# IWSLT-06: experiments with commercial systems and lessons for subjective evaluations

Christian Boitet, Youcef Bey, Mutsuko Tomokiyo, Wenjie Cao, Hervé Blanchon

# Motivations

- Participate to this CSTAR initiative, although we don't work on CE, JE, IE, AE
  - Work for the subjective evaluation (CE)
  - Run some commercial (hand-crafted) systems after "tuning" them to the campaign (user dictionaries)
- Study interesting questions/hypothesis
  - Can commercial wide-coverage text-MT systems be used for speech-MT?
  - Is it true that the subjective evaluation can be made less expensive by changing its setting ?
  - How does the set of reference translations influence the evaluation scores produced by BLEU, NIST, …?

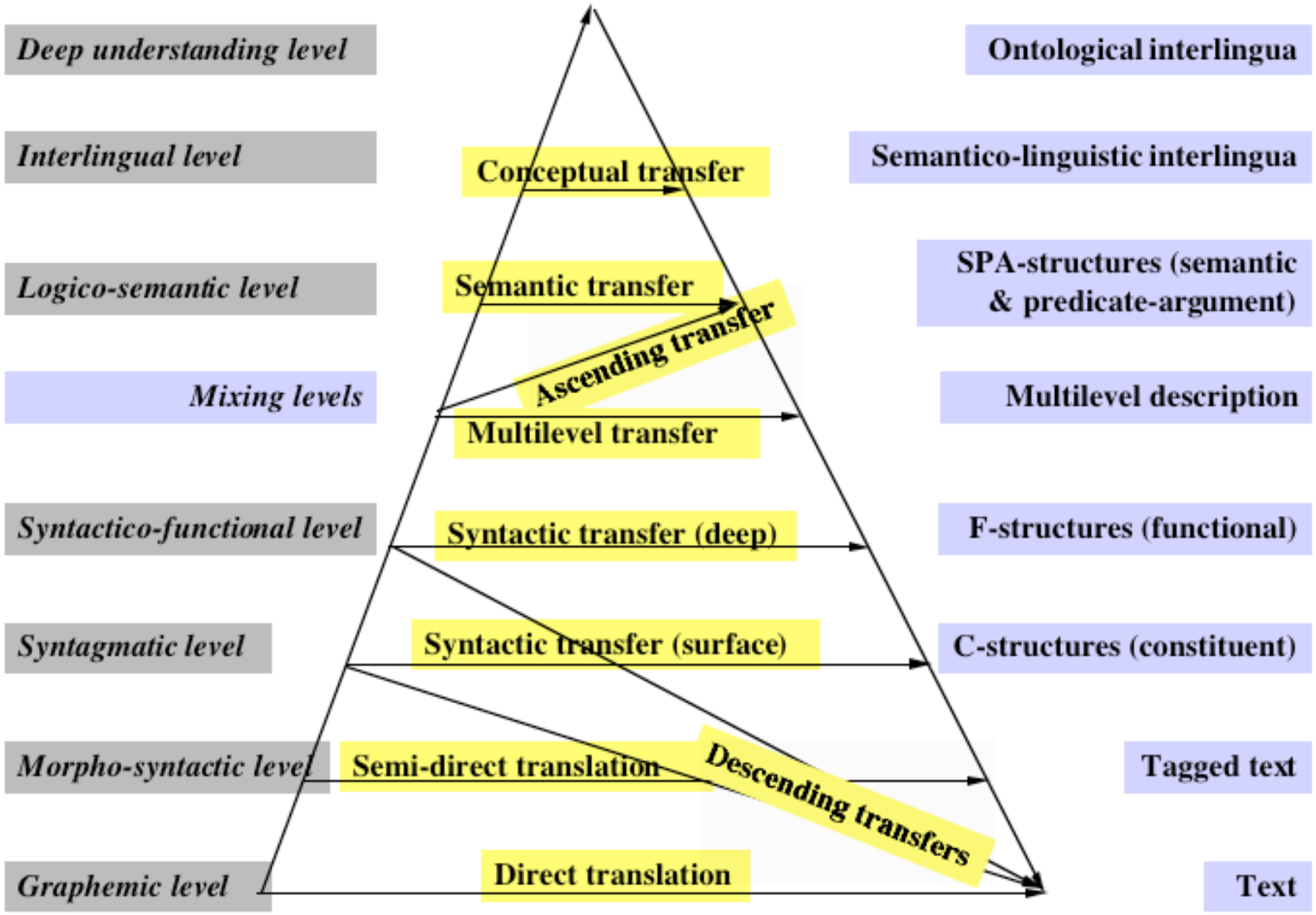
# Some commercial systems

- At least 6 JE/EJ on the shelves at Akihabara
  - Fujitsu (ATLAS v13), IBM (honyaku no o sama v5), Toshiba (The honyaku), Logovista (Logovista-Pro-2007), TechnoCraft (robofuudo v8.2), CROSS (med-transfer v5, pc-transfer+honyaku-studio)…
- and others in Japan (commercial or in-house)
  - Sharp, Oki (Pensée), NTT (ALT/JE, ALT/Flash), CSK (?), Hitachi (HICAT)…
- Elsewhere
  - West: Systran (35 pairs, building more), Softissimo (Reverso), Linguatec/Lingenio (PC-translator, based on LMT), WordMagic, Comprendium (based on METAL)…
  - East: many CE/EC systems, notably Xiamen (Néon, Pr. Shi)…
- For more, see the Compendium of MT systems
  - EAMT, J. Hutchins

# Types of MT systems

- One should distinguish between
  - OBJECTS (intermediate representations)
    - linguistic architecture (see Vauquois' triangle)
  - PROCESSES (how to compute them)
    - computational architecture
- PROCESSES can be
  - basically HAND-CRAFTED
    - RBMT, KBMT
    - ± corpus-induced data (terminology, phraseology)
  - basically MACHINE-LEARNED from // corpora
    - SMT, P-SMT, EBMT
    - corpus with ± deep linguistic annotations (seg.→sem.)
  - BOTH: e.g. Microsoft MTS (transfer only learned)
- ≈all commercial MT systems are basically hand-crafted
  - ==> interesting to compare with machine-induced MT

