Japanese translation and the computer

the past, the present and the future

Jim Breen

June 20, 2007

Machine Translation - The Beginnings

- Postulation on a Universal Language (Pascal, Leibnitz, Wilkins)
- Early Patents (Artsouni, 1933; Troyanskii, 1933,1937)
- The Weaver Memo (1949)
- First Full-time Researcher (Bar-Hillel, MIT, 1951)
- First MT Conference (1952)
- Demonstration of a Simple Russian-English System (1954)
- First Japanese System (Kuno, Harvard, 1960)

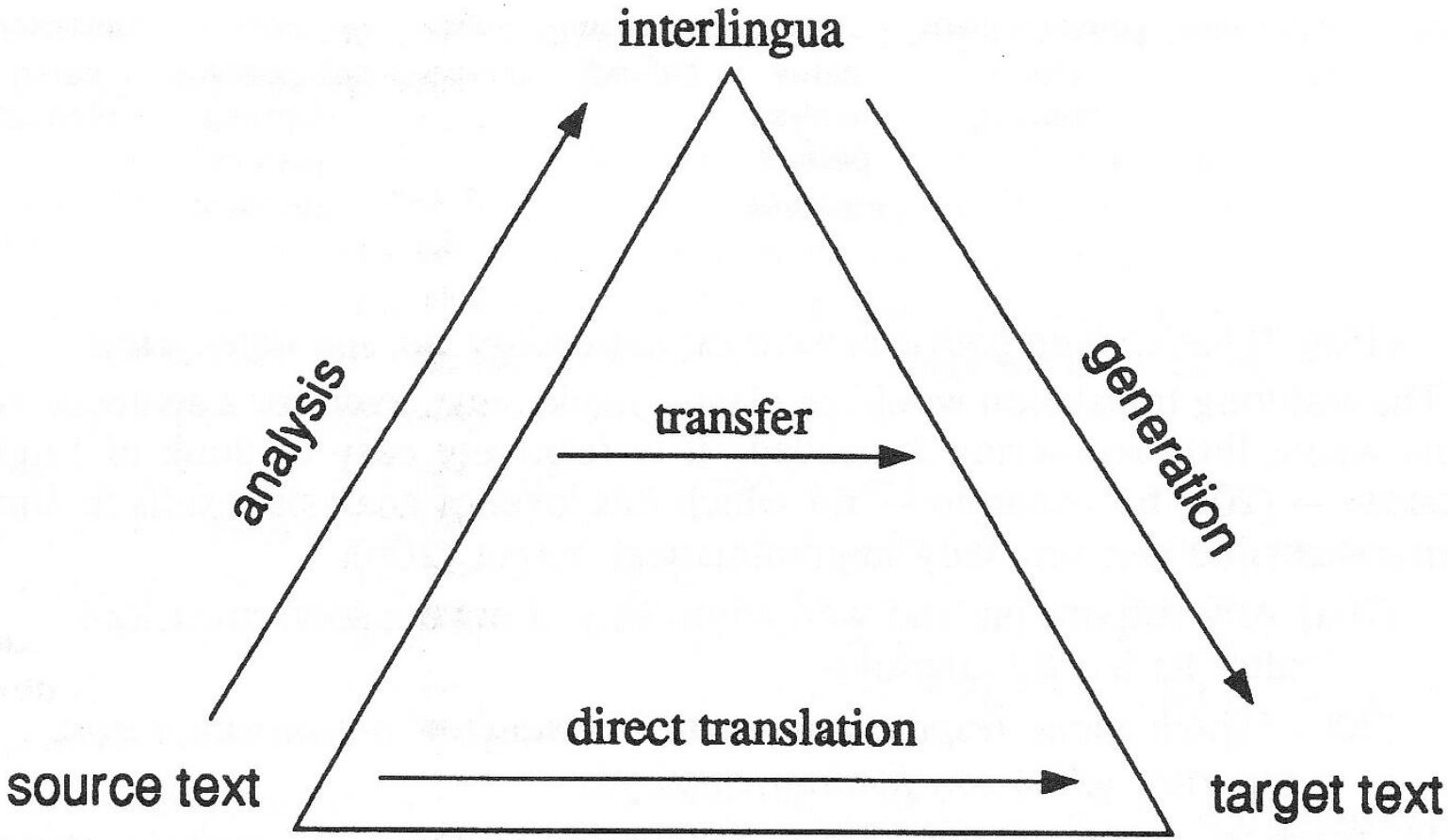
Growing Disappointment

- Early work in MT was actually pioneering CL, AI, etc.
- Formal linguistic theory and analysis was lagging
- Conflicts: brute-force vs perfectionist; high-quality vs pragmatism
- Optimism faded - could quality translation be achieved? (FAHQT)
- US Government appoints Automatic Language Processing Advisory Committee (1964)
- ALPAC reports: *"there is no immediate or practicable prospect of useful Machine Translation"* (1966)
- MT research declines in the US; continues in Europe, Canada, etc.

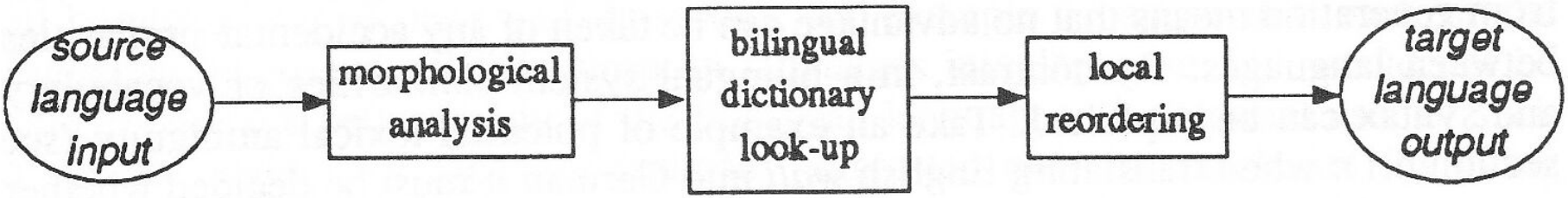
MT Varieties (1)

- Direct Translation
- Transfer
- Interlingual
- Example-based
- Statistical

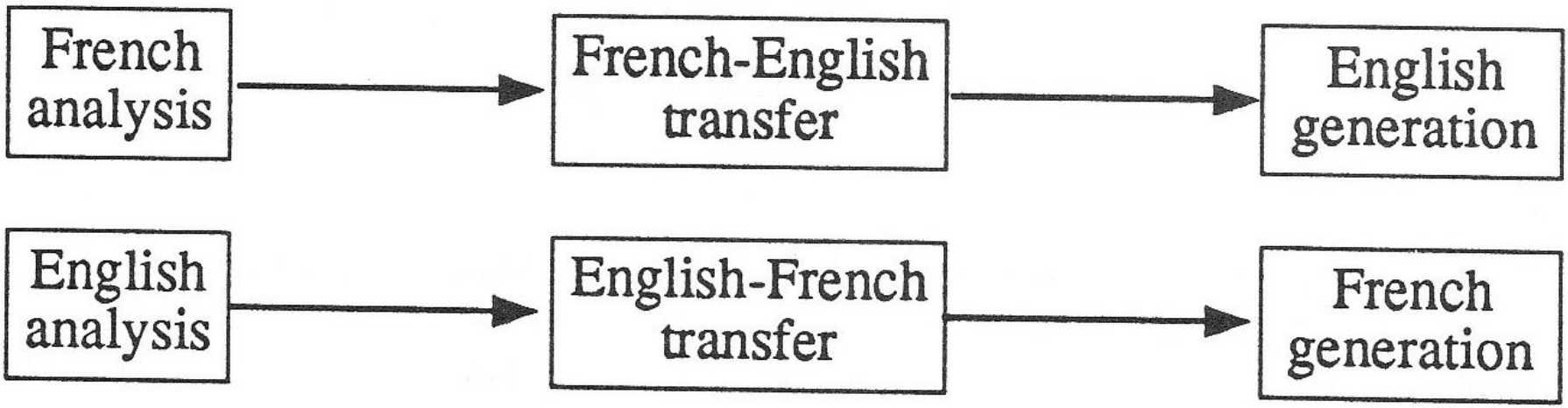
MT Varieties (2)



