北京大學
PEKING UNIVERSITY

计算辅助翻译原理与实践　俞敬松

第5课 翻译记忆

Lecture 5: Translation Memory

第1节翻译记忆的缘起

- 翻译记忆？翻译记忆！
- “让历史告诉未来”

从机器翻译开始，可惜……

- 机器翻译是计算机处理自然语言的最直接最根本的动力
- 当然，代替人手工翻词典，也可算是动力之一吧…

- 机器翻译非常复杂
- 规则式机器翻译；统计式机器翻译；n种模型…

它，却成为了译员们工作中真正使用着的利器！

- 翻译记忆很简单。
- 记住做过的翻译，随后应用…

翻译记忆

Translation memory

From Wikipedia, the free encyclopedia

A **translation memory**, or **TM**, is a database that stores "segments", which can be sentences, paragraphs or sentence-like units (headings, titles or elements in a list) that have previously been translated, in order to aid human translators. The translation memory stores the source text and its corresponding translation in language pairs called “translation units”.

使用翻译记忆工具最简单最直接的动力

- 对个体： 节省劳动，提高劳动生产率
- 对翻译工作： 提高一致性
- 对翻译公司：

  整体生产率提高，生产成本降低；

  语言资产，是翻译公司核心竞争力的一部分

- 效率！ 质量！

翻译记忆的最初形态

- 计算机逐渐从实验室，计算中心变为个人的大玩具和工作平台：

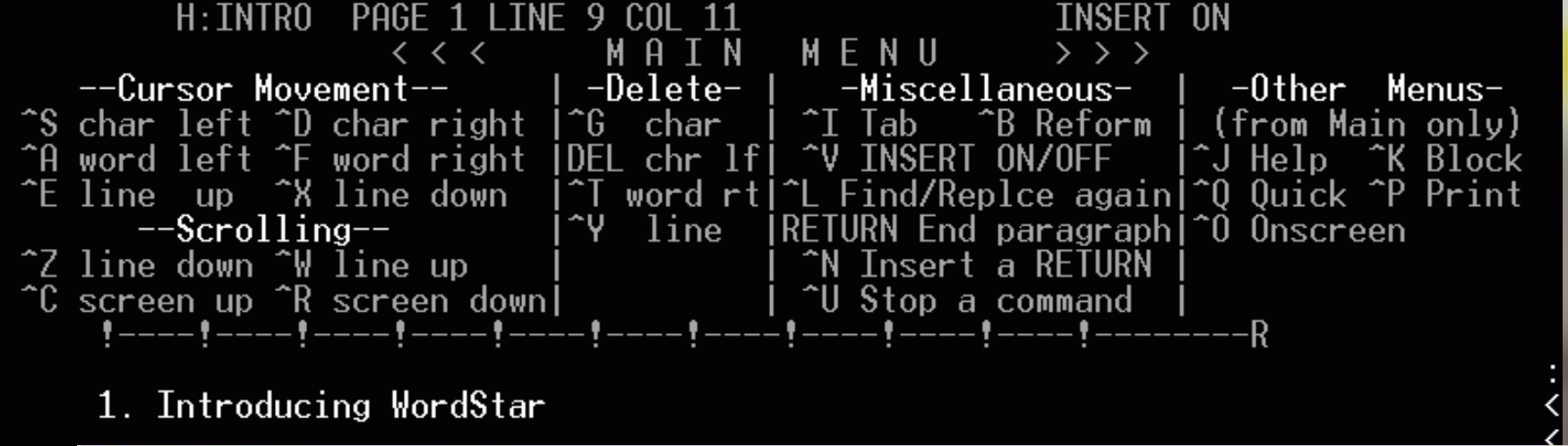
  译员们可以扔掉打字机 -> 字处理软件

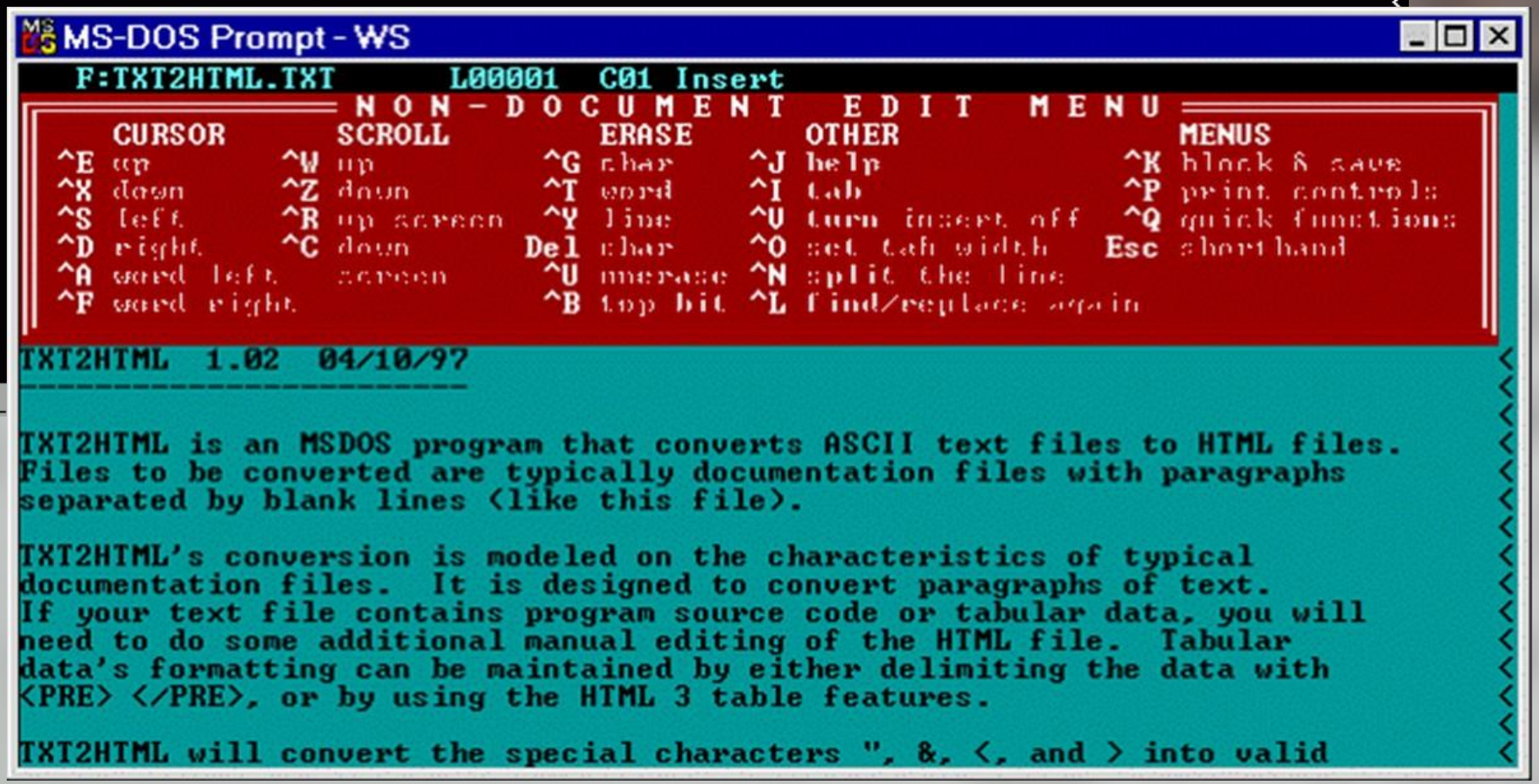
- 翻译工作量的增加
- 翻译工作的性质，工程文档越来越多，文字重复是这类文档区别于传统文学翻译的显著特点之一