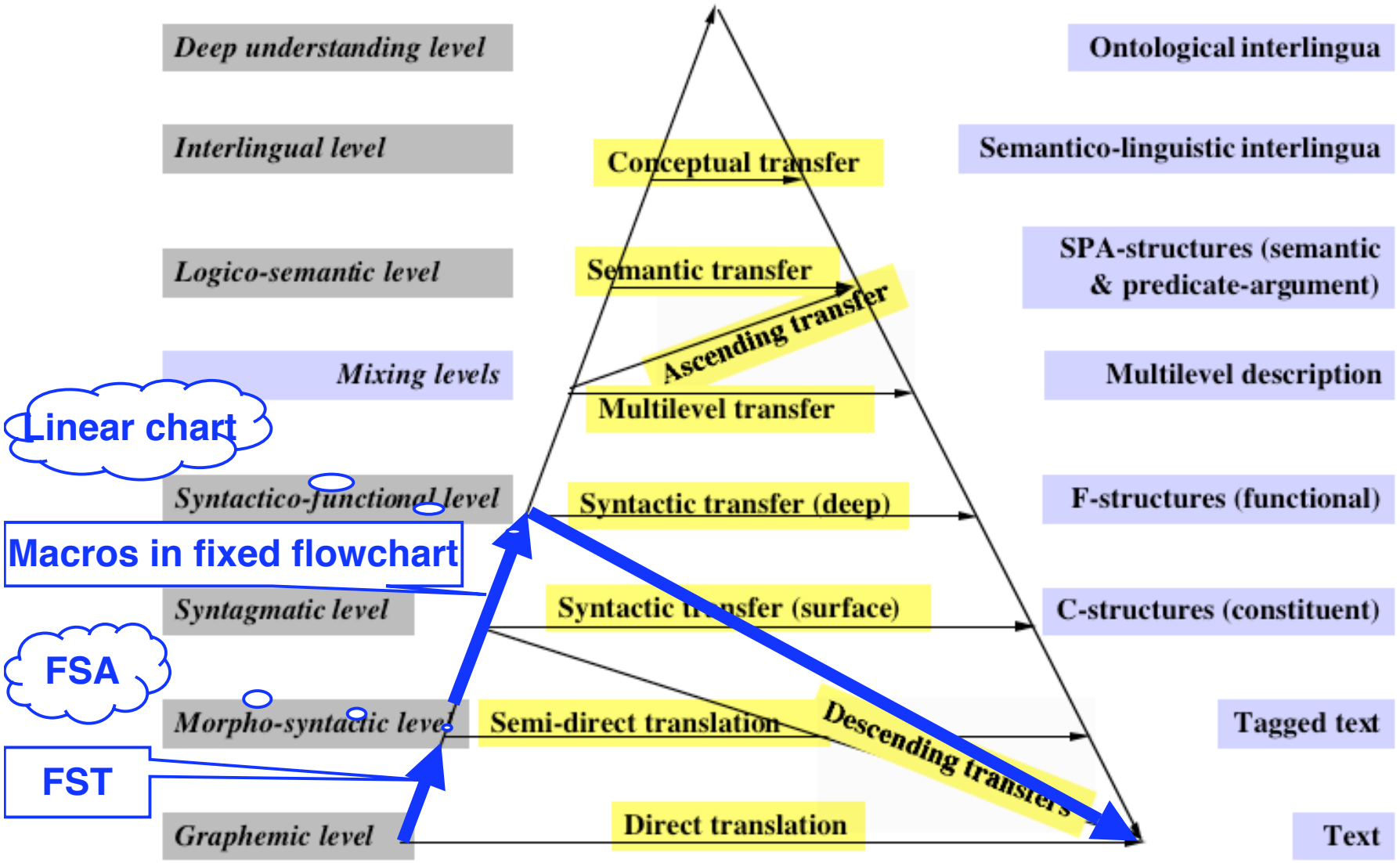
# Vauquois's triangle



# Systems run for IWSLT-04

| Pair | System(s) | Tuning (on training only) | Linguistic analysis |
|---|---|---|---|
| AE | Systran-5 | - | - |
| CE | Systran-5 | User dict. | - |
| IE | Systran-5 | User dict. (50%) | - |
| JE | Systran-5 | - | Tomokiyo |
| JE | ATLAS-2 | - | Tomokiyo |

