以来; 适用于
-with *和；带有

•Parts of Speech

- cover * 投保; 后盖
- rear header *养育；后

syntactic

Past participle

-in the direction indicated by arrow A
- *在箭A 表明的指示里;
如箭头A所示方向

Word order

-This is the display signal circuit from the multi-display controller to the TV display.
*这显示信号巡回从多显示器控制器到电视展示。

/这是自多功能显示屏控制器至电视显示屏的显示屏信号电路。

typographic

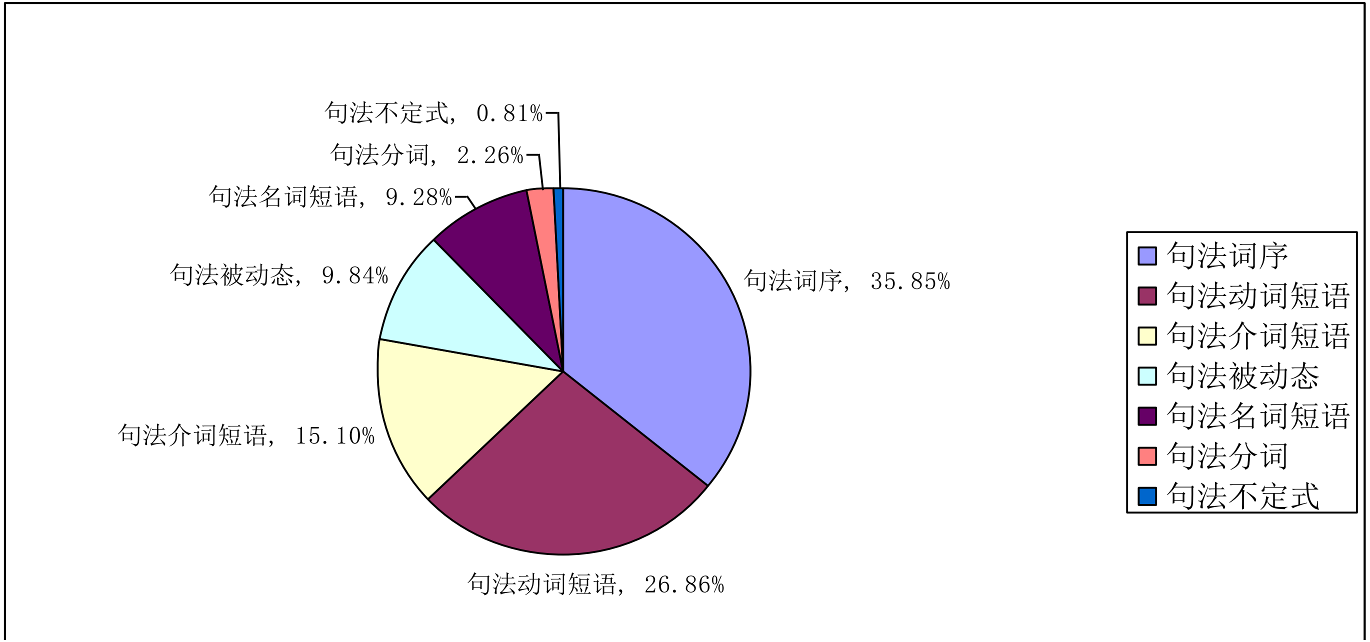
Brackets

-Remove the rear seat for Ottoman seat (see page 1024)

*为长椅位子(参阅页除去靠后的座位1024)。

/拆下带 Ottoman 座椅的后排座椅（参见1024 页）。

Syntactic errors



example

8)a Then the user tries to move away from the vehicle.

M: 然后用户努力从车辆移开。
H: 然后用户试图离开车辆。

subject control

8)b Then the user$_i$ tries PRO$_i$ to move away from the vehicle.

9)a Allow the engine to idle and check DTC.

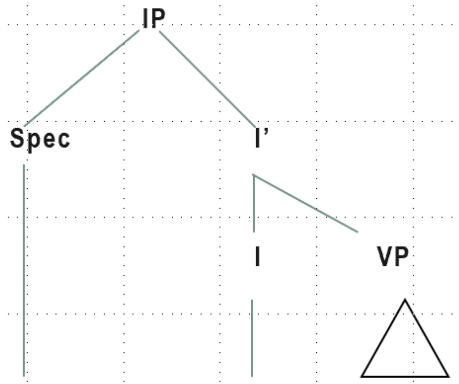
M: 对闲置允许发动机并且检查DTC。
H: 让发动机怠速运转并检查DTC。

object control

9)b Allow the engine$_i$ PRO$_i$ to idle and check DTC.