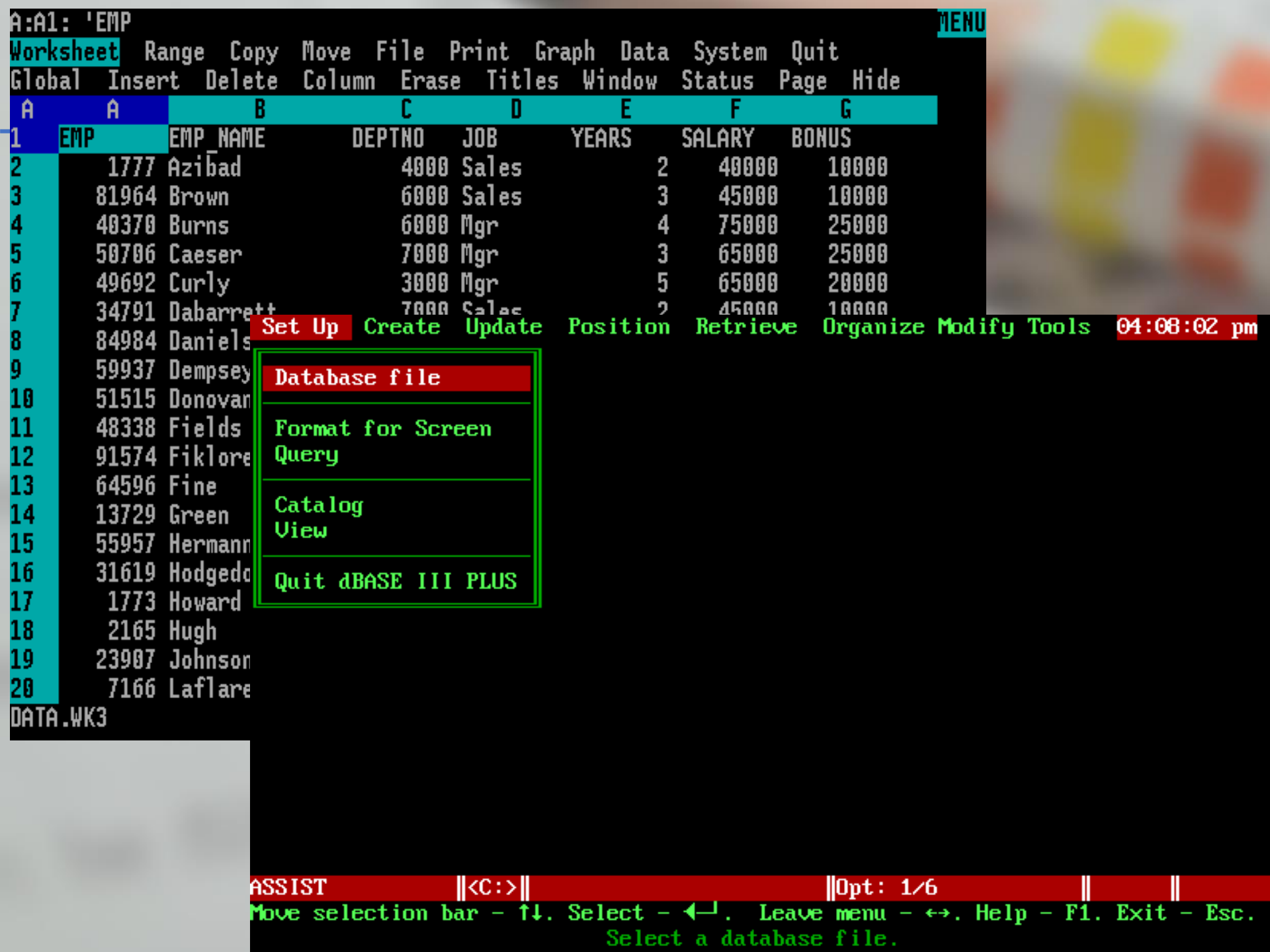
Apple –II

Sharp MZ-8

IBM-PC and IBM PC-XT

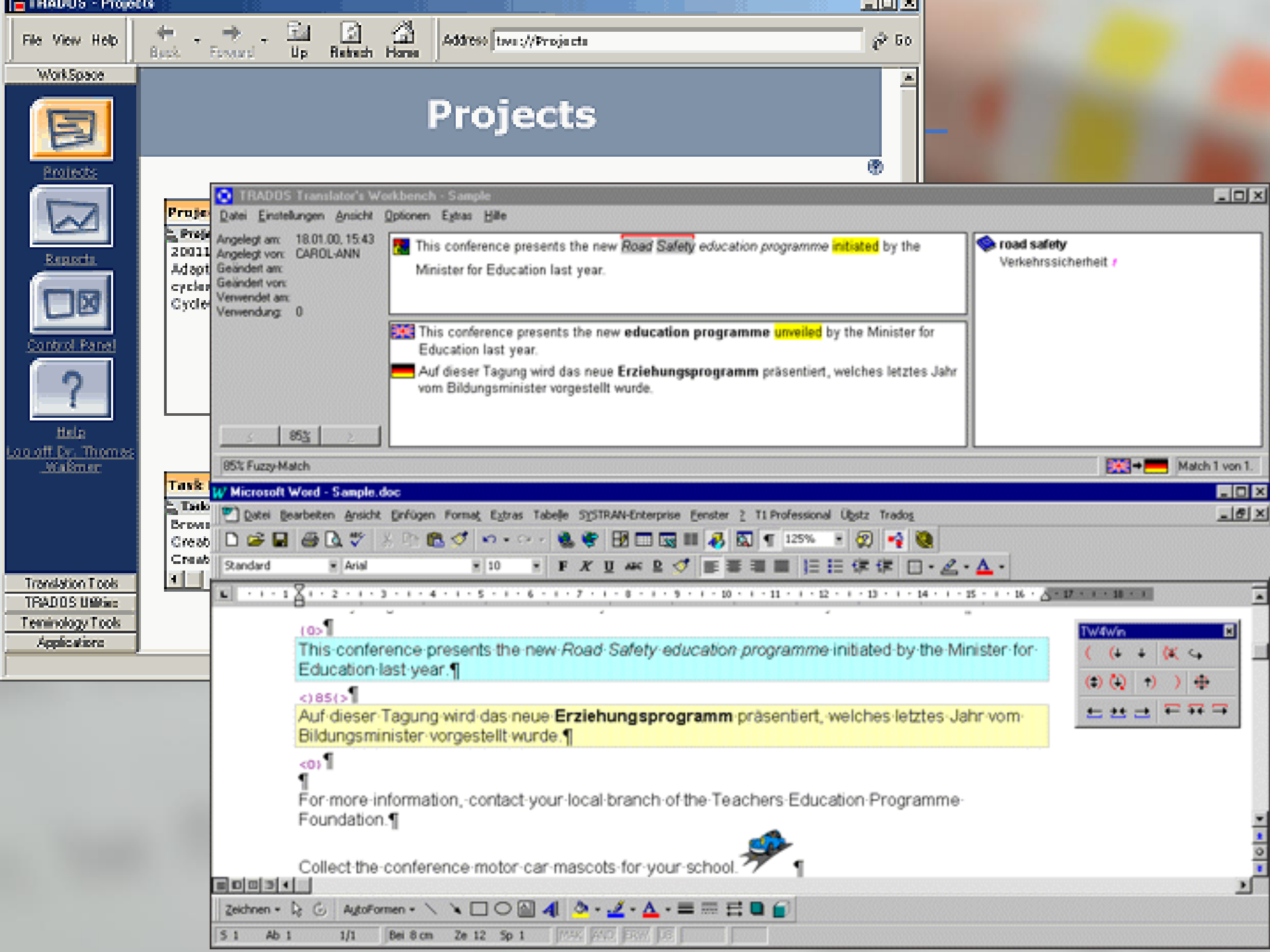

H:INTRO PAGE 1 LINE 9 COL 11 INSERT ON
< < < MAIN MENU > > >
--Cursor Movement-- | -Delete- | -Miscellaneous- | -Other Menus-
^S char left ^D char right | ^G char | ^I Tab ^B Reform | (from Main only)
^A word left ^F word right | DEL chr lf | ^V INSERT ON/OFF | ^J Help ^K Block
^E line up ^X line down | ^T word rt | ^L Find/Replce again | ^Q Quick ^P Print
--Scrolling-- | ^Y line | RETURN End paragraph | ^O Onscreen
^Z line down ^W line up | | ^N Insert a RETURN |
^C screen up ^R screen down | | ^U Stop a command |
1. Introducing WordStar