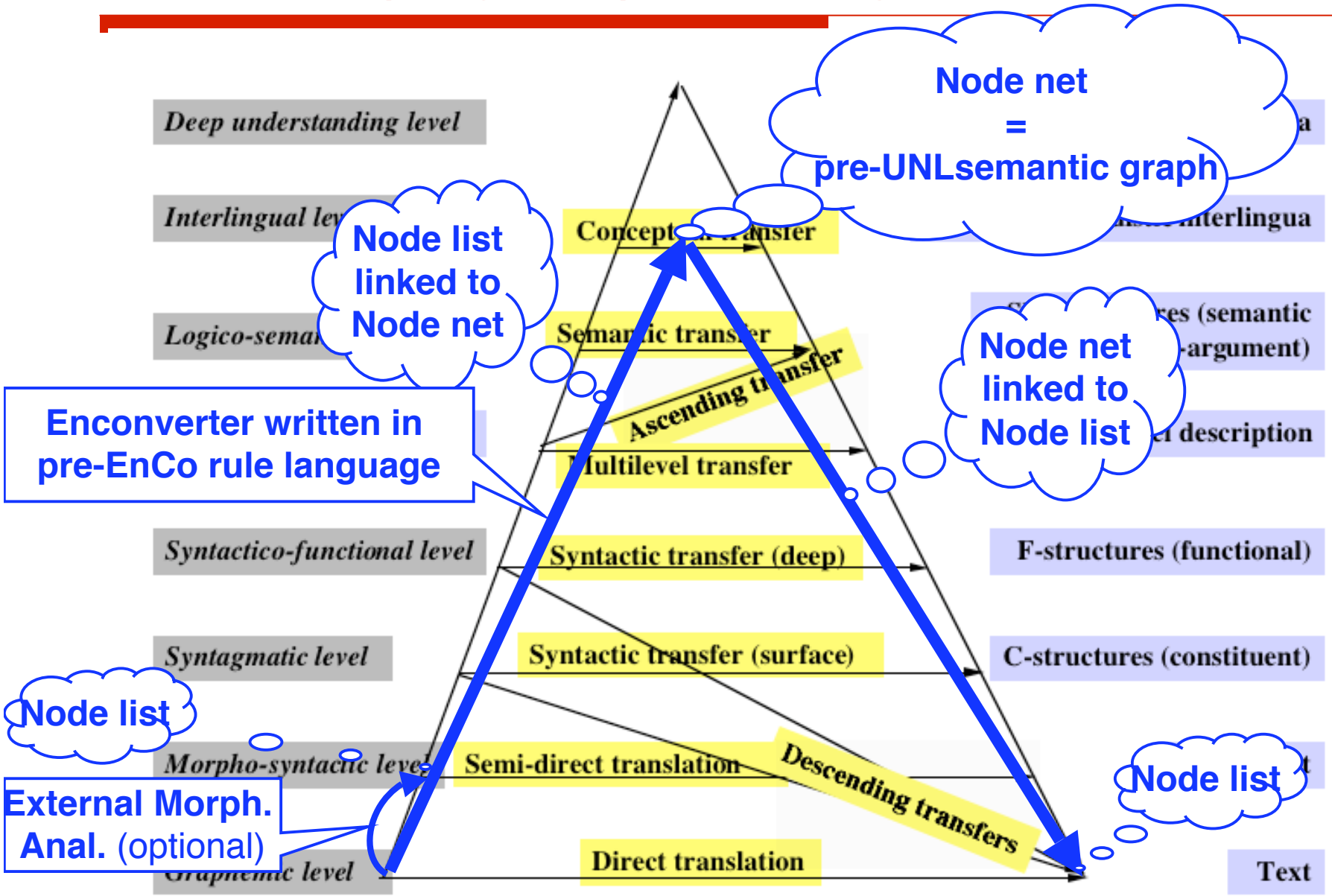
# SYSTRAN v5 — diagram



# SYSTRAN v5 — text

- Descending transfer
  - Morpho-syntactic analysis (MA)
    - FST used since ≈1996 (from Paris VII, Gross & al.)
    - Output is a wFSA
    - 1 "solution" (trajectory) chosen
  - Syntactic analysis (SA)
    - 1 variant per target language (≠ decisions)
    - "rules" in fixed procedural framework (C macros)
    - works on a kind a linear "chart" but not "chart parsing"
    - deterministic: no back-track, one-path.
  - Transfer+generation (T+G)
    - also procedural ("rules"), one-path
- Modern features
  - XML-based workflow
  - Interactive disambiguation of wFSA possible (since v5)

# ATLAS (Fujitsu) — diagram



# ATLAS (Fujitsu) — text

- Interlingual pivot *(pre-UNL, by same author, H. Uchida)*
  - Enconversion
    - 1 integrated component written in a *rule language*
      - (ancestor of current EnCo, see UNDL web site, book)
  - Deconversion
    - 1 integrated component written in a *rule language*
      - (ancestor of current DeCo, see UNDL web site, book)
- Heuristic programming, but low-level *(large # of rules necessary)*
  - non-deterministic: depth-first, back-tracking, one result only
  - variables/features: boolean only (∪ ∩ ⊇ ⊆ unavailable)
- Impressive dictionary size
  - 586,000 entries at MTS-01, 1.5M entries at ACL-03
  - **>5,440,000 entries now!** Our version: 950,000 entries
  - Used corpus-based techniques to multiply dictionary size
- Very good integration *(translation memory, editor…)*
- (One of the) best text MT system(s) in Japan for > 20 years

# Tuning done on Systran

- (CE, IE, JE, AE)
  - preprocessing
    - encoding
    - separating the ids from the text
  - choice of batch parameters
    - choice and priority ordering of dictionaries *(user, general)*
    - handling of capitalized words *(don't translate)*
    - handling of not found words *(NFW)*
    - presentation if multiple lexical translations *(1 only)*
  - building a user dictionary
    - CE: 97%
      - 400 NFW in training corpus *(dev forgotten!)*
      - 12 NFW in test corpus
    - IE: ≈50%
      - 1200 NFW in training corpus *(dev forgotten!)*
      - ≈30 NFW in test corpus
    - JE, AE: none

# Tuning done on ATLAS

- (JE only)
  - preprocessing
    - encoding
    - separating the ids from the text
  - postprocessing
    - removing of annotations and NFW marks
  - we did not build/use
    - user dictionary (no time)
    - translation memory (did not know how!)

