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

Deep understanding level

Interlingual level

Logico-semantic level

Mixing levels

Syntactico-functional level

Syntagmatic level

Morpho-syntactic level

Graphemic level

Ontological interlingua

Semantico-linguistic interlingua

SPA-structures (semantic & predicate-argument)

Multilevel description

F-structures (functional)

C-structures (constituent)

Tagged text

Text
# Systems run for IWSLT-04

<table>
<thead>
<tr>
<th>Pair</th>
<th>System(s)</th>
<th>Tuning (on training only)</th>
<th>Linguistic analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>AE</td>
<td>Systran-5</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CE</td>
<td>Systran-5</td>
<td>User dict.</td>
<td>-</td>
</tr>
<tr>
<td>IE</td>
<td>Systran-5</td>
<td>User dict. (50%)</td>
<td>-</td>
</tr>
<tr>
<td>JE</td>
<td>Systran-5</td>
<td>-</td>
<td>Tomokiyo</td>
</tr>
<tr>
<td>JE</td>
<td>ATLAS-2</td>
<td>-</td>
<td>Tomokiyo</td>
</tr>
</tbody>
</table>
SYSTRAN v5 — diagram

- **Deep understanding level**
- **Interlingual level**
- **Logico-semantic level**
- **Mixing levels**
- **Syntactico-functional level**
- **Syntagmatic level**
- **Morpho-syntactic level**
- **Graphemic level**

**Linear chart**

**Macros in fixed flowchart**

**FSA**

**FST**

- **Conceptual transfer**
- **Semantic transfer**
- **Multilevel transfer**
- **Syntactic transfer (deep)**
- **Syntactic transfer (surface)**
- **Semi-direct translation**
- **Direct translation**

**Ontological interlingua**

**Semantico-linguistic interlingua**

**SPA-structures (semantic & predicate-argument)**

**Multilevel description**

**F-structures (functional)**

**C-structures (constituent)**

**Tagged text**

**Text**
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

Node list linked to Node net

Enconverter written in pre-EnCo rule language

Deep understanding level

Interlingual level

Logico-semantic level

Syntactico-functional level

Syntagmatic level

Morphosyntactic level

Graphemic level

External Morph. Anal. (optional)

Node net = pre-UNL semantic graph

Node net linked to Node list

Conceptual transfer

Semantic transfer

Multilevel transfer

Syntactic transfer (deep)

Syntactic transfer (surface)

Semi-direct translation

Direct translation

Ascending transfer

Descending transfers

F-structures (functional)

C-structures (constituent)

Text

Nodes (semantic argument)
ATLAS (Fujitsu) — text

- Interlingual pivot  
  - 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  
  - 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

<table>
<thead>
<tr>
<th>Italian</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sì, abbiamo la Where, and The City Guide.</td>
<td>Yes, we have the Where, and The City Guide.</td>
</tr>
<tr>
<td>Congratualzioni, Henry. Sono felice di sentire del Suo fidanzamento con Jane.</td>
<td>Congratulations, Henry. I'm delighted to hear of your engagement to Jane.</td>
</tr>
<tr>
<td>Deve essere stato un grande shock per Lei.</td>
<td>It must have been a great shock to you.</td>
</tr>
<tr>
<td>Potrebbe pagare alla reception, prego?</td>
<td>Could you pay at the front desk, please?</td>
</tr>
<tr>
<td>Sono contento di averLa conosciuta. Grazie.</td>
<td>I'm glad I met you. Thank you.</td>
</tr>
<tr>
<td>Qui parla l'operatore dell' International Telephone Call Service.</td>
<td>This is the operator for International Telephone Call Service.</td>
</tr>
<tr>
<td>Facendo lo spelling è G-O-R-O-H.</td>
<td>It's spelled G-O-R-O-H.</td>
</tr>
</tbody>
</table>
### Objective evaluation (Systran)