MS-DOS Prompt - WS
F:TXT2HTML.TXT L00001 C01 Insert
NON-DOCUMENT EDIT MENU
CURSOR SCROLL ERASE OTHER MENUS
^E up ^W up ^G char ^J help ^K block & save
^X down ^Z down ^T word ^I tab ^P print controls
^S left ^R up screen ^Y line ^U turn insert off ^Q quick functions
^D right ^C down Del char ^O set tab width Esc shorthand
^A word left screen ^U unerase ^N split the line
^F word right ^B top bit ^L find/replace again
TXT2HTML 1.02 04/10/97
TXT2HTML is an MSDOS program that converts ASCII text files to HTML files.
Files to be converted are typically documentation files with paragraphs
separated by blank lines (like this file).
TXT2HTML's conversion is modeled on the characteristics of typical
documentation files. It is designed to convert paragraphs of text.
If your text file contains program source code or tabular data, you will
need to do some additional manual editing of the HTML file. Tabular
data's formatting can be maintained by either delimiting the data with
<PRE> </PRE>, or by using the HTML 3 table features.
TXT2HTML will convert the special characters ", &, <, and > into valid


A:A1: 'EMP
MENU
Worksheet Range Copy Move File Print Graph Data System Quit
Global Insert Delete Column Erase Titles Window Status Page Hide
A A B C D E F G
1 EMP EMP_NAME DEPTNO JOB YEARS SALARY BONUS
2 1777 Azibad 4000 Sales 2 40000 10000
3 81964 Brown 6000 Sales 3 45000 10000
4 40370 Burns 6000 Mgr 4 75000 25000
5 50706 Caeser 7000 Mgr 3 65000 25000
6 49692 Curly 3000 Mgr 5 65000 20000
7 34791 Dabarrett 7000 Sales 2 45000 10000
8 84984 Daniels
9 59937 Dempsey
10 51515 Donovan
11 48338 Fields
12 91574 Fiklore
13 64596 Fine
14 13729 Green
15 55957 Hermann
16 31619 Hodgedo
17 1773 Howard
18 2165 Hugh
19 23907 Johnson
20 7166 Laflare
DATA.WK3
Set Up Create Update Position Retrieve Organize Modify Tools
04:08:02 pm
Database file
Format for Screen
Query
Catalog
View
Quit dBASE III PLUS
ASSIST
<C:>
Opt: 1/6
Move selection bar - ↑↓. Select - ←┘. Leave menu - ↔. Help - F1. Exit - Esc.
Select a database file.

翻译记忆技术发展的推动者们

- Martin Kay: “Proper Place”, Peter Arthern
- Alan Melby, Sumita , Tsutsumi
- ALPS(Automated Language Processing Systems) Tools
- Sadler and Vendelmans' Bilingual Knowledge Bank
- Brian Harris "Bi-text"
- **The first commercial TM tool： Trados** (SDL Trados nowadays)


TRADOS - Projects
File View Help
Back Forward Up Refresh Home
Address tws://Projects
Go
WorkSpace
Projects
Reports
Control Panel
Help
Logoff for Thomas Wallner
Translation Tools
TRADOS Utilities
Terminology Tools
Applications
Projects
TRADOS Translator's Workbench - Sample
Datei Einstellungen Ansicht Optionen Extras Hilfe
Angelegt am: 18.01.00, 15:43
Angelegt von: CAROL-ANN
Geändert am:
Geändert von:
Verwendet am:
Verwendung: 0
This conference presents the new Road Safety education programme initiated by the Minister for Education last year.
road safety
Verkehrssicherheit
This conference presents the new education programme unveiled by the Minister for Education last year.
Auf dieser Tagung wird das neue Erziehungsprogramm präsentiert, welches letztes Jahr vom Bildungsminister vorgestellt wurde.
85%
85% Fuzzy-Match
Match 1 von 1.
Microsoft Word - Sample.doc
Datei Bearbeiten Ansicht Einfügen Format Extras Tabelle SYSTRAN-Enterprise Fenster ? TI Professional Ubstz Trados
Standard
Arial
10
125%
This conference presents the new Road Safety education programme initiated by the Minister for Education last year.
Auf dieser Tagung wird das neue Erziehungsprogramm präsentiert, welches letztes Jahr vom Bildungsminister vorgestellt wurde.
For more information, contact your local branch of the Teachers Education Programme Foundation.
Collect the conference motor car mascots for your school.
TW4Win
Zeichnen
AutoFormen

翻译记忆工作的流程

计算机怎么判断“句子”？

怎么检索？

- 读取待翻译的**句子**；
- 将待翻译的句子送到翻译记忆库中**检索**，看看有无翻译过；
- 如果不存在已经翻过的句子，就等待译员翻译并打字输入。翻译结束后，将原文和译文存入翻译记忆库；
- 如果存在即有的翻译句，那么就输出既往的译文，供译员使用或适当的予以订正；
- 循环重复，直到译完最后一句待翻译句，翻译工作全部完成。