# Remarks on IE training corpus

- At some places, there seem to be

  English chunks instead of their Italian translations

| | |
|---|---|
| IE_TRAIN_12108\Sì, abbiamo la Where, and The City Guide. | IE_TRAIN_12108\Yes, we have the Where, and The City Guide. |
| IE_TRAIN_01045\Congratualzioni, Henry. Sono felice di sentire del Suo fidanzamento con Jane. | IE_TRAIN_01045\Congratulations, Henry. I'm delighted to hear of your engagement to Jane. |
| IE_TRAIN_01049\Deve essere stato un grande shock per Lei. | IE_TRAIN_01049\It must have been a great shock to you. |
| IE_TRAIN_01726\Potrebbe pagare alla reception, prego? | IE_TRAIN_01726\Could you pay at the front desk, please? |
| IE_TRAIN_02516\Sono contento di averLa conosciuta. Grazie. | IE_TRAIN_02516\I'm glad I met you. Thank you. |
| IE_TRAIN_06501\Qui parla l'operatore dell' International Telephone Call Service. | IE_TRAIN_06501\This is the operator for International Telephone Call Service. |
| IE_TRAIN_09747\Facendo lo spelling è G-O-R-O-H. | IE_TRAIN_09747\It's spelled G-O-R-O-H. |

# Objective evaluation (Systran)

| | official (with case + punctuation) | | | | |
|---|---|---|---|---|---|
| | **BLEU4** | **NIST** | **METEOR** | **WER** | **PER** |
| **Spontaneous speech** | 0.0344 | 2.7374 | 0.3178 | 0.87129 | 0.743063 |

| | additional (without case + punctuation) | | | | |
|---|---|---|---|---|---|
| | **BLEU4** | **NIST** | **METEOR** | **WER** | **PER** |
| **Spontaneous speech** | 0.0406 | 2.8625 | 0.3184 | 0.880529 | 0.720287 |

| | official (with case + punctuation) | | | | |
|---|---|---|---|---|---|
| | **BLEU4** | **NIST** | **METEOR** | **WER** | **PER** |
| **Read Speech** | 0.0536 | 3.7390 | 0.3210 | 0.805919 | 0.687017 |

| | official (with case + punctuation) | | | | |
|---|---|---|---|---|---|
| | **BLEU4** | **NIST** | **METEOR** | **WER** | **PER** |
| **CRR** | 0.0366 | 2.685 | 0.3178 | 0.858339 | 0.726484 |

| | additional (without case + punctuation) | | | | |
|---|---|---|---|---|---|
| | **BLEU4** | **NIST** | **METEOR** | **WER** | **PER** |
| **CRR** | 0.0749 | 4.4256 | 0.3694 | 0.780118 | 0.643764 |

# Chinese segmentation problems

- (Systran)

| | |
|---|---|
| 对 历 史 ( h i s t o r y ) 感 兴 趣 ( to be interested) | be interested in history |
| 职 员 ( e m p l o y e r ) 会 ( c a n ) 轮 流 放 假 | employee can take several days off by turns |
| 艾 凡 斯 顿 | Evanston |
| 我 就 要 替 你 喝 完 秋 葵 荚 汤 ( s o u p e ) 了 。 | gumbo |
| 雕 塑 ( s c u l p t u r e ) 感 兴 趣 ( b e interested) | interested in sculpture |
| 孟 斐 斯 ( p r o p e r n a m e ) | Memphis |
| 理 察 德 ( R i c h a r d ) 波 尔 曼 | Richard Paulman |

# Additional translation runs

- Remember Systran is "bad" for CE, JE, AE
  - insufficient investment
  - no tuning at all on spoken utterances (quite ≠ !)

(i) Read speech: J-E translation by SYSTRAN

| | | BLEU | NIST |
|---|---|---|---|
| **SYSTRAN** | **ASR output (Read speech)** | 0.0755 | 3.7685 |

(ii) Read speech: J-E translation by ATLAS

| | | BLEU | NIST |
|---|---|---|---|
| **ATLAS** | **ASR output (Read speech)** | 0.1084 | 4.4295 |

(iii) Read speech: A-E translation by SYSTRAN

| | | BLEU | NIST |
|---|---|---|---|
| **SYSTRAN** | **ASR output (Read speech)** | 0.049 | 3.6202 |

(iv) Read speech: I-E translation by SYSTRAN

| | | BLEU | NIST |
|---|---|---|---|
| **SYSTRAN** | **ASR output (Read speech)** | 0.1368 | 5.1528 |