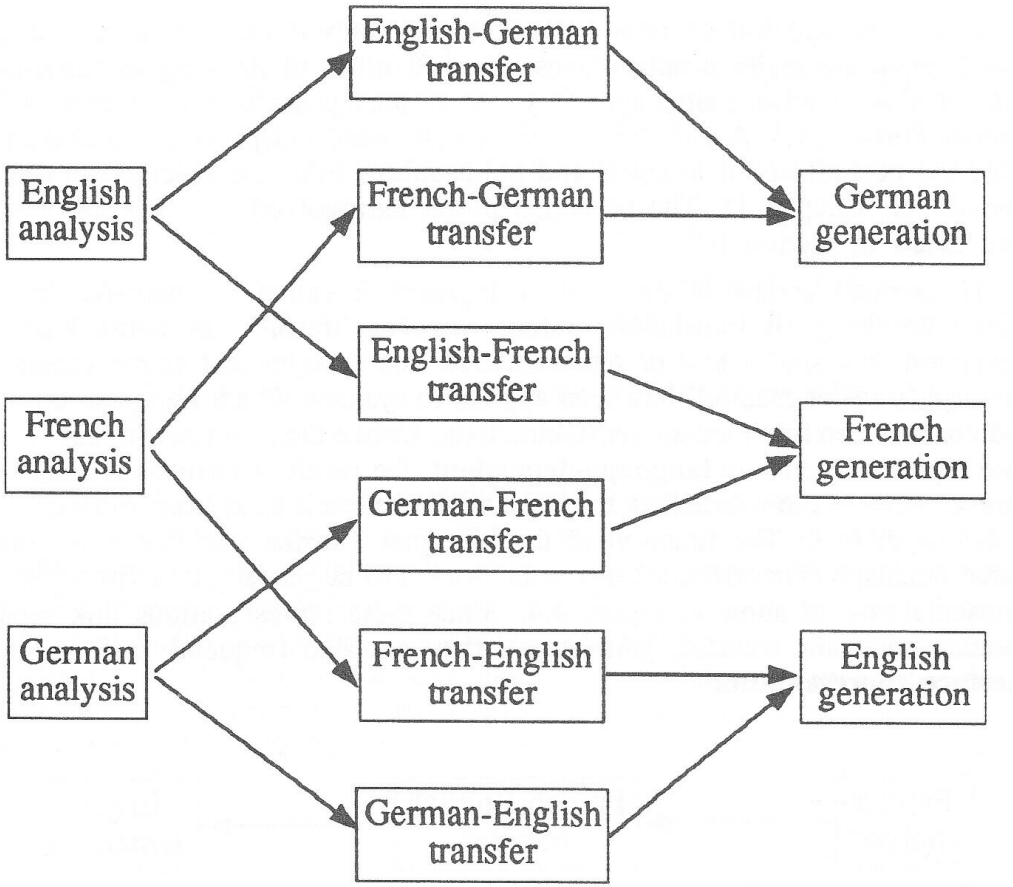
Direct Translation



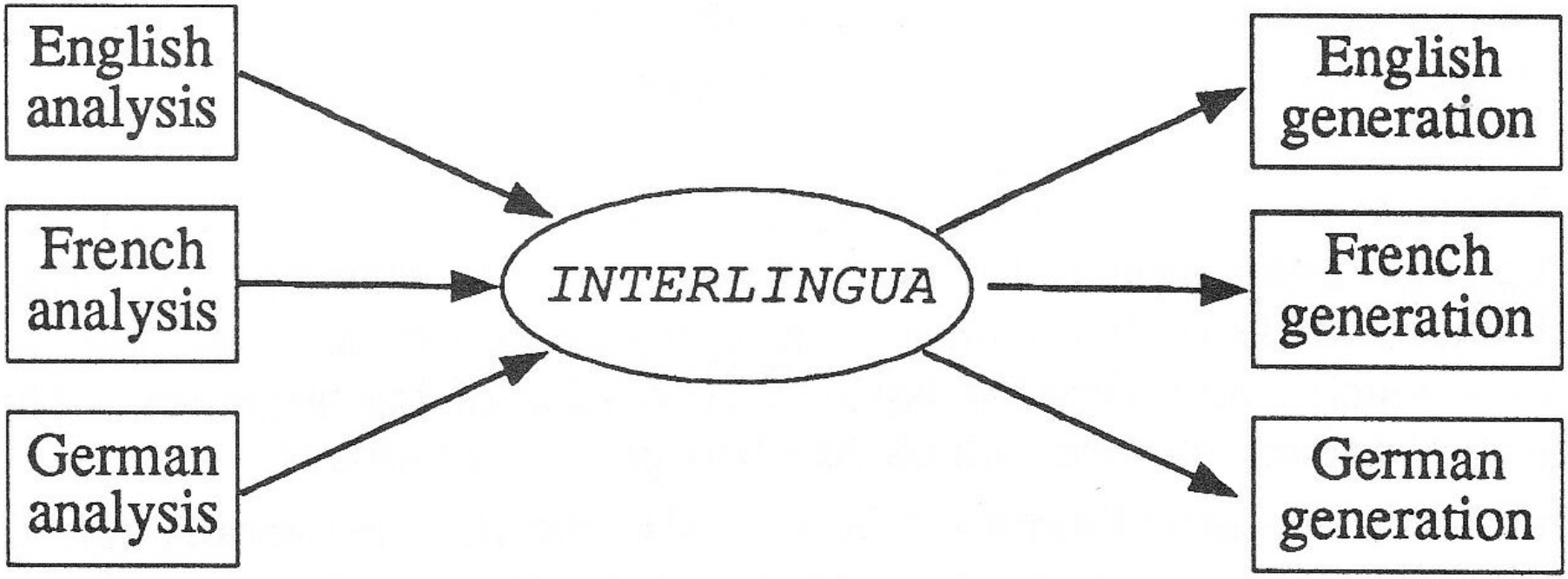
Transfer (Simple)



Transfer (Complex)



Interlingual



Example-based

- First proposed by Nagao (1984)
- uses a large collection of phrase/sentence templates
- system trained with a bilingual corpus
- used by many low-cost Japanese systems

Example-based (example)

あの赤い傘はいくらですか。

あの小さいカメラはいくらですか。

Statistical MT

- First suggested by Weaver in 1949
- Based on information theory, uses calculated probability that a target string is the translation of a source string
- Serious work begun by IBM in 1991
- Relies on large quantities of parallel bilingual texts, and significant computing power
- Language-independent
- Can be word, phrase, sentence-based
- Major focus of current research

Commercial Systems

Metéo:

- Used to translate weather report/forecasts in Canada (1976)
- Special "sub-language"

Systran:

- Descended from early Russian-English systems of the 1960s
- Adopted by the EEC (EU) in 1976
- Now has many language pairs
- Many WWW-based systems (Babelfish, Yahoo, etc.)

Atlas:

- Developed by Fujitsu (late 70s)
- Well-regarded for E-J.
- Very large lexicon

Translator Survey - Questionnaire & Responses

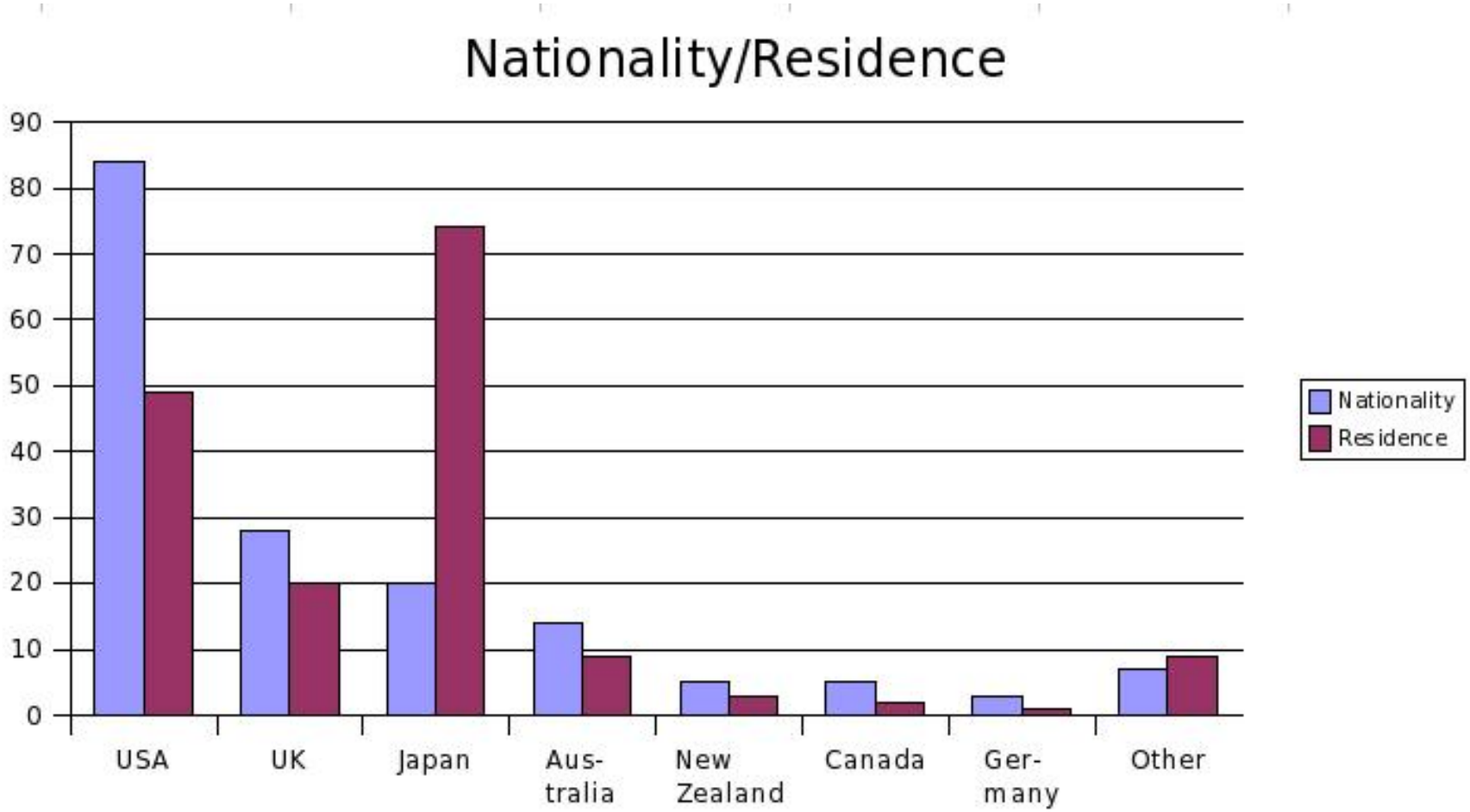
Goal: to find out:

- what translators of Japanese were doing with computers
- what their views were of current and future impacts of ICT on translation

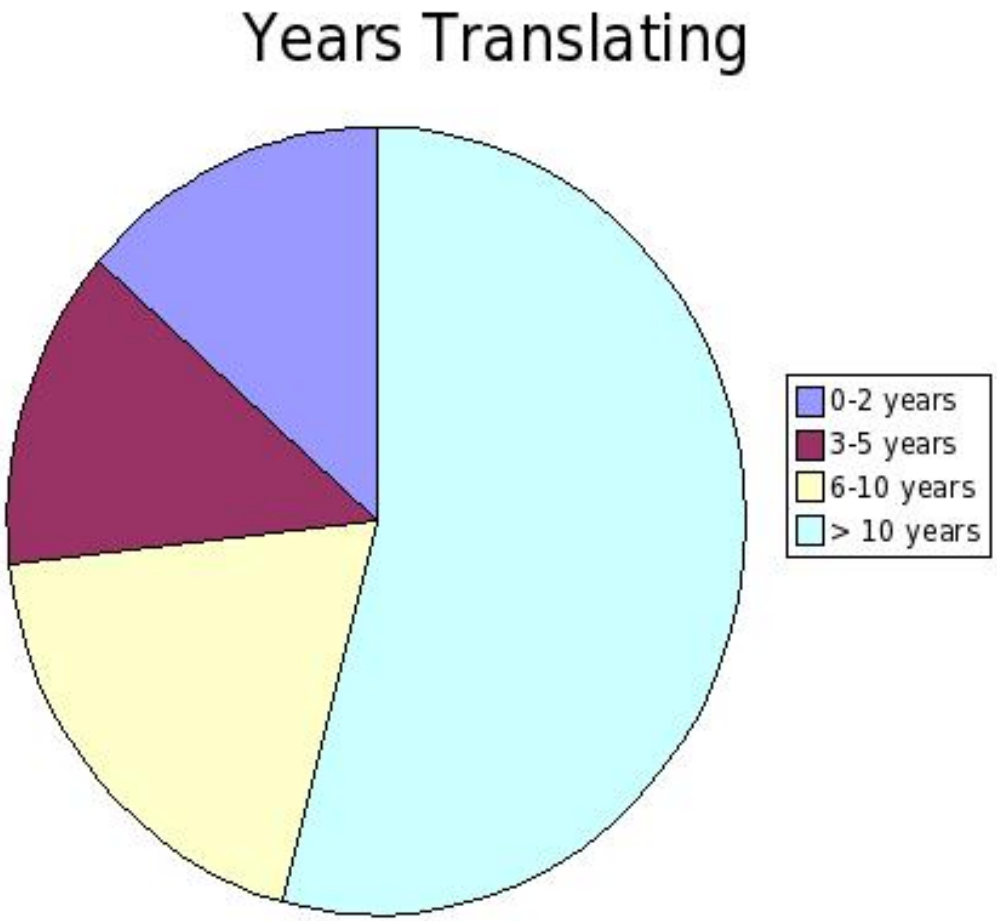
Questionnaire:

- WWW-based questionnaire (CGI program & templates)
- Open for 3 weeks in May 2007
- 171 useable responses