第2节翻译记忆若干关键问题之一

- 翻译记忆存储单位——翻译单元
- 获得翻译记忆

翻译记忆存储单位——翻译单元

- 工作 – 拆解 > 子工作
- 翻译工作 – 拆解 > 子翻译工作

- 子翻译工作？

字、词、短语、片段、子句、句子、句群、段落乃至篇章

单位大些还是小些？

- 必要性（可重复利用吗？）

  重复出现的可能性

  重复出现时，翻译是否可以使用呢？

  使用翻译记忆必须检查后再利用！

- 可行性 （能否做到？）

  不是所有的自然语言单位都有清晰定义

  不是所有的单位都容易分辨处理

翻译单元的最佳选择

- 亚句级别的单位：词、词组、语块、子句
- 句子级别
- 句子以上级别：句组和段落

句子的自动识别和切分

- 句子识别也是问题吗？

  对人来说， 如果达到数学般精确，显然是

  对计算机来说，肯定是

- 分隔标志：标点符号

  英语：句点（ . ）、问号（ ？ ）以及叹号（ ！ ）

句点的功能

http://en.wikipedia.org/wiki/Period_(punctuation)

- 数字　3.1415926535　192.168.0.1
- 缩写　U.S.A.　,　Rev.
- 称呼　Mr. Mrs. , George W. Bush.
- 省略号的一部分
- 分隔号 <-> 逗号 对某些语言来说可能有些互换
- www.cnn.com　,　http://cattp.pkucat.com

- "Carefree" means "free from care or anxiety."
- "Carefree" means "free from care or anxiety".

识别句子的方法

- 规则方法

  穷尽所有的可能性， 例如诸如正则表达式这样的工具进行识别

  写得越多，越要注意重复和冲突

- 统计方法

  给定一个语料库，人工判断哪些句点是句子的分隔符做好标记；

  机器学习程序： 训练 -> 获得模型 -> 处理其他文章（解码）

亚句级别的翻译记忆

- 确实有用，因为重复出现的可能性高多了；
- 但不好做，因为虽然重复出现得多，但出现多种翻译的可能性也增加了
- 与术语管理还是有区别的
- 加入此项技术后， 是否给译员增加负担？
- 机器翻译的发展，例如层次短语模型，树模型，EBMT等，都可能在这个技术的中间地带得到利用

另辟蹊径，不划分翻译单元可以吗？

- 似乎可以，利用信息检索技术
- 直接存储原文和译文，大致找到对应关系（对齐）
- 关键词倒排索引
- 检索，相似度计算，找到更相似片段
- 输出原文和译文段，译员自行判断和使用