# Types & sources of errors (Systran-JE)

## ☐ Synthesis

| | |
|---|---|
| When the utterance is **euphemistic** (が), the particle is always translated by "but" | |
| Some of the utterances do not make sense without **context** | 切りますよ。 ➟ "it cuts" ? |
| When the **first person subject** is **omitted in Japanese**, it is always translated as "it" | ここで降ります。<br>➟ "It gets off here." |
| **Interrogative pronouns and adverbs are always** (incorrectly) **shifted at the end of the translation** | オペラ座はどこですか。<br>➟ "Is the opera house where?" |
| Many daily life **idiomatic expressions** are not contained in the SYSTRAN dictionaries | どういたしまして。 ➟ "How doing."<br>もしもし。 ➟ "It does."<br>さようなら。 ➟ "Way if." |
| **Requests or invitations** are not always well translated | 注文したいのです。<br>➟ "It is to like to order."<br>一緒に行きましょう。<br>➟ "It will go together." |
| **When the valency of the verb for two expressions in Japanese and English is different, the translation is almost always wrong** | 寒気がする。 ➟ "Chill does." |
| **Aspect** of Japanese predicates is not correctly rendered in English | 航空券を家に忘れて しまいました。<br>➟ "The air ticket was forgotten in the house." |
| **Positive point:**<br>**lexicalized Japanese politeness is correctly handled** | そのまま切らずにお待ち下さい。<br>➟ "Without cutting that way, please wait." |

# Types & sources of errors (ATLAS-JE)

## ☐ Effects of segmentation errors

| | | |
|---|---|---|
| 申し訳ありません 離陸 し て から でないと テレビ を 御 使い 頂け ません. | The television cannot be ~~had to be~~ used after the take off ~~which apologizes and not is.~~ | The turn is composed of 3 turns, but ATLAS has translated it as two turns with a relative clause” |
| これ は 無鉛 で は あり ません ね が ご 希望 なら 御 取り替え 致し ます. | If sleep which is not no lead is hope, I will change this. | The turn is composed of 3 turns ”これ は 無鉛 で は あり ません ね”, ”ご 希望 なら” and ”御 取り替え 致し ます”, but ATLAS has translated it as two turns with a relative clause, because the sentence final particle ”ね” is not recognized. |

Table 1: Segmentation errors (ATLAS JE)

# Types & sources of errors (ATLAS-JE)

## ☐ Not handled spoken language phenomena

| | | |
|---|---|---|
| 申し訳ありません 離陸 し て から **でないと** テレビ を 御 使い 頂け ません. | The television cannot be had to be used after the take off which apologizes and not is. | Verb “でないと” |
| やっ てみ ます が から ぞ 予約 できる か 保証 し 兼ね ます. | Whether からぞ can be reserved cannot be guaranteed やってみます. | Verb “やる” |
| えーっと それ は 六 百 円 です. | Food っとそれ is 600 yen. | Phatic “えーっと” |
| 以前 は 野球 を する の が 好き でし た でも 今 は スキー の 方 が 興味 が あり ます | It was liked to play baseball and skiing is interesting yet now before. | Conjunction “でも” |
| 切っ て 今 手 が ございます どうぞ ご覧 下さい. | (*S) cuts (*O), and there is a hand now and (*S) sees please. | Polite expression |
| 結構 です けど ね でき ます. | (*S) sleeps though it is excellent. | Modal particle “ね” |
| ドイツ語 の が ある と 一 番 良い の です が 英語 は 読め ない の です. | English cannot be read as German がある though it is the best. | Referential noun “の” |
| はい 洗濯機 の 着席 優しく 払わ なければなりません ので ご 注意 下さい. | Please <払 わなければなり> note (*O) <sit-down> nice of the tile washing machine. | Modal expression “なければなりません” |
| 通常 一 週間 です でも 天気 が 悪い わ えー 少し 遅れる こと も あり ます. | The weather for one usual week it yet might be late of <badness> いわえ least | Phatic “えー” |
| かしこまりました 少々 御 待ち 下さい. | Please wait a little standing on ceremony. | Polite expression “かしこまりました” and Honorific expression “御” |
| 陶器御茶の方御酒を買いましたこれらは全てねでーです. | These by which person 御酒 of earthen 御茶 is bought are all sleeps. | Honorific expression “御” |
| そうですね あと 一 時間 位 で 着陸 し ます. | (*S) <aspect> has, (*S) sleeps, and (*S) will land in about another hour. | ‘’ね’’ in dialogues |
| 御 客 様 こちら です 口頭 その ビル の 男性 の 角 に ございます. | It is in the corner of the man in guest こちらです oral その building | Deictic expression ‘’こちら’’ |