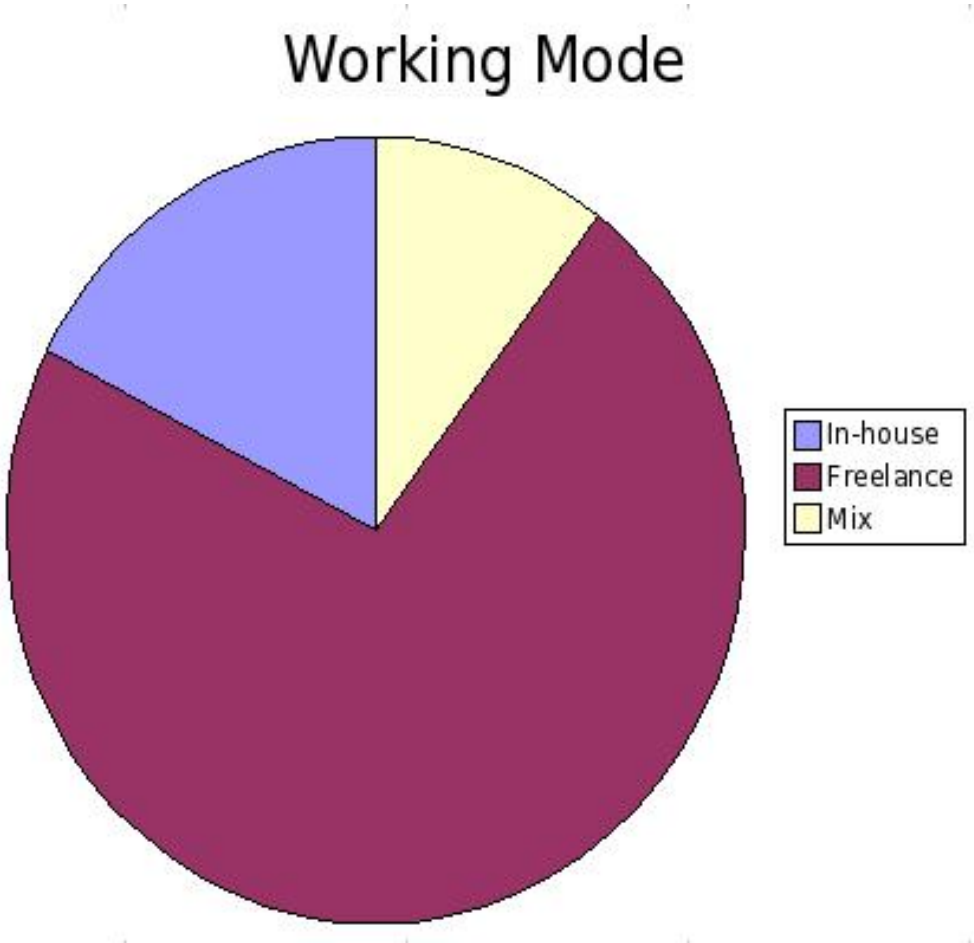
Nationality & Residence



Years Translating



Working Mode



Other Details

- From Japanese: 75.6%, into Japanese 4.5%, Mix: 19.9%
- Sole occupation: 65.3%, main occupation: 22.7%, second occupation: 11.9%
- Windows: 80.7%; Macintosh: 19.3%

The Honyaku Archive

The Honyaku archive of postings (1994-present) provides interesting insights into the evolution of use of computers by translators.

- Email
- Word-processing and document types
- The WWW
- Translation Memories

Email - the changing discourse.

- 1994: The days of TWICS. Almost everything in romaji. Kanji by Nelson numbers.
- 1995: Occasional kanji/kana (usually mojibake). Mentions of Win/V etc.
- 1996: Many kanji/kana postings - almost always repeated in romaji.
- 1998: Most postings using kanji/kana; occasional romaji
- 2000: Most postings using kanji/kana; romaji is rare
- Document exchange with clients went from rare to common

Word-processing and document types

- Early emphasis on JWP, NJStar, Ichitaro, etc.
- Growing concentration on Word. Dominant by 2000.
- 1996: First mention of PDFs. Rapid increase from there.
- PowerPoint translations appearing by 2000.
- Steady expansion of role from **text** translation to **document** translation

The arrival of the WWW

Adam Rice (July 1994)

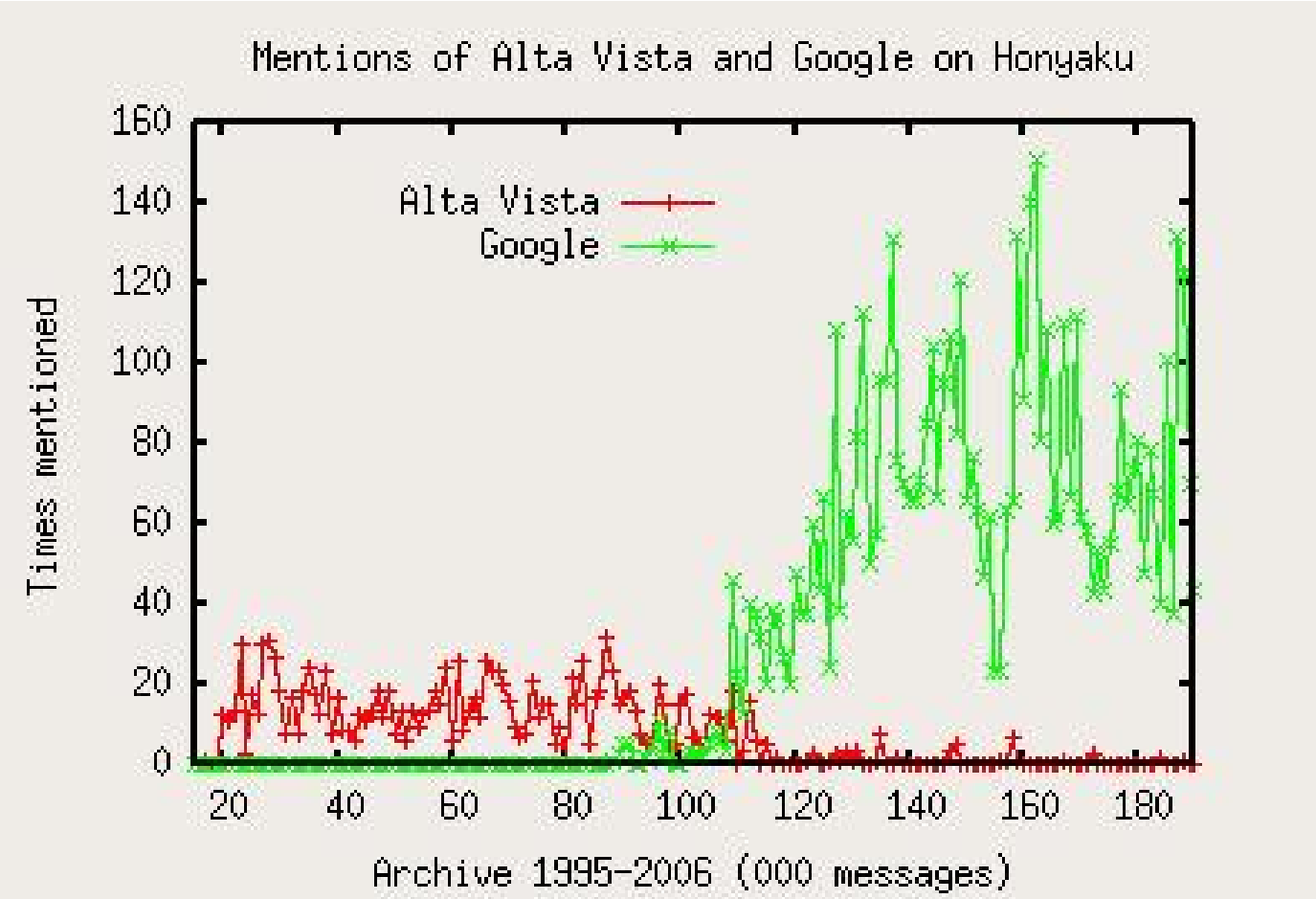
- *... Some of you may be acquainted with the World Wide Web (WWW) ...*
- *The amount and diversity of information available on the Web is staggering....*
- *I also believe it is going to become the primary form of dealing with the Internet in the near future....*
- *I am considering creating a "Honyaku Home Page" on the Web ...*

The Rise of WWW use

Honyaku and the WWW

- 1995: First mention of search engines
- 1996: Alta Vista, Lycos, etc. being used. Problems with Japanese
- 1999: First mention of Google
- 1999: Honyaku moves to Onelist, which merges with eGroup
- 2001: Yahoo takes over eGroup
- 2002: First mention of "St. Google"; James Sparks coins "Googits"
- 2006: Honyaku moves to Google groups
- WWW searching topics have become a common part of the discussion

The Rise of Google



The Arrival of Translation Memories

Trados

- Nov. 1996 first mentioned on Honyaku
- Jun. 1997 E-J demonstration at IJET 8
- Feb. 1998 Trados Tokyo Office, hiring programmers, JE "in beta"
- Jul. 1998 Minoru Mochizuki uses Trados for E-J.
- Nov. 1998 Presentation to JAT
- 2001 - rising numbers of users, many price concerns

The Arrival of Translation Memories

Déja Vu

- Nov. 1998 First mentioned in a job advert
- Sep. 1999 J-E capability "in a month or two"
- 2003 J-E capability "next release"

Wordfast

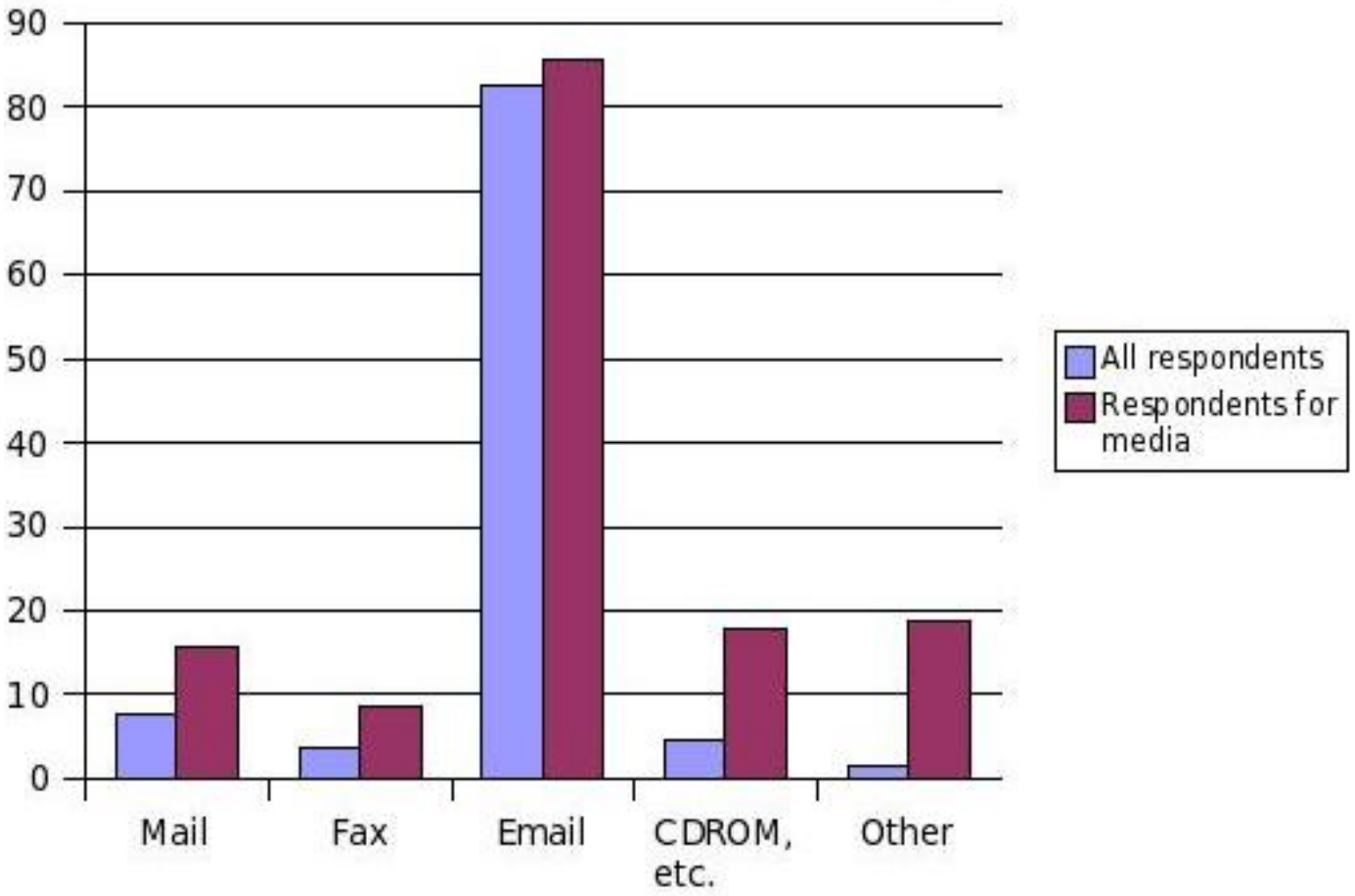
- 2001 First mention (Gururaj Rao tried it)
- 2002 More people trying it