- 程序很好写，用起来也还行，就是效率低 < 致命伤

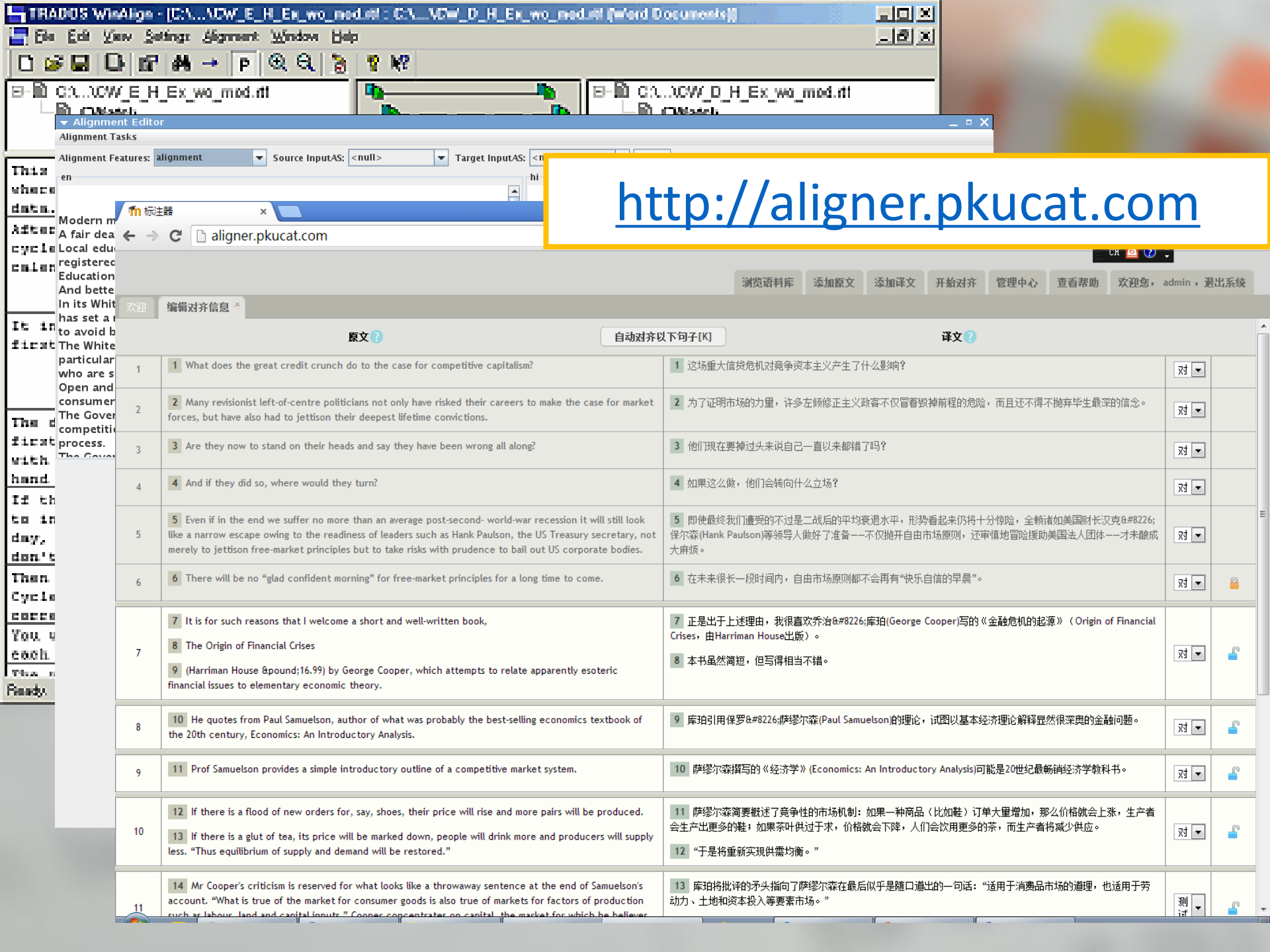
获得翻译记忆

- 一边翻译，一边存储　　最自然，最直接
- 存储成……？

  特定软件的翻译记忆库文件

  开放格式的记忆库文件　　TMX

- 批量导入过去译好的文件建立翻译记忆库


TRADOS WinAlign - [C:\...\CW_E_H_Ex_wo_mod.rtf : C:\...\CW_D_H_Ex_wo_mod.rtf (Read Documents)]
File Edit View Settings Alignment Window Help
C:\...\CW_E_H_Ex_wo_mod.rtf
C:\...\CW_D_H_Ex_wo_mod.rtf
Alignment Editor
Alignment Tasks
Alignment Features: alignment
Source InputAS: <null>
Target InputAS:
http://aligner.pkucat.com
标注器
aligner.pkucat.com
浏览语料库
添加原文
添加译文
开始对齐
管理中心
查看帮助
欢迎您， admin ，退出系统
欢迎
编辑对齐信息
原文
自动对齐以下句子[K]
译文
1 What does the great credit crunch do to the case for competitive capitalism?
1 这场重大信贷危机对竞争资本主义产生了什么影响?
2 Many revisionist left-of-centre politicians not only have risked their careers to make the case for market forces, but have also had to jettison their deepest lifetime convictions.
2 为了证明市场的力量，许多左倾修正主义政客不仅冒着毁掉前程的危险，而且还不得不抛弃毕生最深的信念。
3 Are they now to stand on their heads and say they have been wrong all along?
3 他们现在要掉过头来说自己一直以来都错了吗?
4 And if they did so, where would they turn?
4 如果这么做，他们会转向什么立场?
5 Even if in the end we suffer no more than an average post-second- world-war recession it will still look like a narrow escape owing to the readiness of leaders such as Hank Paulson, the US Treasury secretary, not merely to jettison free-market principles but to take risks with prudence to bail out US corporate bodies.
5 即使最终我们遭受的不过是二战后的平均衰退水平，形势看起来仍将十分惊险，全赖诸如美国财长汉克•保尔森(Hank Paulson)等领导人做好了准备——不仅抛开自由市场原则，还审慎地冒险援助美国法人团体——才未酿成大麻烦。
6 There will be no “glad confident morning” for free-market principles for a long time to come.
6 在未来很长一段时间内，自由市场原则都不会再有“快乐自信的早晨”。
7 It is for such reasons that I welcome a short and well-written book,
8 The Origin of Financial Crises
9 (Harriman House £16.99) by George Cooper, which attempts to relate apparently esoteric financial issues to elementary economic theory.
7 正是出于上述理由，我很喜欢乔治•库珀(George Cooper)写的《金融危机的起源》（Origin of Financial Crises，由Harriman House出版）。
8 本书虽然简短，但写得相当不错。
10 He quotes from Paul Samuelson, author of what was probably the best-selling economics textbook of the 20th century, Economics: An Introductory Analysis.
9 库珀引用保罗•萨缪尔森(Paul Samuelson)的理论，试图以基本经济理论解释显然很深奥的金融问题。
11 Prof Samuelson provides a simple introductory outline of a competitive market system.
10 萨缪尔森撰写的《经济学》(Economics: An Introductory Analysis)可能是20世纪最畅销经济学教科书。
12 If there is a flood of new orders for, say, shoes, their price will rise and more pairs will be produced.
13 If there is a glut of tea, its price will be marked down, people will drink more and producers will supply less. “Thus equilibrium of supply and demand will be restored.”
11 萨缪尔森简要概述了竞争性的市场机制：如果一种商品（比如鞋）订单大量增加，那么价格就会上涨，生产者会生产出更多的鞋；如果茶叶供过于求，价格就会下降，人们会饮用更多的茶，而生产者将减少供应。
12 “于是将重新实现供需均衡。”
14 Mr Cooper's criticism is reserved for what looks like a throwaway sentence at the end of Samuelson's account. “What is true of the market for consumer goods is also true of markets for factors of production
13 库珀将批评的矛头指向了萨缪尔森在最后似乎是随口道出的一句话：“适用于消费品市场的道理，也适用于劳动力、土地和资本投入等要素市场。”
Ready