H: 让发动机怠速运转并检查DTC。

Tree diagram



PRO$_i$ to erase the DTC

PRO$_i$ to move away from the vehicle

Diagnosis:

Problem:

- Object control

Solution:

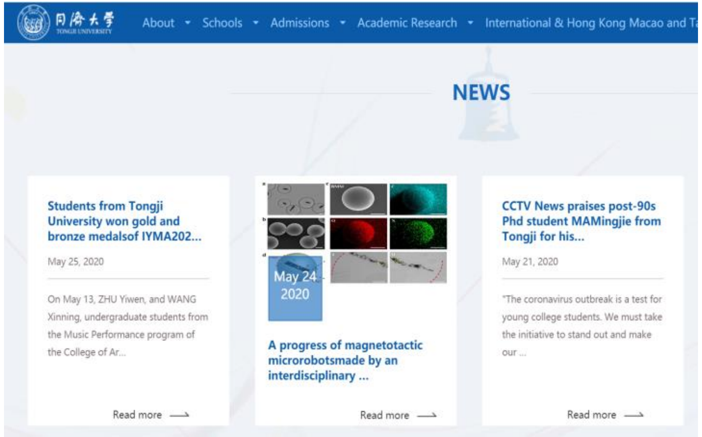
- To use PRO in MT rules

↓

Pre-editing

Example

Tongji University



NEWS

Students from Tongji University won gold and bronze medalsof IYMA202...

May 25, 2020

On May 13, ZHU Yiwen, and WANG Xinning, undergraduate students from the Music Performance program of the College of Ar...

Read more ⟶

May 24 2020

A progress of magnetotactic microrobotsmade by an interdisciplinary ...

Read more ⟶

CCTV News praises post-90s Phd student MAMingjie from Tongji for his...

May 21, 2020

"The coronavirus outbreak is a test for young college students. We must take the initiative to stand out and make our ...

Read more ⟶

- ➢ Term bank
- ➢ database

BUT ...