<table>
<thead>
<tr>
<th></th>
<th>BLEU4</th>
<th>NIST</th>
<th>METEOR</th>
<th>WER</th>
<th>PER</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Spontaneous speech</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>official (with case + punctuation)</strong></td>
<td>0.0344</td>
<td>2.7374</td>
<td>0.3178</td>
<td>0.87129</td>
<td>0.743063</td>
</tr>
<tr>
<td><strong>additional (without case + punctuation)</strong></td>
<td>0.0406</td>
<td>2.8625</td>
<td>0.3184</td>
<td>0.880529</td>
<td>0.720287</td>
</tr>
<tr>
<td><strong>Read Speech</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>official (with case + punctuation)</strong></td>
<td>0.0536</td>
<td>3.7390</td>
<td>0.3210</td>
<td>0.805919</td>
<td>0.687017</td>
</tr>
<tr>
<td><strong>CRR</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>official (with case + punctuation)</strong></td>
<td>0.0366</td>
<td>2.685</td>
<td>0.3178</td>
<td>0.858339</td>
<td>0.726484</td>
</tr>
<tr>
<td><strong>additional (without case + punctuation)</strong></td>
<td>0.0749</td>
<td>4.4256</td>
<td>0.3694</td>
<td>0.780118</td>
<td>0.643764</td>
</tr>
</tbody>
</table>
Chinese segmentation problems

(Systran)

<table>
<thead>
<tr>
<th>汉语</th>
<th>英语</th>
</tr>
</thead>
<tbody>
<tr>
<td>对历史( history )感兴趣 (to be interested)</td>
<td>be interested in history</td>
</tr>
<tr>
<td>职员(employer)会(can)轮流放假</td>
<td>employee can take several days off by turns</td>
</tr>
<tr>
<td>我就要替你喝完秋葵汤(soupe)了。</td>
<td>gumbo</td>
</tr>
<tr>
<td>雕塑(sculpture)感兴趣(be interested)</td>
<td>interested in sculpture</td>
</tr>
<tr>
<td>孟斐斯(proper name)</td>
<td>Memphis</td>
</tr>
<tr>
<td>理察德(Richard)波尔曼</td>
<td>Richard Paulman</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
<th>NIST</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Systran</strong></td>
<td>ASR output (Read speech)</td>
<td>0.0755</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
<th>NIST</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Atlas</strong></td>
<td>ASR output (Read speech)</td>
<td>0.1084</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
<th>NIST</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Systran</strong></td>
<td>ASR output (Read speech)</td>
<td>0.049</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
<th>NIST</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Systran</strong></td>
<td>ASR output (Read speech)</td>
<td>0.1368</td>
</tr>
</tbody>
</table>
# Types & sources of errors (Systran-JE)

## Synthesis

<table>
<thead>
<tr>
<th>Error Type</th>
<th>Example Utterance</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Euphemistic utterances</td>
<td><strong>When the utterance is <em>euphemistic</em> (が), the particle is always translated by “but”</strong></td>
<td>切りますよ。 → “it cuts”?</td>
</tr>
<tr>
<td>Some utterances do not make sense without context</td>
<td></td>
<td></td>
</tr>
<tr>
<td>First person subject omitted in Japanese</td>
<td><strong>When the <em>first person subject</em> is <em>omitted in Japanese</em>, it is always translated as “it”</strong></td>
<td>ここで降ります。 → “It gets off here.”</td>
</tr>
<tr>
<td>Interrogative pronouns and adverbs are always incorrectly shifted</td>
<td><strong>Interrogative pronouns and adverbs are always (incorrectly) shifted at the end of the translation</strong></td>
<td>オペラ座はどこですか。 → “Is the opera house where?”</td>
</tr>
</tbody>
</table>
| Idiom expressions not contained in the SYSTRAN dictionaries               | **Many daily life *idiomatic expressions* are not contained in the SYSTRAN dictionaries** | どういたしまして。 → “How doing.”  
もしもし。 → “It does.”  
さようなら。 → “Way if.” |
| Requests or invitations                                                   | **Requests or invitations are not always well translated**                      | 注文したいのです。 → “It is to like to order.”  
一緒に行きましょう。 → “It will go together.” |
| Verb valency different in two expressions                                 | **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 not correctly rendered in English          | **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。 → “The air ticket *was* forgotten in the house.” |
| Politeness                                                               | **Positive point: *lexicalized Japanese politeness is correctly handled***     | そのまま切らずにお待ち下さい。 → “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 致しません。 | 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 致しません”, 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