自动对齐双语句子的原理

•语言长度的等比例现象

不同语言的对等翻译之间，以字符/词计算的长度比例是近似固定的

英语：汉语　　字符比大约　　3:1

源语言方向会更精炼些

英语词：汉语字符　大约1.6:1

自动对齐双语句子的原理

- 两种语言都先分句，数出每一个句子的字符数
- 动态规划方法：

  对齐模式1:1, 1:0, 0:1, 1:2, 2:1, …

  全局扫描，达到每句话的对应比例最均衡(评价函数)

  各种动态规划算法，提高效率
- 改进：

  词汇信息对应

  先发现对齐最准确的锚点句后再对齐其他句子

  ……

双语句对齐的工作实践

- 纯文本的对齐？带格式文本的对齐！
- 先对齐格式，再对齐文本

  对齐软件必须懂得识别文件格式,例如DOC

- 自动对齐后，为了达到译员可以使用的水平，必须有手工校正的环节

  软件设计的水平，用户体验的高下，多人协同

- 作为翻译公司资产的重要部分

  翻译记忆库的生产和管理

  翻译记忆库的分发，安全，回收，版本，质量……

第3节翻译记忆若干关键问题之二

- 翻译记忆的存储和使用问题
- 句子之间的相似度计算
- 匹配

翻译记忆库

- 直观感觉：

一句一句的存储，但是语言的存在需要环境，即上下文，才有更明确意义！

翻译记忆库越大越好，重复率才能高啊…

翻译记忆库的存储顺序

- 文本的意义确实与上下文环境相关
- 前面反复强调译员使用翻译记忆的规则：

  译员在沿用翻译记忆之前必须检查核对！

- 检查的是什么？上下文！

- 上下文对于机器来说可以理解为：窗口

自然语言处理中的“窗口”

- 窗口就是指待处理对象的周围环境。文字是按照字符序列写作和存储的，“窗口”显然就是前后文字了。
- 待处理对象：词汇，句子，……

- 窗口大小：向前，向后观察的距离

  待翻译的句子的前后句子数量

  对于翻译任务，选择 ±1 ，即前一句，后一句

上下文“匹配”，完美“匹配”

- 存储一个待翻译的句子的原文和译文对的同时，存储前一句和后一句
- 如果下次复用的时候，发现，不但当前翻译句和原来存储的句子一致，而且前后句子都一模一样，那么：

“上下文匹配”，“完美匹配”就有了

翻译记忆库的大小

- 越大越好？大就行了吗？
- 翻译记忆库，在某种意义上来说，自然是大些好
- 前提是：

1. 译员使用的翻译记忆库必须是与自己当前的翻译工作是相关的

2. 翻译记忆库运行所需的软件和硬件拥有相当的存储和检索能力

翻译记忆库的使用规则之二

- 相关的判定准则公式：

任务 > 客户 > 领域 > 行业 > 通用

- 翻译记忆库的适当大小

  除了硬件软件能力之外

  译员的认知负担是最需要考虑的问题

句子之间的相似度

- 人可以模糊判定两个句子是否相似，但计算机不行
- 完全一致，最容易，即字符对字符的一致
- 但此规则有时候会显得过于严格

相似度计算的意义和目标

- 检验待翻译句与翻译记忆库中的存储句子之间的相似度：

  必须有一定的弹性和宽容度

- 翻译句子的相似程度最好能够告诉译员，便于译员进行判断

相似度计算有难度

•待翻译句：

昨天，张先生欣喜地来到了恭王府

•翻译记忆库：

(1) 1997年3月5日，张先生高兴的来到了北京城。

(2) 大前天，乔峰欣喜地来到了石家庄。

(3) 庚子年八月的一天，玉格格沉着脸来到了恭王府。

先不考虑其他，只在字符级别计算

- 莱文斯廷编辑距离（Levenshtein distance）算法

http://en.wikipedia.org/wiki/Levenshtein_distance

http://zh.wikipedia.org/wiki/%E7%B7%A8%E8%BC%AF%E8%B7%9D%E9%9B%A2

编辑距离，又称**Levenshtein距离**，是指两个字串之间，由一个转成另一个所需的最少编辑操作次数。许可的编辑操作包括将一个字符替换成另一个字符，插入一个字符，删除一个字符。

当然，编辑所需的原子操作中，替换也可以用删除和插入替换，或者增加更多的操作种类，但不影响我们的讨论。

编辑距离计算的经典例子

•For example, the Levenshtein distance between "kitten" and "sitting" is 3, since the following three edits change one into the other, and there is no way to do it with fewer than three edits:

“kitten” —> "sitting"

kitten → **s**itten (substitution of "s" for "k")

sitt**e**n → sitt**i**n (substitution of "i" for "e")

sittin → sitting (insertion of "g" at the end).