term

5月17日，同济大学基本完成了2020年全日制硕士研究生的网络远程复试工作，并顺利举行了2020年博士研究生的第一场网络远程复试。

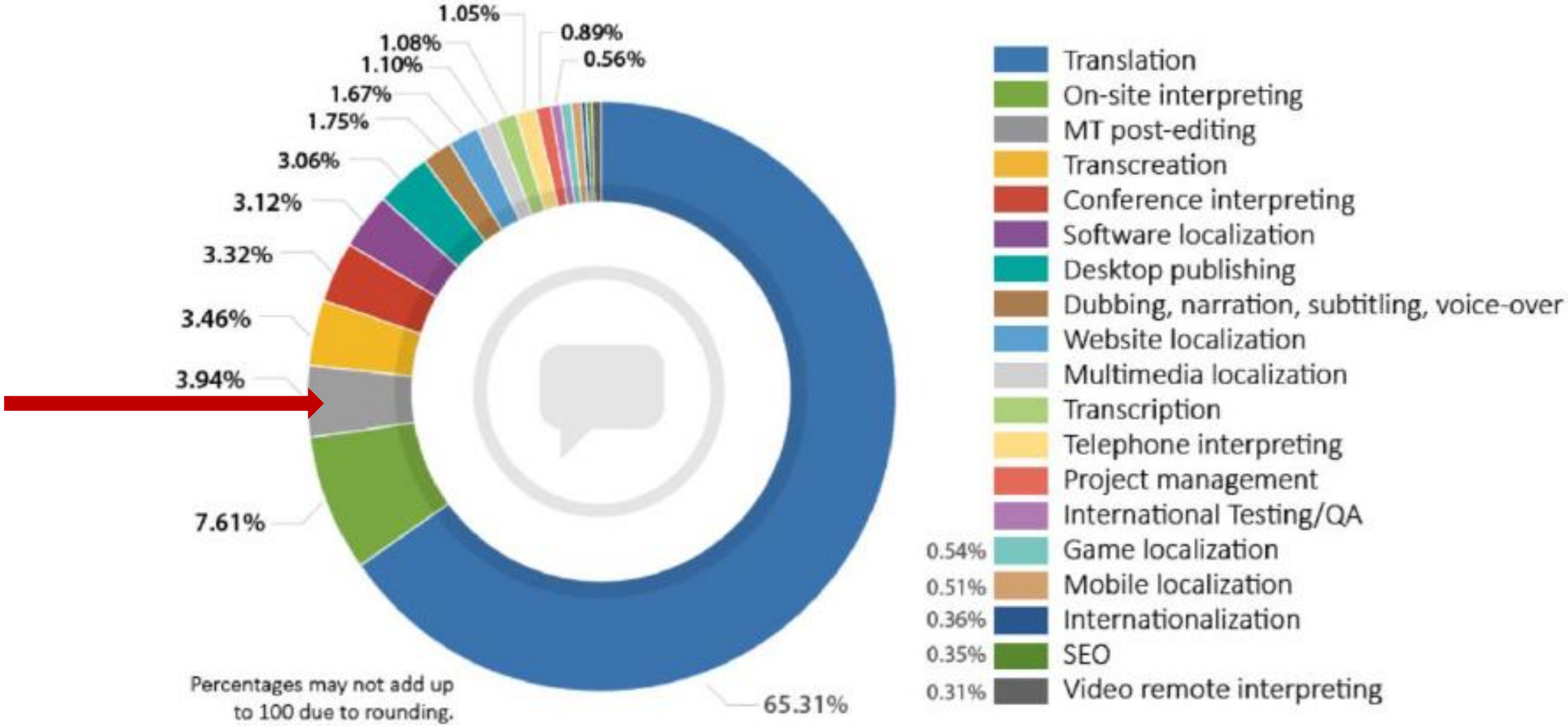
-- 同济新闻 20200517

复 试

the second round entrance examinations

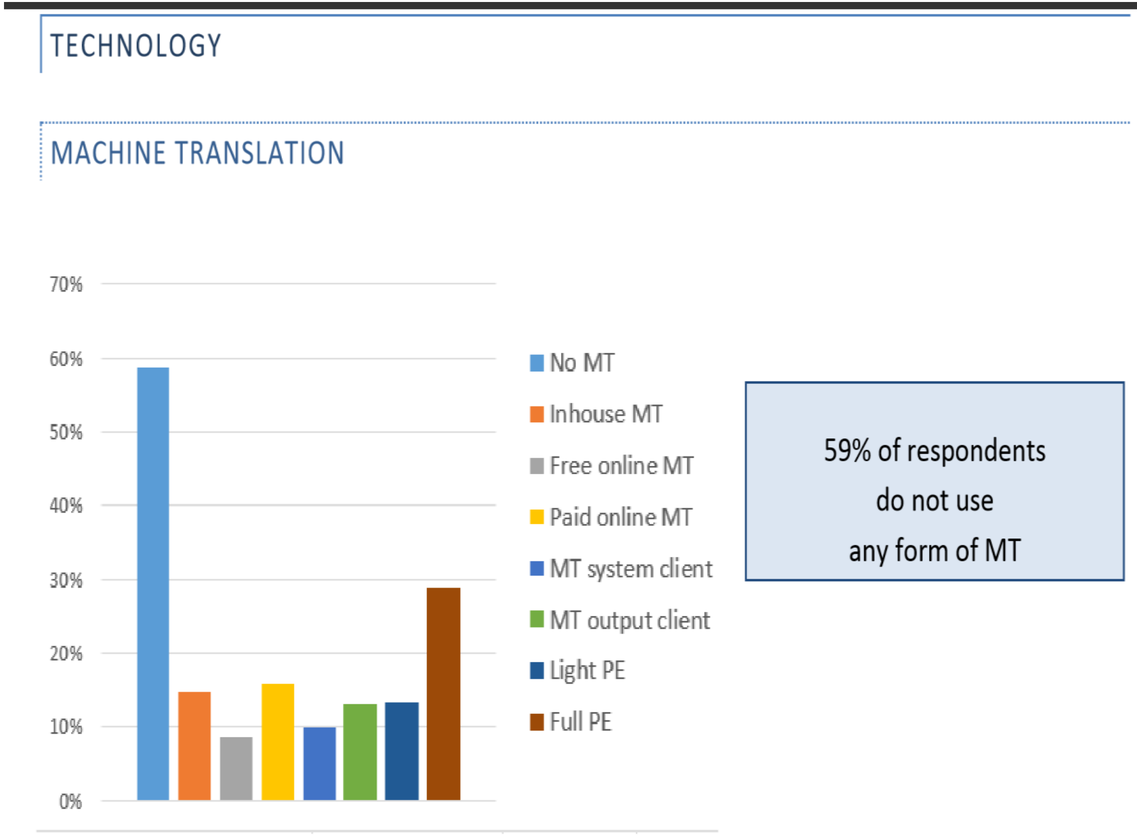
Google: retest
Baidu： retest, reexamine
Tencent： retest
Sogo： second interview
Youdao： the second interview,
secondary examination，
reexamination