# Types & sources of errors (ATLAS-JE)

## Problems coming from the dictionary

### 850,000 entries still not enough (maybe 4.5 M are?)

| | | |
|---|---|---|
| 赤青緑黄色がございますどの色が御好みですか. | Which color with 赤青緑 yellow is favor? | Special words "赤,青,緑" |
| いいえ その ドア を 出 て から 右 に 曲がら なければなりません. | It is necessary to turn right after (*S) goes out of the door of いいえそ. | Deictic and anaphoric word Mots déictique et anaphorique "その" |
| こんにちは 御 客 様 の フライトナンバー と 宿泊 を 取る 名前 を 書い て下さい | The name by which the flight-number and staying of 御客様 hello are taken | Honorific word "御 客 様" |
| ラジオ の 電源 スイッチ は 一人 が ですし の つまみ は 音量 を 調節 する 為 の 物 です. | つまみ of <one person> ですし is a thing to adjust the volume. the power supply switch of the radio | Special word "つまみ" |
| 御 客 様 もう しばらく 御 待ち 下さい 一 週間 以内 に は 御 返事 差し上げ ます. | Guest もうしばらく is waited and I present the answer within one week. | Special word "もうしばらく" |
| . 御 客 様 こちら です 口頭 その ビル の 男性 の 角 に ございます. | It is in the corner of the man in guest こちらです oral その building | "御 客 様"<br>Deictic word "こちら" |
| あちらの大きな連中は記念ように保存されています. | A big party there is preserved in the commemoration way. | Deictic word "あちら" |
| いいえ まだ です. | いいえまだです. | Special word "いいえ" |
| 一 番 近く の レストラン は 車 で もう 三十分 近く 掛かり ます. | The nearby restaurant hangs in the vicinity for 30 another minutes in the car. | Semantic ambiguity of verb "掛かる" |

# Types & sources of errors (ATLAS-JE)

## ☐ Problems in the input Japanese text

### and consequences on ATLAS translation results

| | | |
|---|---|---|
| 精神は三名ドルほどです. | The soul is about three person dollar. | ? |
| 私の国は中国のりんご君日本です. | My country is apple 君日本 of China. | ? |
| 離陸 を 三十分 以内 に は 昼食 を 御 出し 致し ます. | The take off is served and I will serve lunch within 30 minutes. | 離陸 を → 離陸 後 |
| トイレ は 機内 高校 です ご 案内 致し ます. | It will be a guide of the rest room that an in-flight high school has (*O). | 高校 → 後方 |
| はい クレジットカード を ご 利用 頂け ます し 帰る カード は ビザ マスター アメリカンエクスプレス です. | The yes credit card can be had to be used and the card where (*S) returns is visa | 帰る → 使える |
| はい 車 で 十分 ほど と 頃 に 一つ ございます. | It is a tile car and there is one every about ten minutes. | と 頃 に → のところに |
| こちら です 化粧 品 は 二 階 です え で データ で 上がっ て下さい. | Cosmetics which have (*O) <here> must rise by data in placing by the second floor. | えで　データ → エレベータ |
| やっ てみ ます が から ぞ 予約 できる か 保証 し 兼ね ます. | Whether からぞ can be reserved cannot be guaranteed やってみます. | から ぞ → 必ず |
| 申し訳ありません 今 の 所 に を 五 チャンネル は ございません. | There are no place にを five channels now since (*S) apologizes and (*S) does not exist. | にを → には |

# Types & sources of errors (ATLAS-JE)

- Most Japanese spoken language characteristics are not processed by ATLAS
  - (of course as it is prepared for "clean texts"

| | | |
|---|---|---|
| 結構 です けど ね でき ます. | (*S) sleeps though it is excellent. | Back channel particle “ね” is not recognized, but is interpreted as the verb “寝る”. |
| ドイツ語 の が ある と 一 番 良い の です が 英語 は 読め ない の です. | English cannot be read as German が ある though it is the best. | Anaphoric pronoun “の” is not recognized. |
| はい 洗濯機 の 着席 優しく 払わ なければなりません ので ご 注意 下さい. | Please <払 わなければなり> note (*O) <sit-down> nice of the tile washing machine. | Modal expression “なければなりません” is not recognized. |
| 通常 一 週間 です でも 天気 が 悪い わ えー 少し 遅れる こと も あり ます. | The weather for one usual week it yet might be late of <badness> いわえ least | Phatic “えー” is not recognized. |
| かしこまりました 少々 御 待ち 下さい. | Please wait a little standing on ceremony. | However, politeness expression “かしこまりました” and honorific particle ”御” are recognized. |