Translator Survey

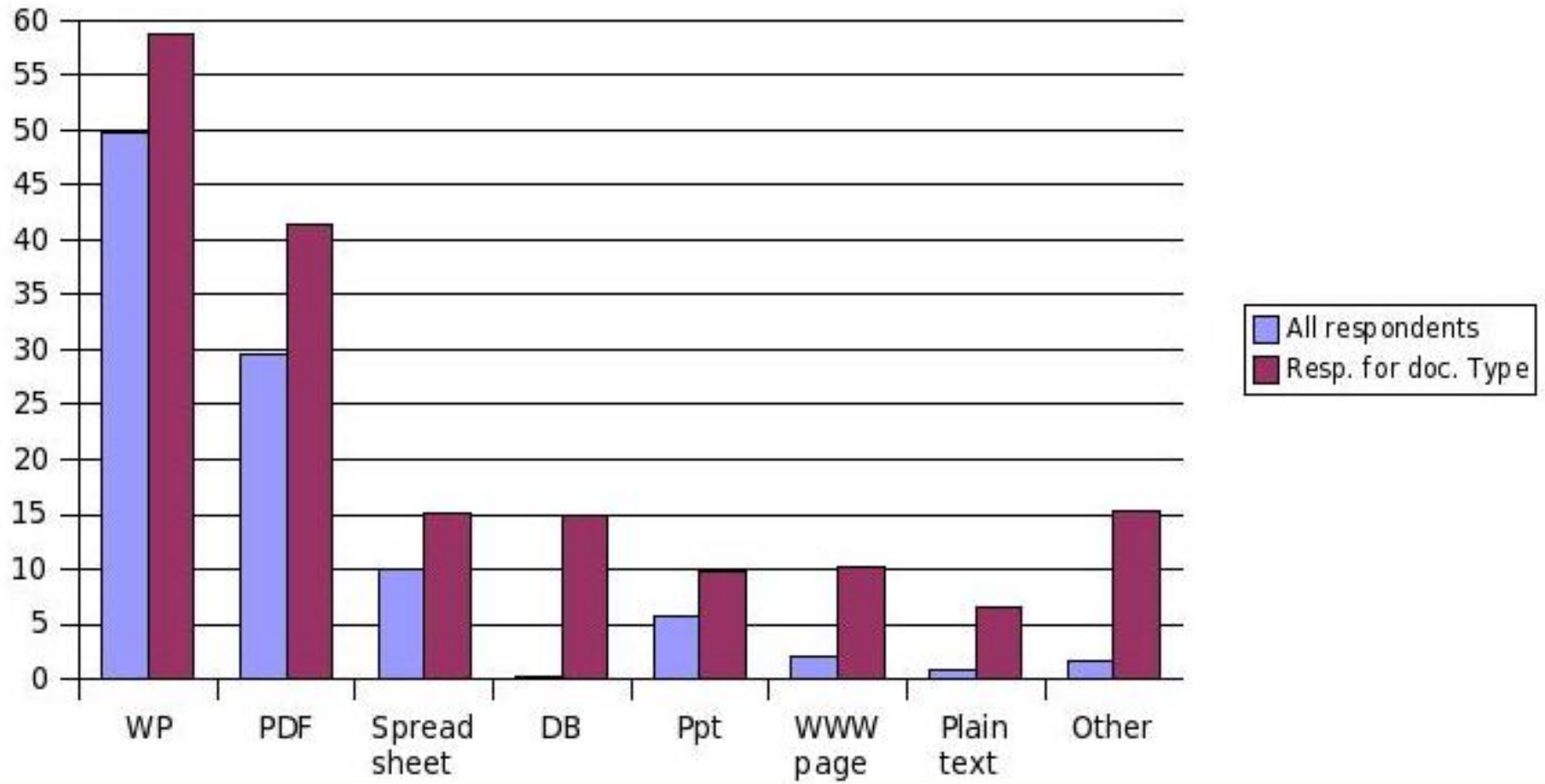
The results from the Translator Survey provide information on the present state of computer usage, and expectations for the future:

- Document Delivery
- Document Types
- Dictionary Usage
- WWW Searching
- Translation Memories
- Machine Translation
- Mail Archives