全球语言服务的类型及比例



Percentages may not add up to 100 due to rounding.

The Language Services Market: 2016，Common Sense Advisory



TECHNOLOGY

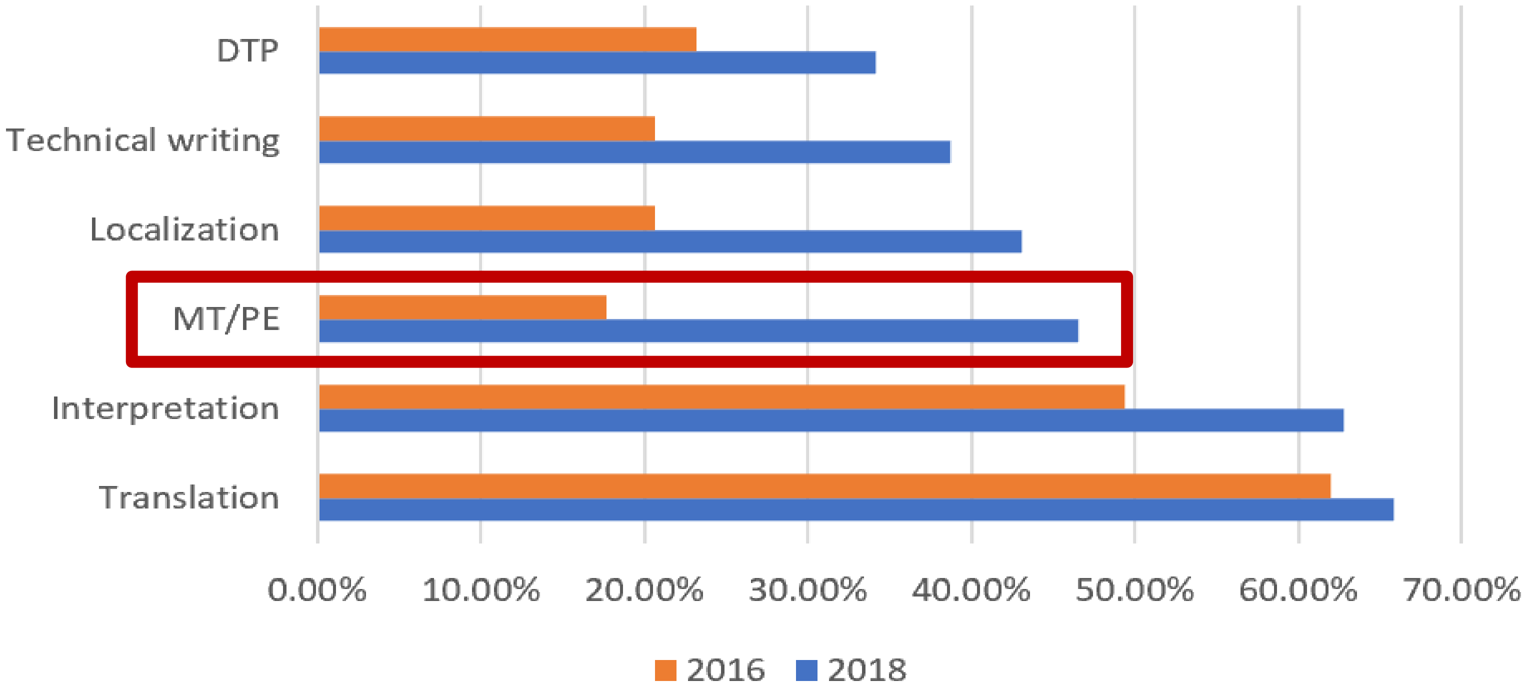
MACHINE TRANSLATION

59% of respondents
do not use
any form of MT

LSPs that are using MT usually provide their clients with fully post-edited output.