显然，用前述方法计算任意两个句子的“距离”，即相似度差值，是完全可以做到了。

编辑距离计算的各种优化方法

- 计算对象，可以是：字符，也可以是词
- 编辑动作的权重可以调整
- 部分语言部件可以捆绑为一个计算单位，或根本不进行计算。“非译元素”：数字，网址；格式标签

编辑距离计算的各种优化方法

- 词汇间的语义差值计算需要依赖：

  普林斯顿大学 WordNet

  北京大学中文概念辞书 CCD

- 词汇语义的差值，可以修正编辑距离的计算

  比如： 高兴 vs. 欣喜

编辑距离计算的各种优化方法

- 命名实体的检测：http://en.wikipedia.org/wiki/Named_entity

- 命名实体(named entity)：所谓的命名实体就是人名、机构名、地名以及其他所有以名称为标识的语言单位。广义的命名实体还可以包括表达数字、日期、货币、地址等等的语块。

编辑距离计算的各种优化方法

- 计算编辑距离前，先进性命名实体的识别

- 昨天，张先生欣喜的来到了恭王府
- [日期]，[人名]欣喜地来到了[地名]

- 1997年3月5日，张先生高兴的来到了北京城。
- [日期]，[人名]高兴的来到了[地名]

编辑距离计算的各种优化方法

- 句法结构上的相似度
- 子句及其他亚句级的对齐、相似度计算

- 识别源语言句子的句法结构，比如结构树，或者依存树都行，考虑句子的相似度先考虑结构，再考虑词汇；识别出子句或其他亚句级单位，进行相似度计算

- 事实证明，某些实验室工作很有“创新”意义；但在商业软件中很难得到应用：

译员们工作时的认知负担！越小的负担，越好的体验！

相似度计算—>匹配问题

- 字符串级别的一致：Exact Match
- 商业翻译记忆系统的标配：非译元素

- 非译元素也不大可能解决的问题：

a. There are 4 goats on the hill. 4：非译元素

b. There are 2 goats on the hill. 2：非译元素

c. There is a goat on the hill. a：……? Goat vs. Goats ?

模糊匹配（Fuzzy Match）

- 句子相似度计算算法得到的可能对译员有参考价值的句子**排序输出**；
- 相似度算法的要点就在于让相似度数值的排序，与译员心目中翻译记忆库中的可能参考的句子的可用性成稳定的**正比例关系**；
- 相似度的数值有多种算法，对于相似度数值的解释也可能各不相同。

模糊匹配的相似度排序

- 极值： 100%， 精确匹配， 0%，没有任何相似

- 模糊匹配的阈值，通常由译员自行设定。
- 记忆库大小（越大的翻译记忆库，越可能有更多的相似句子；认知负担高低（输出句子越多，阅读负担越大）

第4节翻译记忆工具的优缺点

- 适用性
- 优缺点分析

翻译记忆工具的适用性

- 工程性质的文档，适用性最高
- 科技类次之
- 人文社科、文学类最低

- 重复现象是否存在？
- 重复出现的句子，是否能使用同样的译文输出？

翻译任务的交付

- 原文，翻译要求
- 原文，翻译要求，曾经的翻译文档（过去的版本），相关文献等
- 原文，翻译要求，翻译记忆库，术语库

- 翻译记忆库的批量建立
- 翻译记忆库的选择性使用

翻译工作的交付

- 译文
- 翻译记忆库
- 术语库

商业翻译记忆工具的优点

- “做过的事情就不要再做了”，节省翻译的重复劳动
- 增加了翻译的一致性

  相同句段译文的一致

  原文与译文格式一致

  与术语系统结合，保证术语翻译的一致性
- 提高翻译速度：不仅仅是重复发生，优秀的适合翻译工作的译员工作平台，节省了大量的繁琐工作
- 降低了翻译成本：可惜的是，翻译记忆未必增加译员的个体工资
- 方便了翻译语言资源的查找与管理：语言资产的建立和维护

翻译记忆工具的缺点

- 翻译记忆工具的理论缺陷
  - 翻译记忆库提供的参考译文的可用性难以保证
  - 翻译记忆的质量管控：谁有资格写入，谁有资格读出和使用？
  - 翻译记忆系统割裂了原文的连续性（曾经）
  - 基于句段的翻译模式造成了翻译中句序调整的困难
- 技术环境下工作带来的问题
  - 翻译人员能否适应，培训的需要，学习的代价和负担
  - 逐句段翻译的习惯带来的弊病，缺乏上下文观念，依赖性、惰性
  - 翻译记忆管理的专业化需求，翻译公司对专业化分工管理的需要
- 成本问题
  - 商业软件价值比较高昂，但分摊成本还是非常有限的
  - 管理维护带来的人工成本
  - 不断地升级带来的学习成本
  - 与其他管理系统的融合，也会带来成本

翻译记忆工具的发展方向

- 自然语言处理技术的发展

  相似度计算

  机器翻译/交互式机器翻译

- 用户界面、用户体验
- 网络化工作平台
- 翻译、翻译工作管理、翻译工程管理、翻译公司管理系统的融合
- 新的云计算技术、大数据技术的引入

软件学习建议

- 仔细观看软件操作演示录像：**一边看，一边做**
- 读软件说明书、读帮助
- 保持好奇心，多尝试
- 有不可解决的问题，多多参与讨论

谢谢观看！