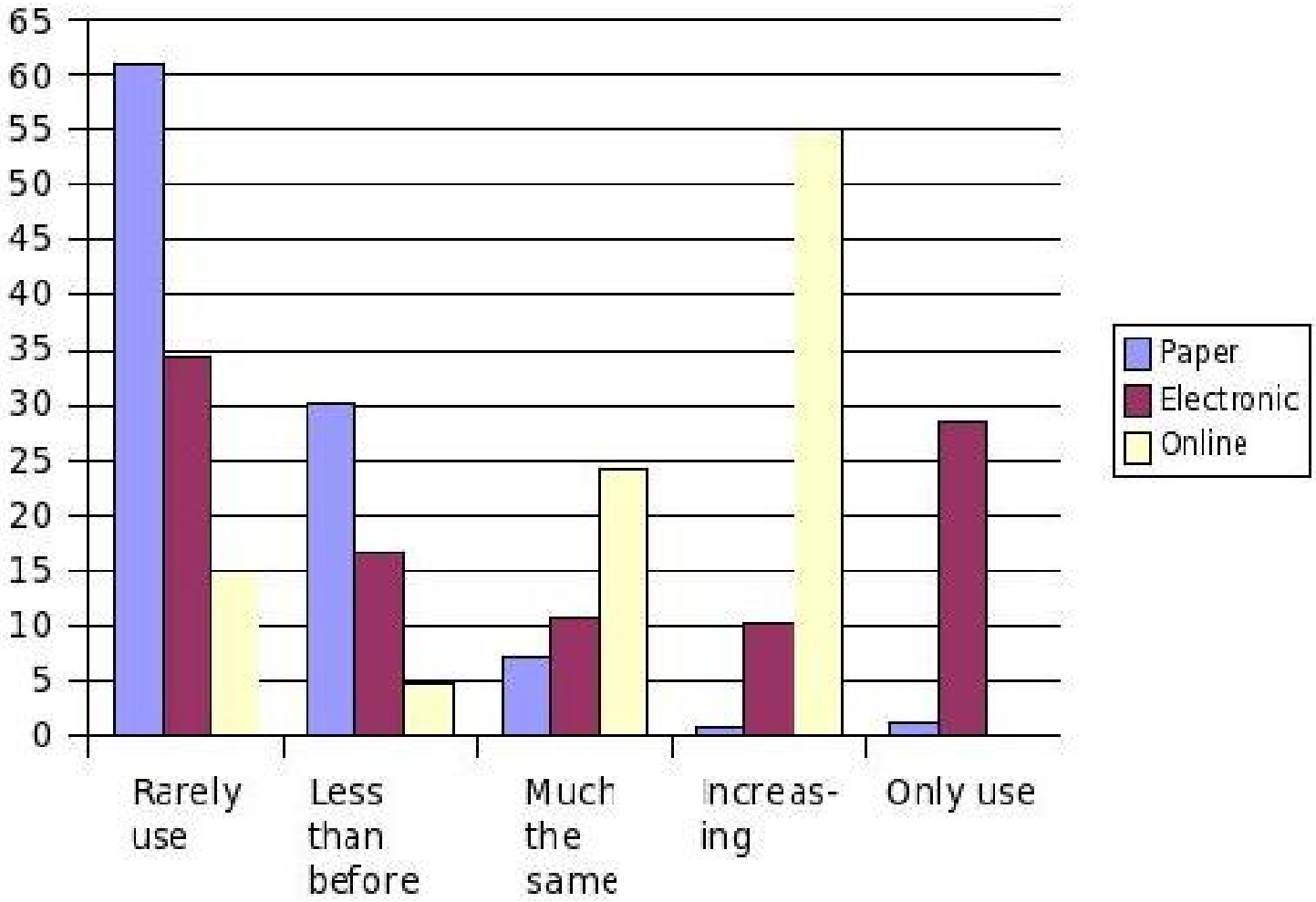
Document Delivery by Type



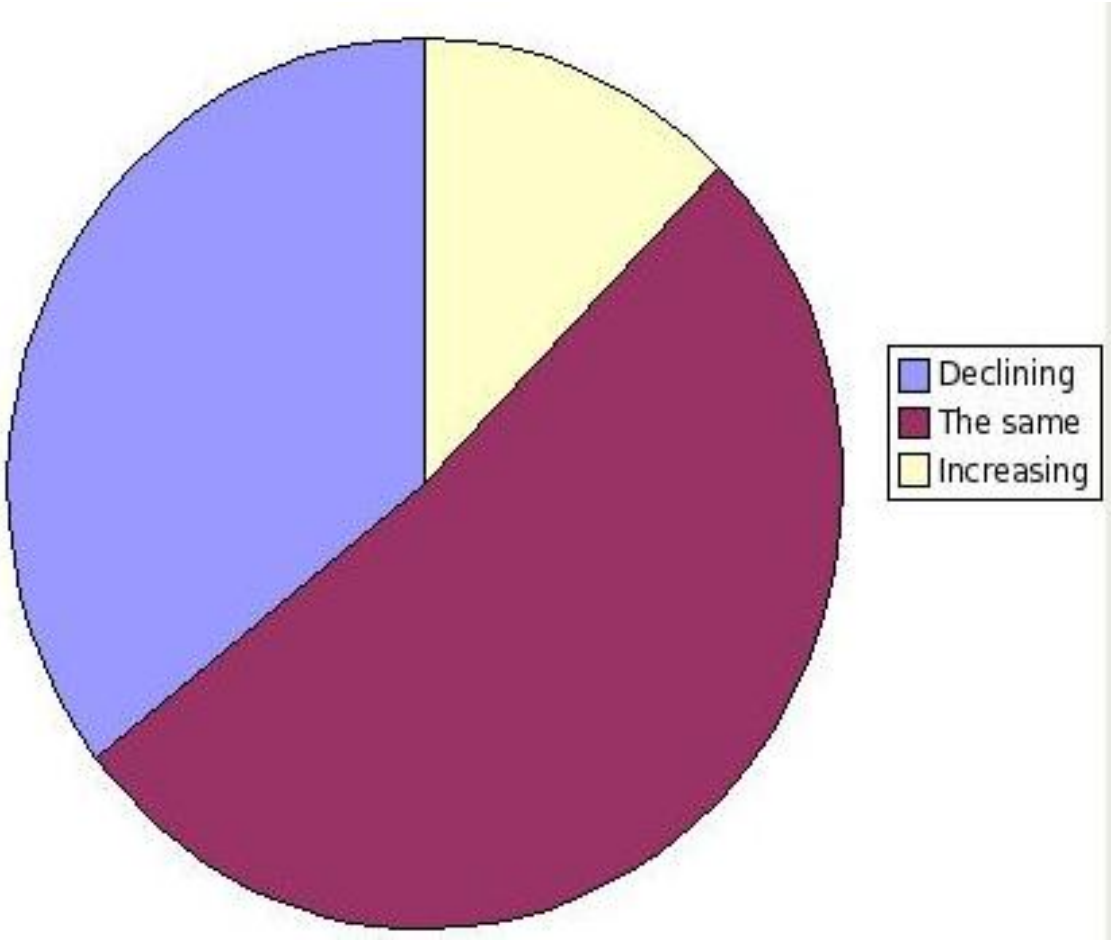
Electronic Document Types



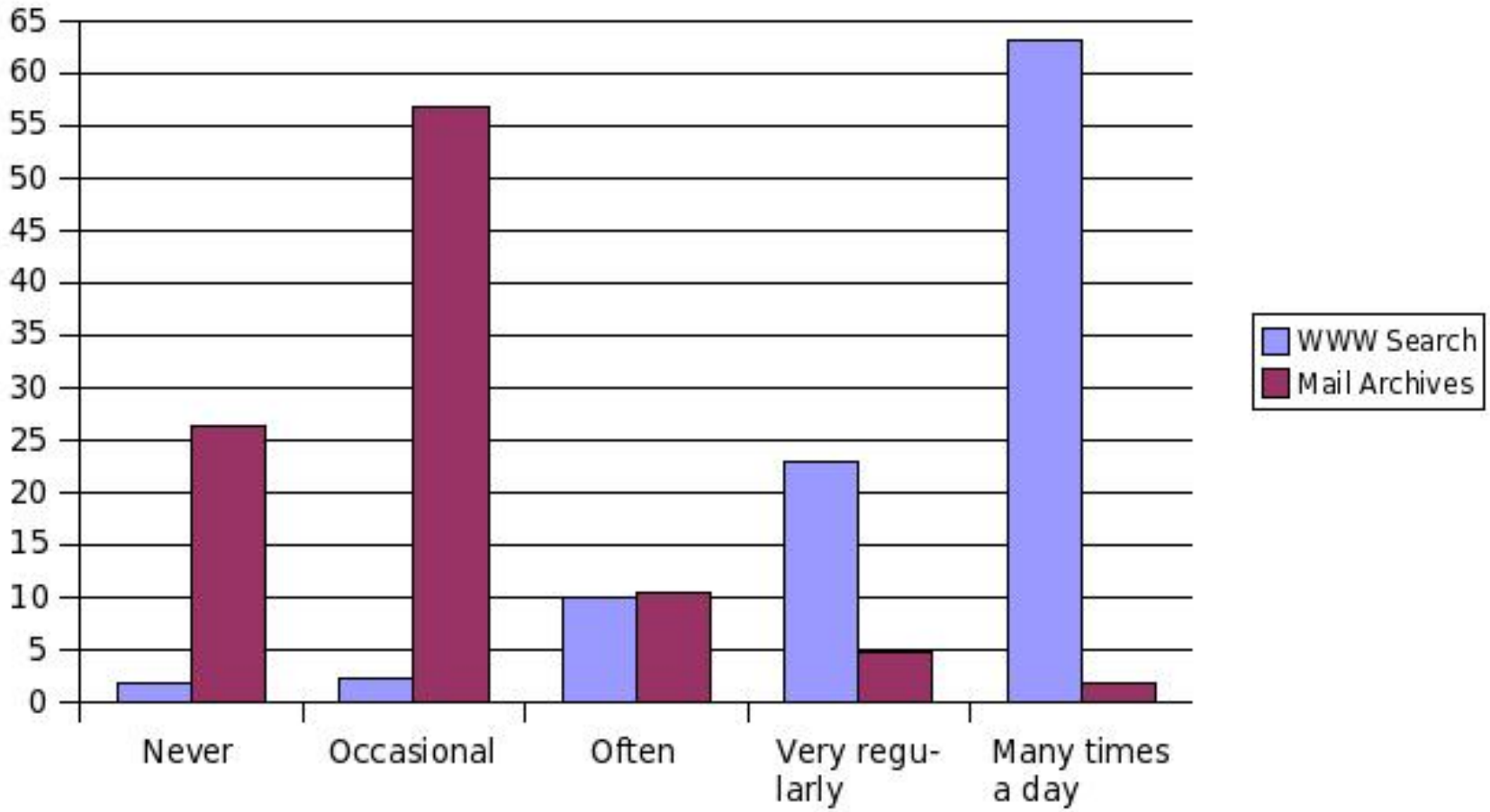
Dictionary Usage by Type