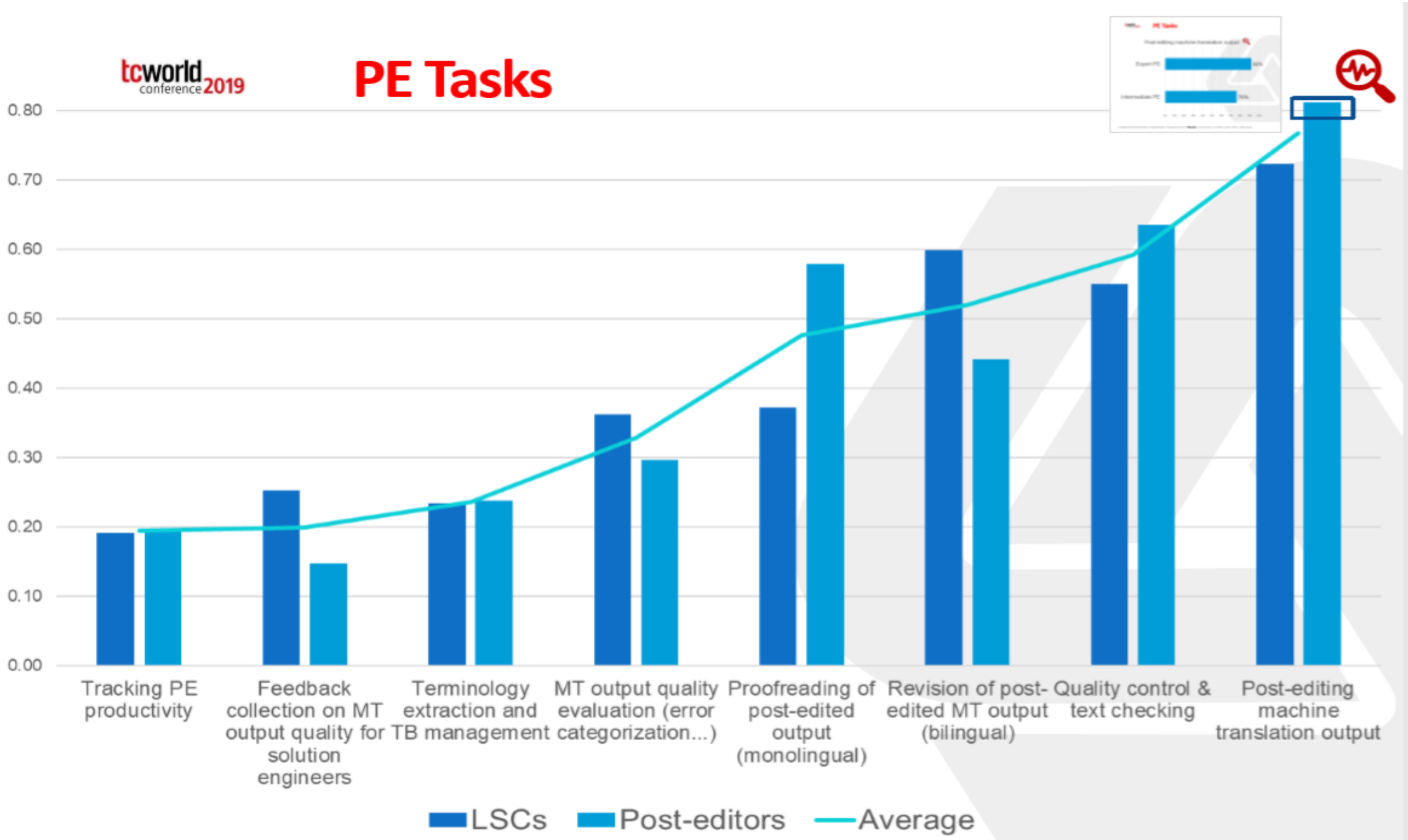
- 2016 Language Industry Survey

Increase of Language Service Demands in China