# Participation to subjective evaluation

- Setting
  - Fluency
    - 2 English teachers, native speakers
    - + a French to help 1 of them (agreement on grades)
  - Adequacy
    - 1 Chinese Master student planned (<31/8) Wei W.
    - Some delay → 1 Chinese PhD student (Cao WJ)

# "all results in parallel": costly setting

- Initial suspicion: comparisons slow the process
  - there can be N log N comparisons (≈100 if N≈20)
- For fluency & adequacy
  - Time divided by >5 if no comparisons done
    - Don't present several outputs of same input together

| task / cost | grading (1 res.) | comparing 2 results | grading screenful | max # of comparisons | Total time *T* | Hypothesis |
|---|---|---|---|---|---|---|
| **Suspicion** | | | | | | |
| N outputs on screen | u | v | $Tg = N\,u$ | $C = N \log_2 N$ | $(N + C)\,u$ | 1.5 u ≤ v≤ 2u |
| N = 20 | u | v | $Tg = 20\,u$ | $\approx 100$ | $T \approx 200\,u$ $=[8..11]\,Tg$ | 170 u ≤ T ≤ 220 u |
| Worst (real) case (C/2) | u = 3-9 s | v = 20 s | *Tg = 3 mn* | *≈50—80—100* | *20—30—40 mn* | ≈ 60—180 + 1000—1600—2000 s |
| **If C/2** | u = 3-9 s | v = 20 s | *Tg = 3 mn* | *≈ 50* | *≈ 20 mn* | |
| Grading without compar. | | | *Tg = 3 mn* | *0* | *3 mn* | |
| **Grading CE** | **Turns** | **Screens** | | | **Total time** | **Gain** |
| Cao WJ | *5400* | *270* | *Tg = 3 mn* | *0* | *13.5 hours* | |
| IWSLT05 figures | *u = 3 s* | | *T = 9—10 mn* | *≈ 20 ?* | *4—5 days≈ 36—40 hours* | *≈ 3* |

# Remarks on adequacy evaluation

- Judgments are
  - biased by bad fluency
  - not task-oriented
- In the future: multiple choice understanding questions?
  - [Mitkov 2006] 3mn/question with machine help
  - If 10 questions/page (BTEC: 1 page = 40 sentences)
    - 30mn preparation
    - If 1mn to answer 1 question
    - then 10mn/page/evaluator
    - and ≈ 5mn/screen instead of 3mn/screen
    - but
    - better measure
    - 3 evaluators might be enough (better agreement)

# Task-related objective measures?

- Reason: n-gram based measures inadequate
  - Callison-Burch C., Osborne M. & Koehn P. (2006) *Re-evaluating the Role of BLEU in Machine Translation Research.* Proc. EACL-06, Trento, April 3-7, 2006, ITC/irst ed., 8 p.
- Fear: task-related measures too costly
- Possibilities
  - If goal = HQ translation
    - measure postedition time (cf. METEO, Spanam)
    - no added cost (beyond adapting translation editor)
  - If goal = understanding
    - not possible at 0 cost for all situations
    - If Web + e-commerce: measure # "buying acts"

# Conclusion (1/2)

- Experimenting with Systran, ATLAS
  - worse "objective" grades than other systems
  - but
    - all but 1 or 2 got dismal scores anyway
    - inadequacy of scores confirmed by subjective evaluation
      - subjectively, they are not worse than the others !
- Analysis of source of translation errors
  - Systran, ATLAS are built for "clean" texts
    - don't handle most spoken language phenomena
    - tunable only at dictionary level, which is not enough
  - SMT systems
    - lower scores than in IWSLT-05
    - main reasons:
      - lack of data: development set ≈ 25000 w, too small
      - different nature of training set and development set

# Conclusion (2/2)

- Participation in the subjective evaluation
  - proposal: reduce the cost of subjective evaluation
    - by not presenting outputs for same output together
  - proposal: better (task-oriented) & cheap objective measures
    - measure postedition time, or compare number of buying acts
- Objective measures
  - can involve human work
  - have no added cost if embedded in normal workflow
- Initial questions
  - Can commercial wide-coverage text-MT systems be used for speech-MT?
    - no, or developers would have to do a lot of work
  - Is it true that the subjective evaluation can be made less expensive by changing its setting ?
    - yes
  - How does the set of reference translations influence the evaluation scores produced by BLEU, NIST…?
    - not done for lack of time, human resources

# The End!

- Thanks to ATR for organizing this IWSLT-06
  - and to our reviewers