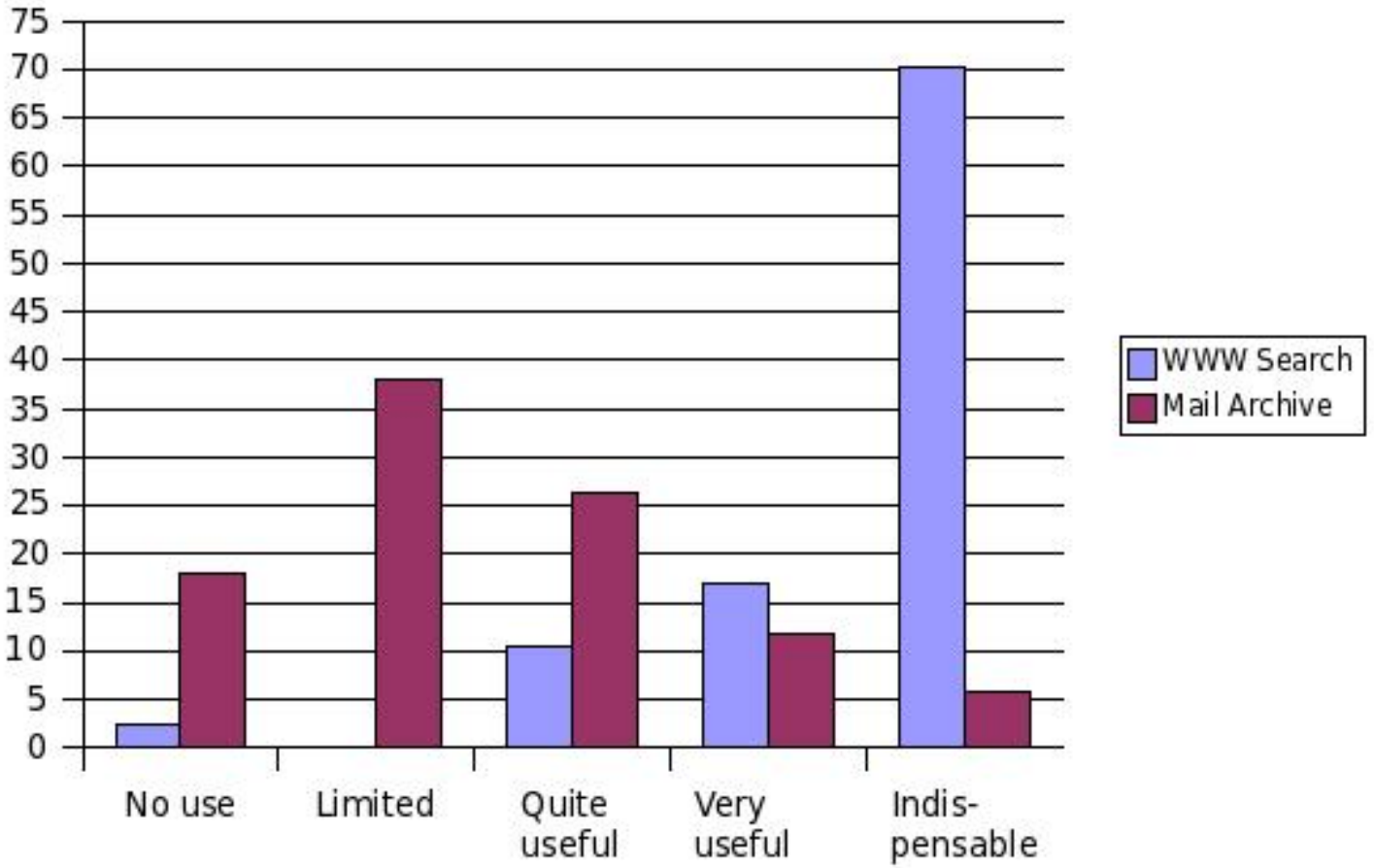
Overall Dictionary Use



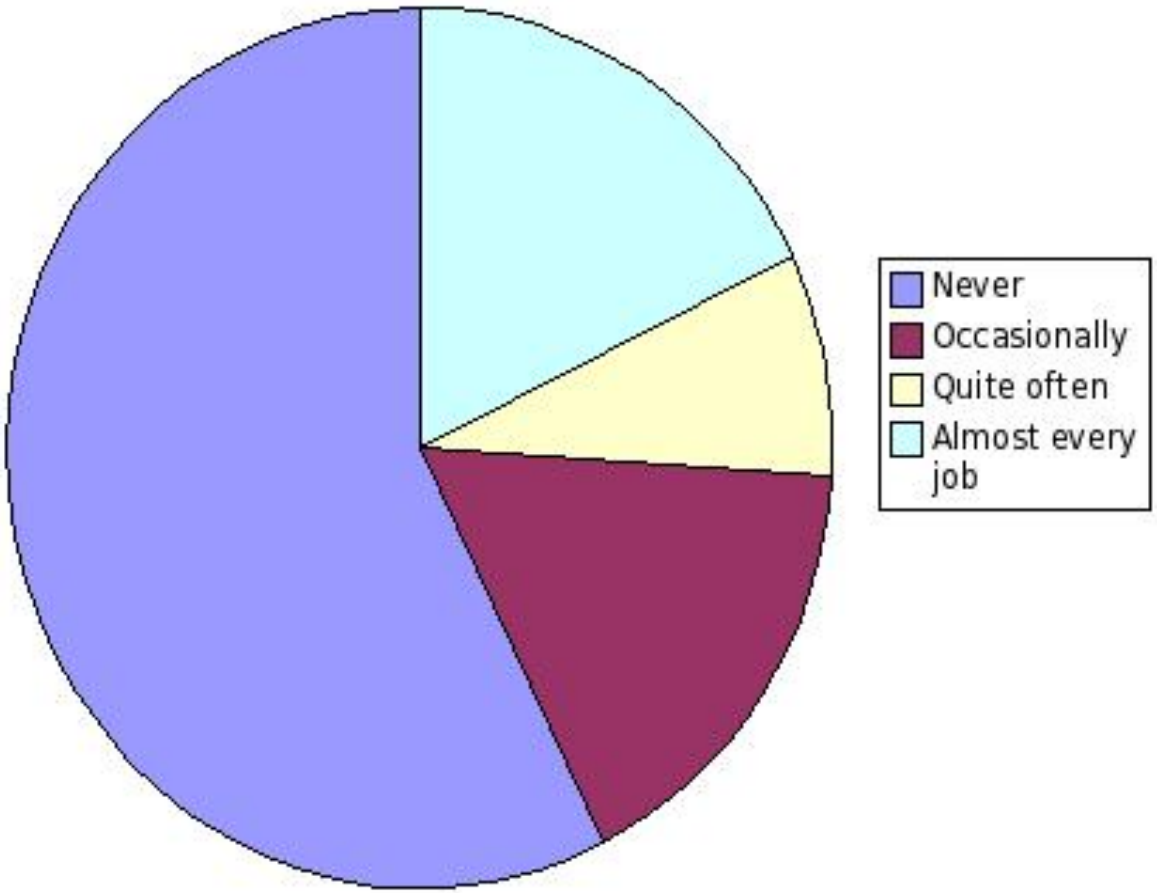
WWW Searching & Mail Archives - Usage



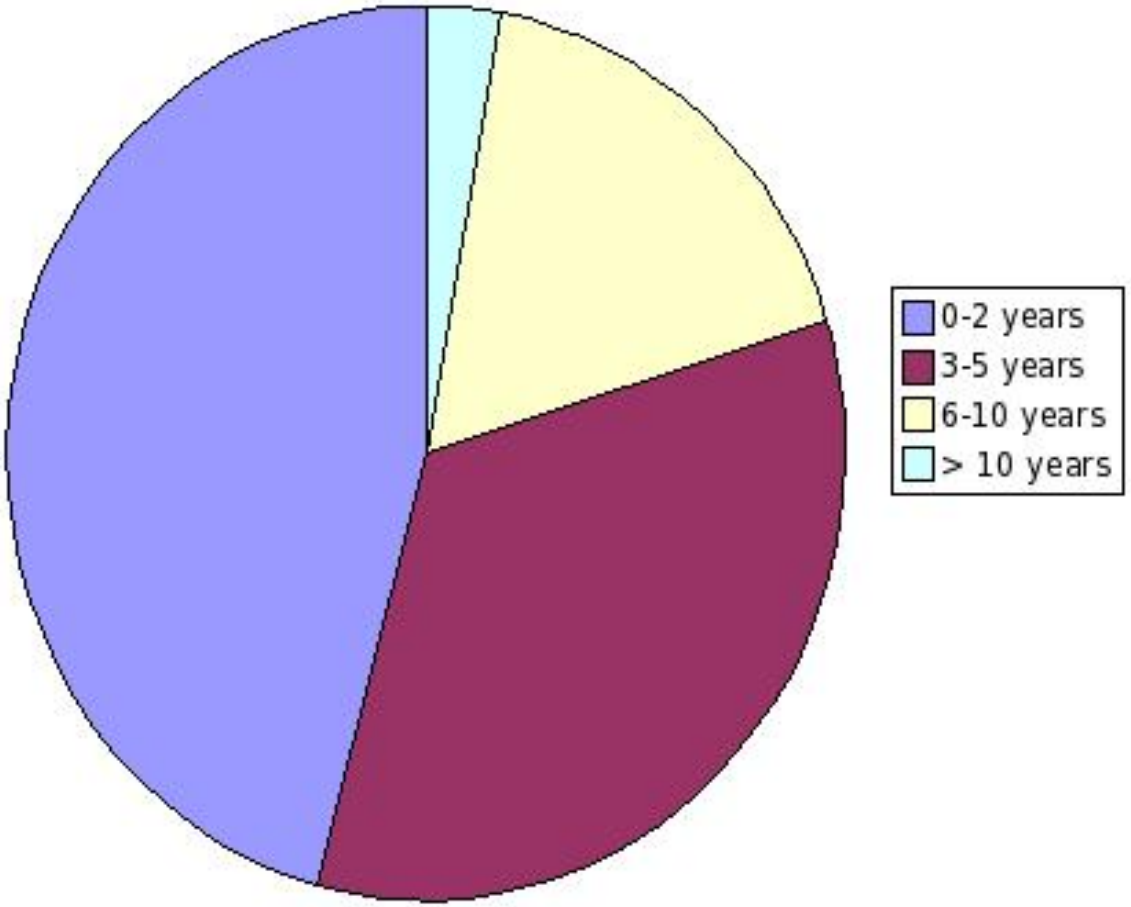
WWW Searching & Mail Archives - Importance