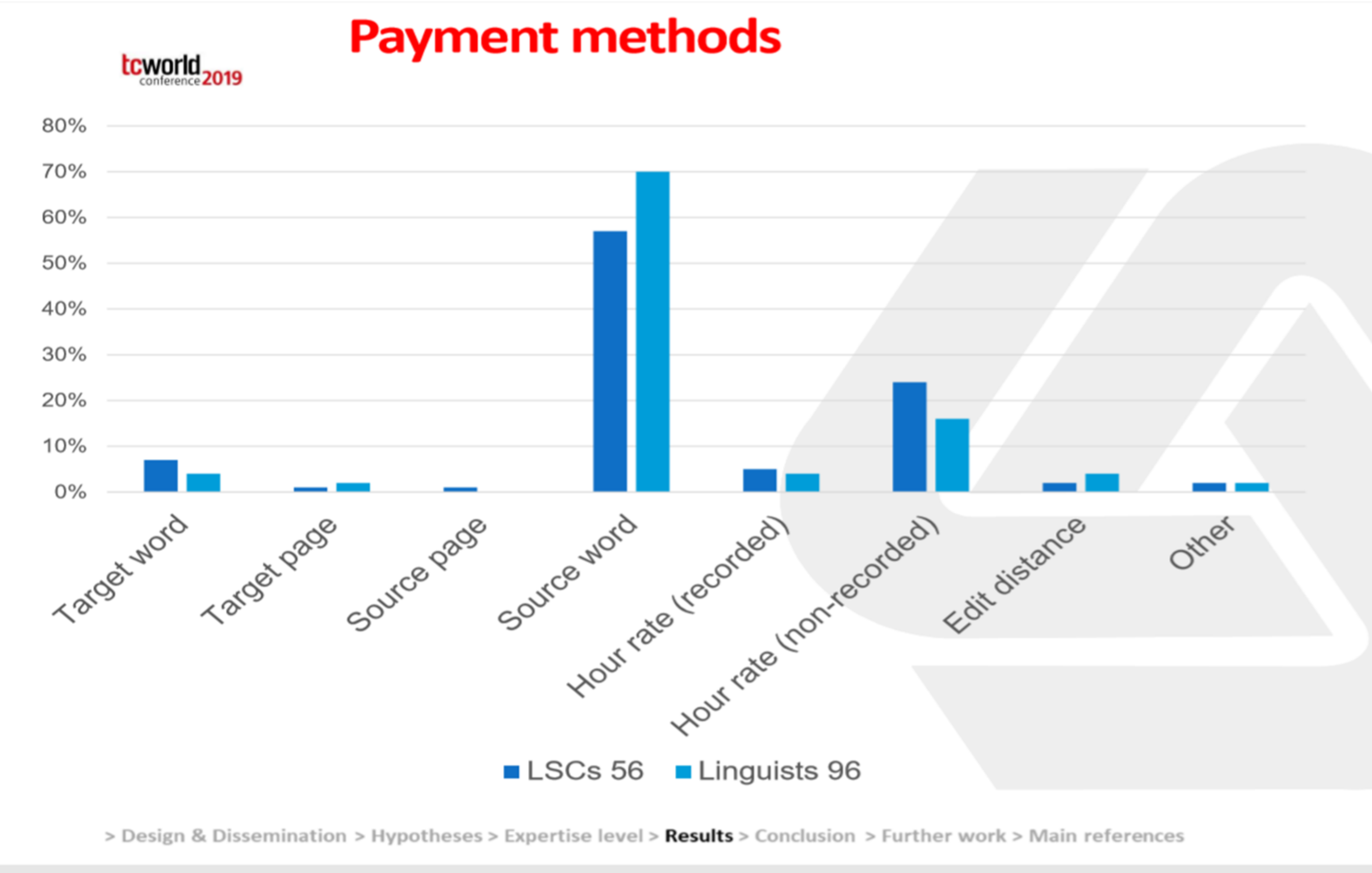
Source: China Language Service Development Report 2018