<table>
<thead>
<tr>
<th>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。</th>
<th>The television cannot be had to be used after the take off which apologizes and not is.</th>
<th>Verb “でないと”</th>
</tr>
</thead>
<tbody>
<tr>
<td>やってみますがから予約できるか保証し兼ねます。</td>
<td>Whether can be reserved cannot be guaranteed やってみます。</td>
<td>Verb “やる”</td>
</tr>
<tr>
<td>えーっとそれは六百円です。</td>
<td>Food っとそれ is 600 yen.</td>
<td>Phatic “えーっと”</td>
</tr>
<tr>
<td>以前は野球をするのが好きでした でも今 スキーの方が興味があります</td>
<td>It was liked to play baseball and skiing is interesting yet now before.</td>
<td>Conjunction “でも”</td>
</tr>
<tr>
<td>切って今手がございますどうぞご覧下さい。</td>
<td>(*S) cuts (O), and there is a hand now and (*S) sees please.</td>
<td>Polite expression</td>
</tr>
<tr>
<td>結構ですけどねできます。</td>
<td>(*S) sleeps though it is excellent.</td>
<td>Modal particle “ね”</td>
</tr>
<tr>
<td>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</td>
<td>English cannot be read as German がある though it is the best.</td>
<td>Referential noun “の”</td>
</tr>
<tr>
<td>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。</td>
<td>Please &lt;払わなければなりません&gt; note (O) &lt;sit-down&gt; nice of the tile washing machine.</td>
<td>Modal expression “なければなりません”</td>
</tr>
<tr>
<td>通常一週間でも天気が悪いведение少し遅れることもあります。</td>
<td>The weather for one usual week it yet might be late of &lt;badness&gt; いわえleast</td>
<td>Phatic “えー”</td>
</tr>
<tr>
<td>かしこまりました少々御待ち下さい。</td>
<td>Please wait a little standing on ceremony.</td>
<td>Polite expression “かしこまりました” and Honorific expression “御”</td>
</tr>
<tr>
<td>陶器御茶の方御酒を買いましたこちらは全ておで一です。</td>
<td>These by which person 御酒 of earthen 茶 is bought are all sleeps.</td>
<td>Honorific expression “御”</td>
</tr>
<tr>
<td>そうですねあと一時間で着陸します。</td>
<td>(*S) &lt;aspect&gt; has, (*S) sleeps, and (*S) will land in about another hour.</td>
<td>“‘ね’’ in dialogues</td>
</tr>
<tr>
<td>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。</td>
<td>It is in the corner of the man in guestこちらです oral その building</td>
<td>Deictic expression “こちら”</td>
</tr>
</tbody>
</table>
## Types & sources of errors (ATLAS-JE)