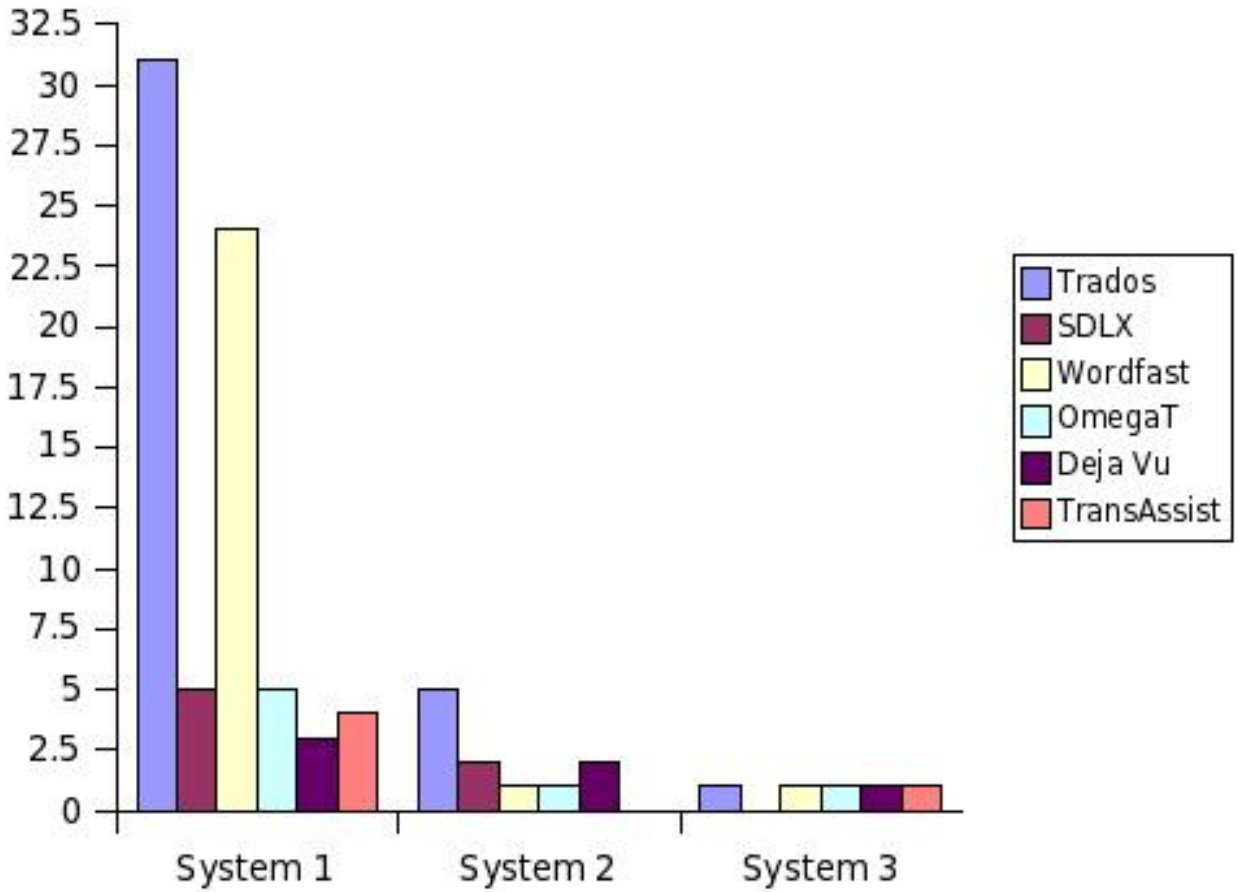
Translation Memory Usage



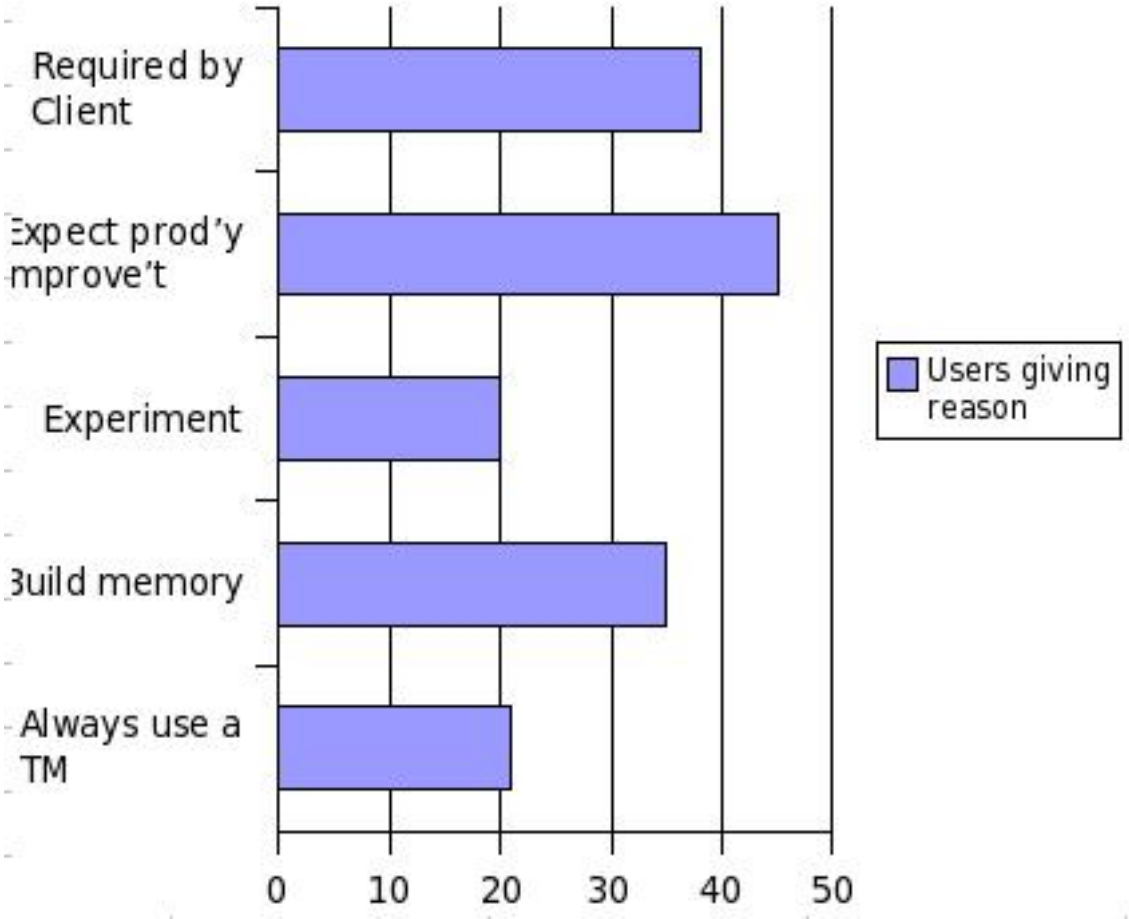
Translation Memory - period of use



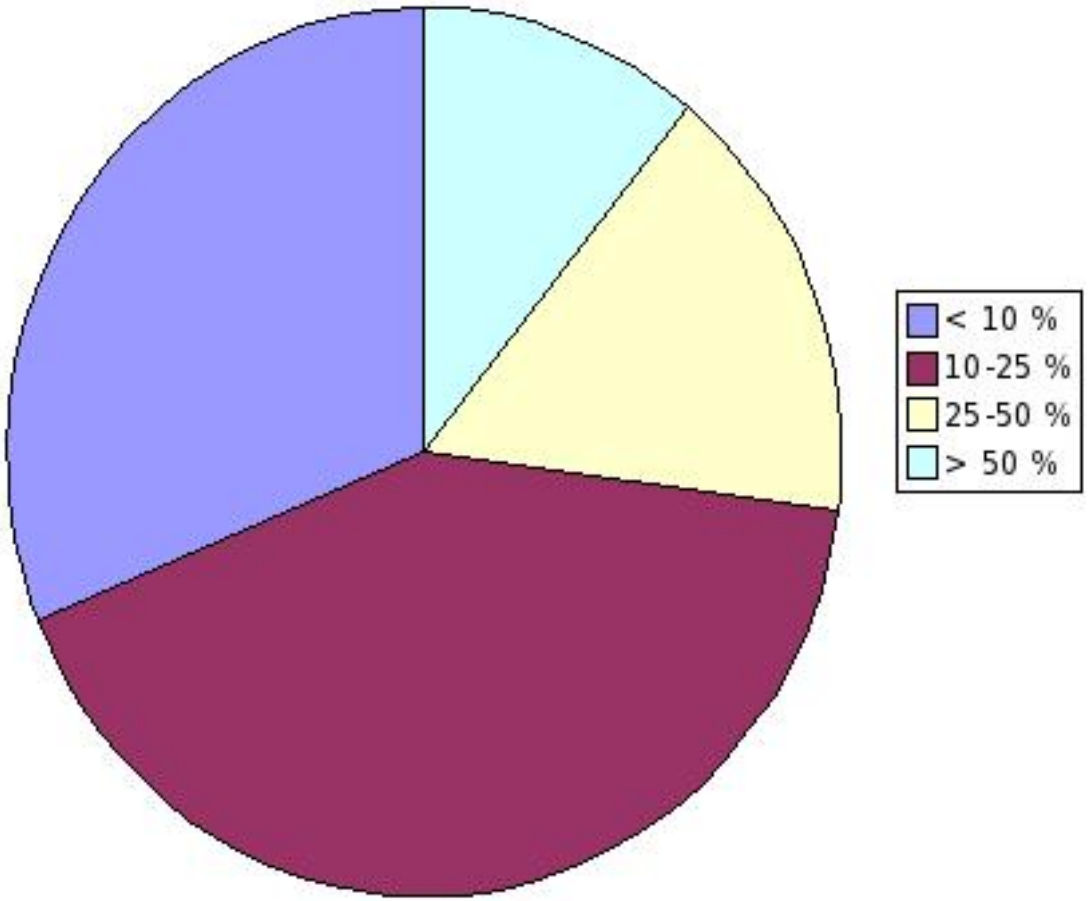
Translation Memory Systems Used



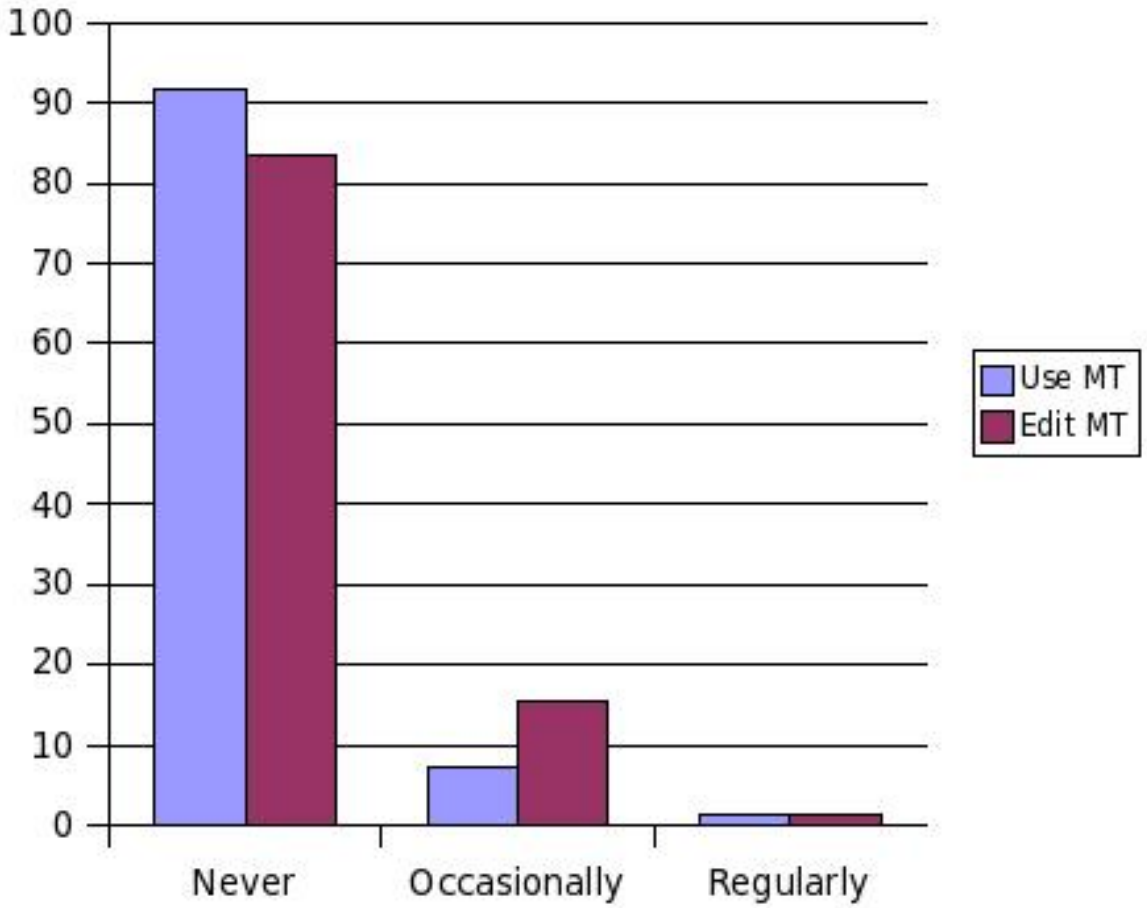
Reasons for Translation Memory Use



Productivity Improvement with Translation Memory



Machine Translation Usage



Machine Translation Quality

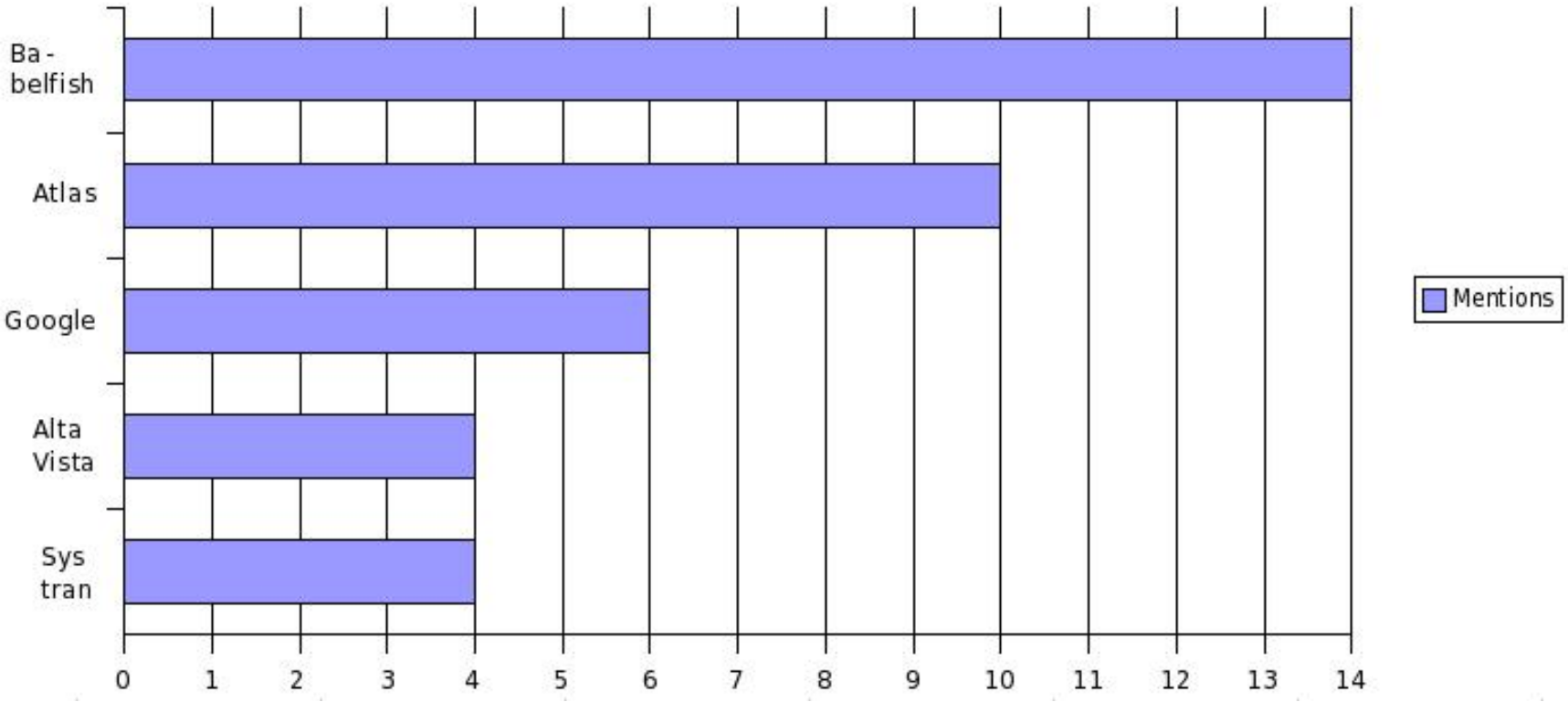
Current Quality

- 57.3% very poor
- 12.9% marginal

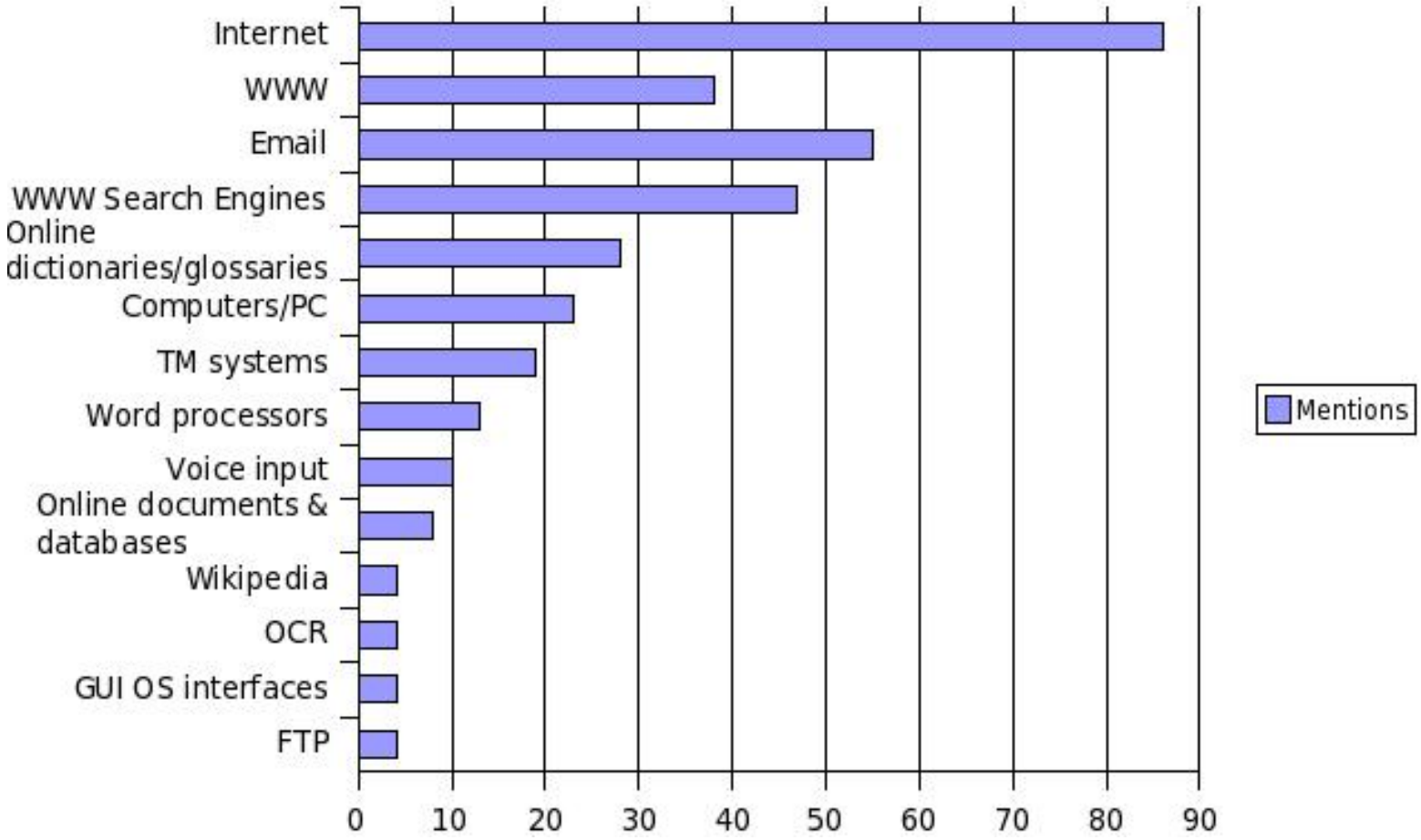
Any recent improvement?