PE tasks



tcworld conference 2019

PE Tasks



tcworld conference 2019

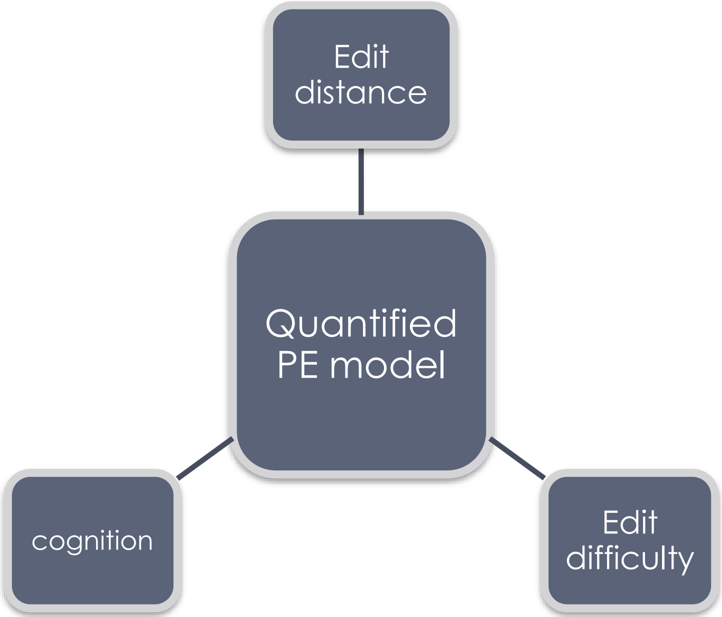
Payment methods

> Design & Dissemination > Hypotheses > Expertise level > **Results** > Conclusion > Further work > Main references