- 27.5% no sign
- 27.5% slight improvement

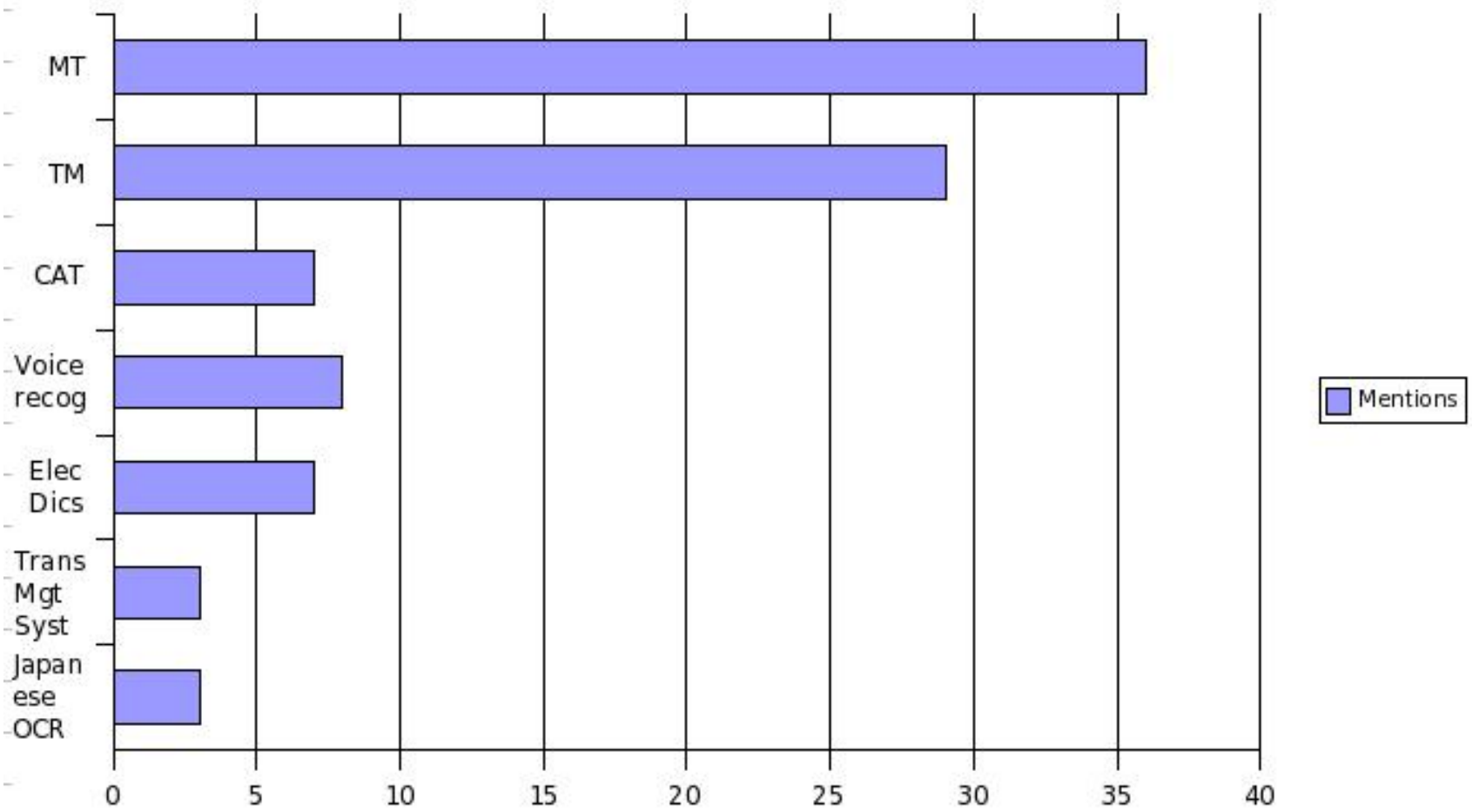
MT - Awareness of Systems



Technology having Impact



Technology That Disappointed



General Observations

- Clearly significant changes in the way translators work
- Dramatic changes in the "tools of the trade"
- Translators of Japanese appear to be "light" users of TM systems
- Translators now heavily involved in the "presentation" of text

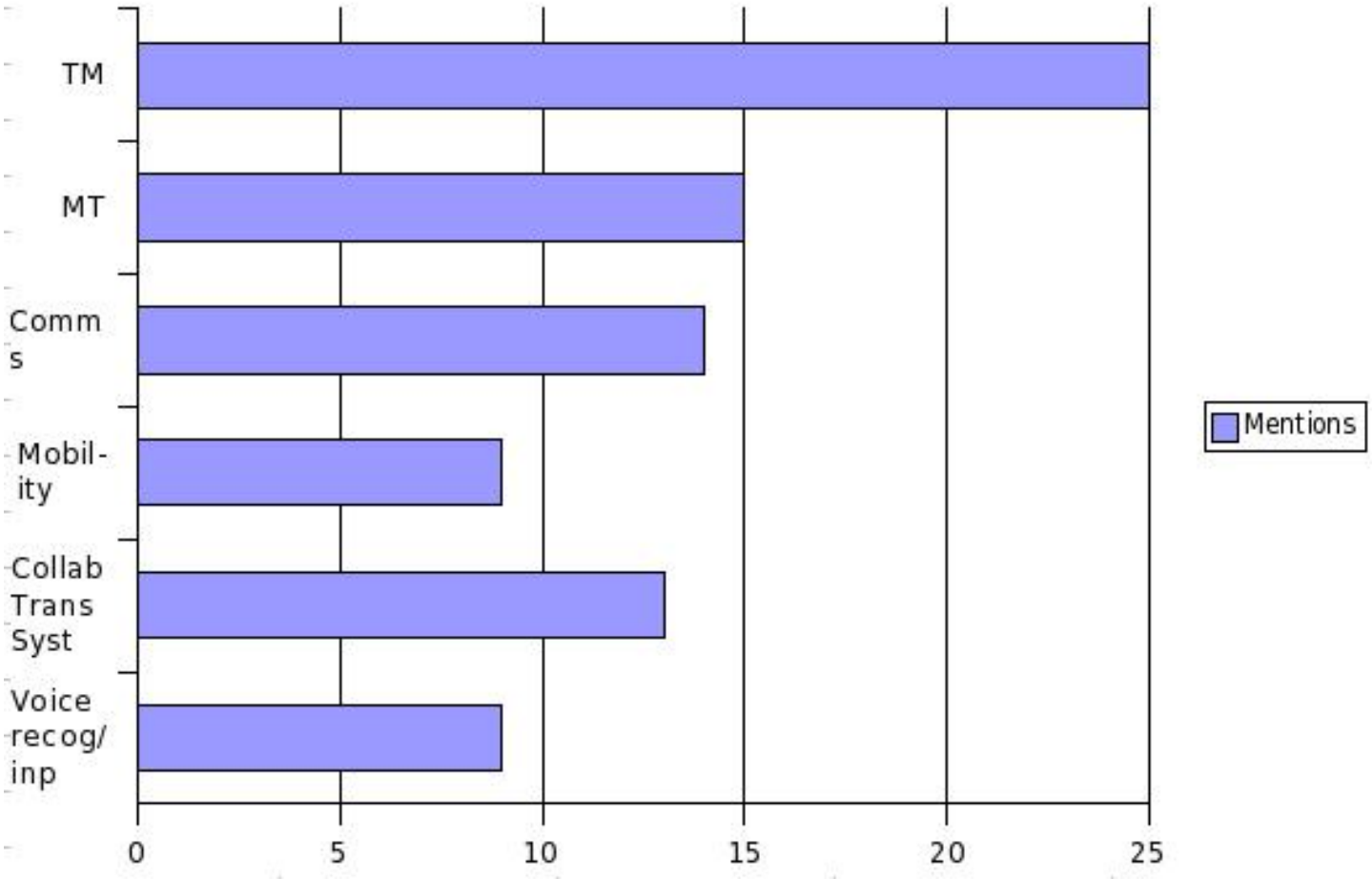
The Future - Introductory Comments

- Predictions about IT developments are usually wrong
- – we tend to extrapolate from the present situation
- – we tend to focus on improving currents systems/techniques
- – IT is very susceptible to paradigm shifts, e.g.
  - the PC: effectively destroyed the "mainframe" industry
  - the Internet: totally overturned traditional networking
  - the WWW: revolutionized the interaction between people, computers and information sources

Introductory Comments (2)

- Underlying technology usually develop incrementally, BUT ....
- – the movement can be rapid (Moore's Law)
- – development/production lead times are relatively short
- – economies of scale in manufacture can be huge
- – end-user capitalization is relatively light
- Developments are largely market-driven, BUT ....
- – translation has been a relatively small market

Survey - Technology with Future Impact



Machine Translation

- MT **IS** getting a lot of use worldwide
  - draft preparation using controlled language (EU, etc.)
  - WWW (Translate this page)
- the smart money is going into statistical translation
- – parallel texts for training is an issue
- a lot of activity in application-specific systems (eg travel industry)
- Google and Microsoft becoming major players

Machine Translation - Predictions

- Steady improvement in Statistical MT quality
- – will probably never reach a "high" quality
- – main application will be in the WWW and restricted domains
- – may even be a "standard" option with Windows, etc.
- Increase of controlled language in business documentation
- – widely used in the EU, and a few multinationals
- – can dramatically improve the quality of MT output
- Arrival of MT-based travel guides
- – already working well in the lab.
- – speech recog. and synthesis

Dictionaries

- The last decade has seen major computerization of dictionaries
  - handheld
  - CDROM/file
  - online
- – still often mirroring the paper originals
- Scope for significant rethink of the dictionary concept (Atkins, 1997)
- – internal hyperlinking
- – customization of user view
- – less language distortion, etc.
- Not a lot of progress
- EPWING/JIS X 4081 good for its day, but book-oriented

Dictionaries - Predictions

- Continued trend from paper to electronic and online
- New more flexible and useful structures
- Integration into other systems:
- – the Desktop
- – TM/CAT systems
- Potential interworking of multiple dictionaries

TM/CAT Systems

- Continued movement to overall CAT systems of which TM is part
- Increase in "2nd gen" TM, using advanced NLP techniques (e.g. Similis)
- Movement to more server-based systems with shared memories and glossaries
- Pressure to share/sell memories and glossaries
- – potential/threat(?) to interwork with Statistical MT

Communications and Mobility

- Expect an even more "networked" future - it hasn't plateaued
- – continued blurring of the voice/data/video boundaries
- – buzzwords: convergence, pervasive, ubiquitous, embedded, seamless, smart devices
- Work-from-anywhere potential even stronger
- Expect the PDA to morph into a powerful multimedia comms/processing tool
- Expect greater integration into clients' systems
- – change in the concept of "freelance"

Envoi

- Computers and related technologies have had a massive impact on translation
- – dictionaries: paper to electronic forms
- – new tools: TM, glossary systems, WWW searching, etc.
- – greater involvement with client documents
- Expect just as massive changes in the future
- Watch out for the **next** paradigm shift ....
- – and the one after that ....
- – and the one after that ....