PE

payment

PE research



Quantified PE model

Edit distance

cognition

Edit difficulty

NMT-based Quantified PE Model

National Social Science Foundation 2019

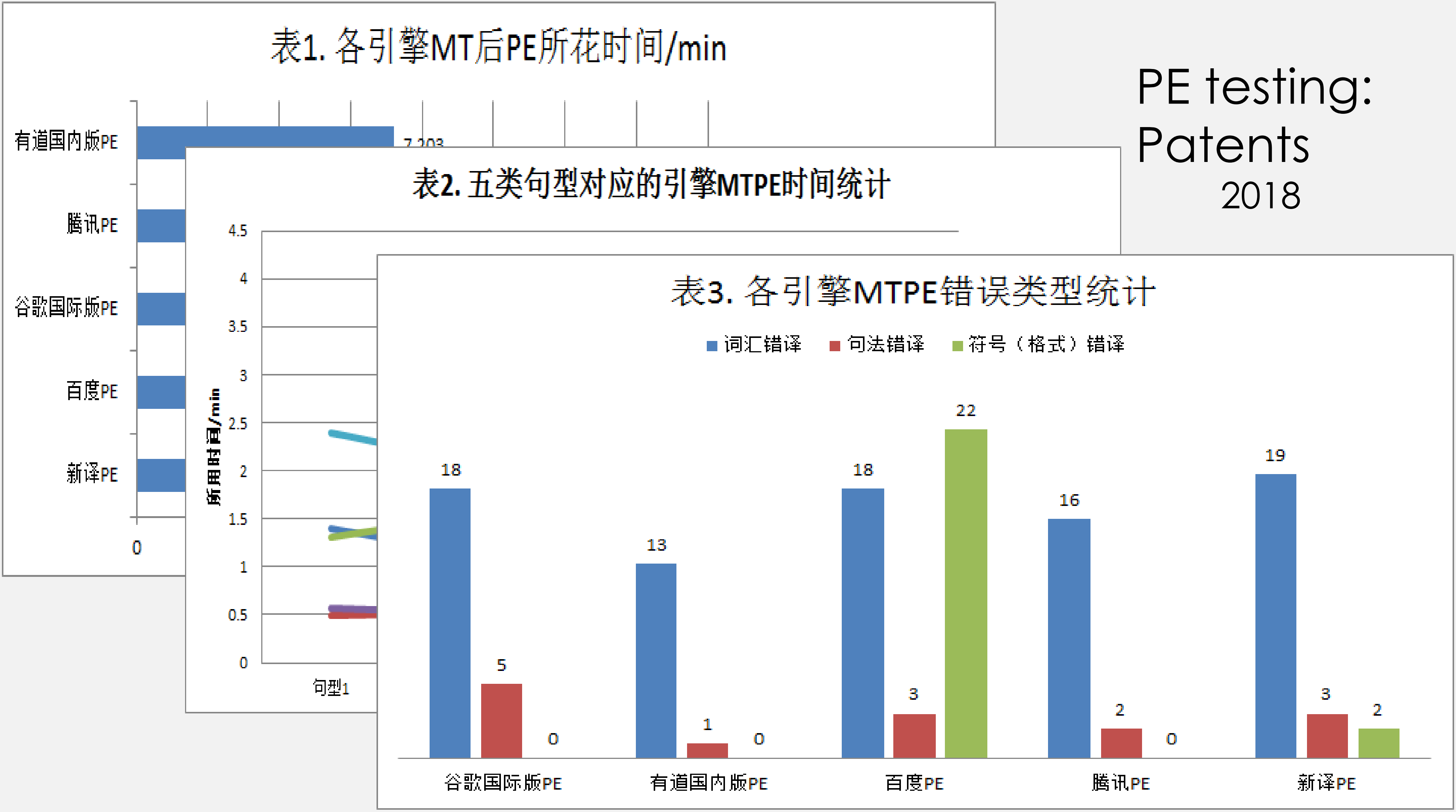
PE

productivity

Which MT?

PE perspective

PE testing: Patents
2018



PE testing: biology

2018

	谷歌国际	有道国内	百度	腾讯	新译
句型1：主谓定宾句					

	谷歌国际	有道国内	百度	腾讯	新译
句型2：无 句					

		谷歌国际	有道国内	百度	腾讯	新译
句型3：被动句	MT星级	★★★★☆	★★★★☆	★★★★☆	★★★☆☆	★★★☆☆

		谷歌国际	有道国内	百度	腾讯	新译
句型4：形容词谓语句	MT星级	★★★★	★★★★	★★★☆	★★★☆	★★★☆
	PE时间（分钟）					
	漏译					
	增译					
	误译					

		谷歌国际	有道国内	百度	腾讯	新译
句型5：动词补语句	MT星级	★★★★☆	★★★★☆	★★★☆☆	★★★☆☆	★★★☆☆
	PE时间（分钟）	15	14	17	12	12
	漏译	0	0	1	2	0
	增译	1	2	3	3	2
	误译	1	1	3	2	3

5 sent types

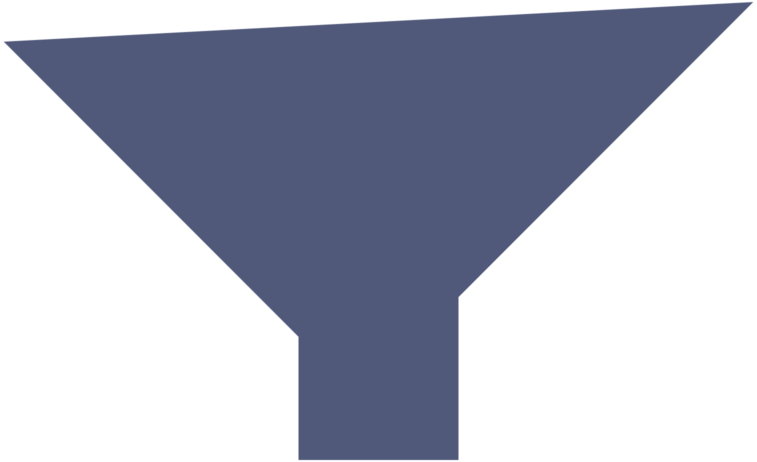
AI： to turn machines into men

- **past 30 years**
  - Turn men into machines
- **future 30 years**
  - Turn machines into men

- 马云：2017第四届世界互联网大会

- **eventually**
  - machines like men
  - men more like men

Thank you



Q&A

References

Chomsky, Noam. 2001. Preface by Chomsky. *Introducing Transformational Grammar: From Principles and Parameters to Minimalism.* London: Edward Arnold (Publishers) Limited, F13-19.

Haegeman, Liliane. 1995. *Introduction to Government & Binding Theory.* Oxford: Blackwell Publishers Ltd.

Alliance Clita. 2016 Language Industry Survey-Expectations and Concerns of the European Language Industry.

陈娜 2012. 机器译文与人工译文不定式句法结构的对比研究

李梅 2013. 译后编辑自动化的英汉机器翻译新探索，《中国翻译》第一期：83-87 （第一作者）

李梅 2013. 英汉机译错误分类及其数据统计分析，《上海理工大学学报》第四期：201-207

李梅 2012. 机器翻译译文错误分析，《中国翻译》第五期(第二作者)：84-89

李梅 2012. 中国名园英译策略探讨，《中国翻译》第一期：83-86