### Problems coming from the dictionary

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

<table>
<thead>
<tr>
<th>Problems coming from the dictionary</th>
<th>850,000 entries still not enough</th>
<th>maybe 4.5 M are?</th>
</tr>
</thead>
<tbody>
<tr>
<td>赤青緑黄色がございますどの色が御好みですか。</td>
<td>Which color with 赤青緑 yellow is favor?</td>
<td>Special words “赤,青,緑”</td>
</tr>
<tr>
<td>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。</td>
<td>It is necessary to turn right after (*S) goes out of the door of いいえそ。</td>
<td>Deictic and anaphoric word Mots déictique et anaphorique “その”</td>
</tr>
<tr>
<td>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</td>
<td>The name by which the flight-number and staying of 御客様 hello are taken</td>
<td>Honorific word “御客様”</td>
</tr>
<tr>
<td>ラジオの電源スイッチは一人がです</td>
<td>つまみ of &lt;one person&gt; です is a thing to adjust the volume. the power supply switch of the radio</td>
<td>Special word “つまみ”</td>
</tr>
<tr>
<td>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。</td>
<td>Guest もうしばらく is waited and бро present the answer within one week.</td>
<td>Special word “もうしばらく”</td>
</tr>
<tr>
<td>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。</td>
<td>It is in the corner of the man in guestこちらです oral その building</td>
<td>“御客様” Deictic word “こちら”</td>
</tr>
<tr>
<td>こちらの大きな連中は記念ように保存されています。</td>
<td>A big party there is preserved in the commemoration way.</td>
<td>Deictic word “こちら”</td>
</tr>
<tr>
<td>いいえまだです。</td>
<td>いいえまだです。</td>
<td>Special word “いいえ”</td>
</tr>
<tr>
<td>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。</td>
<td>The nearby restaurant hangs in the vicinity for 30 another minutes in the car.</td>
<td>Semantic ambiguity of verb “掛かる”</td>
</tr>
<tr>
<td>Problems in the input Japanese text and consequences on ATLAS translation results</td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------------------------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>精神は三名ドルほどです。</td>
<td>The soul is about three person dollar.</td>
<td>?</td>
</tr>
<tr>
<td>私の国は中国のりんご君日本です。</td>
<td>My country is apple 君日本 of China.</td>
<td>?</td>
</tr>
<tr>
<td>離陸を三十分以内には昼食を御出し致します。</td>
<td>The take off is served and I will serve lunch within 30 minutes.</td>
<td>離陸を → 離陸後</td>
</tr>
<tr>
<td>トイレは機内高校ですご案内致します。</td>
<td>It will be a guide of the rest room that an in-flight high school has (*O).</td>
<td>高校 → 後方</td>
</tr>
<tr>
<td>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。</td>
<td>The yes credit card can be had to be used and the card where (*S) returns is visa</td>
<td>帰る → 使える</td>
</tr>
<tr>
<td>はい車で十分ほどと頃に一つございます。</td>
<td>It is a tile car and there is one every about ten minutes.</td>
<td>と頃に → のところに</td>
</tr>
<tr>
<td>こちらです化粧品は二階ですすぐにデータで上がって下さい。</td>
<td>Cosmetics which have (*O) &lt;here&gt; must rise by data in placing by the second floor.</td>
<td>えで 徳ータ → エレベータ</td>
</tr>
<tr>
<td>やってみますがからぞ予約できるか保証し兼ねます。</td>
<td>Whether からぞ can be reserved cannot be guaranteed やってみます。</td>
<td>からぞ → 必ず</td>
</tr>
<tr>
<td>申し訳ありません今のですにを五チャンネルはございますません。</td>
<td>There are no place にを five channels now since (*S) apologizes and (*S) does not exist.</td>
<td>にを → には</td>
</tr>
</tbody>
</table>
## 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")

<table>
<thead>
<tr>
<th>Japanese</th>
<th>English</th>
<th>ATLAS Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>結構 です けど ね できます。</td>
<td>(*S) sleeps though it is excellent.</td>
<td>Back channel particle “ね” is not recognized, but is interpreted as the verb “寝る”</td>
</tr>
<tr>
<td>ドイツ語 の ある と 一番 良い の です が 英語 は 読め ない の です。</td>
<td>English cannot be read as German がある though it is the best.</td>
<td>Anaphoric pronoun “の” is not recognized.</td>
</tr>
<tr>
<td>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。</td>
<td>Please &lt;払 わなければならない&gt; note (*O) &lt;sit-down&gt; nice of the tile washing machine.</td>
<td>Modal expression “なければなりません” is not recognized.</td>
</tr>
<tr>
<td>通常 一 週間 です でも 天気 が 悪い ええ 少し 遅れること も あります。</td>
<td>The weather for one usual week it yet might be late of &lt;badness&gt; いわえ least</td>
<td>Phatic “ええ” is not recognized.</td>
</tr>
<tr>
<td>かしこまりました 少々 御 待ち下さい。</td>
<td>Please wait a little standing on ceremony.</td>
<td>However, politeness expression “かしこまりました” and honorific particle ”御” are recognized.</td>
</tr>
</tbody>
</table>
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 ($\approx 100$ if $N \approx 20$)

- **For fluency & adequacy**
  - Time divided by >5 if no comparisons done

- **Don't present several outputs of same input together**

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Total time $T$</th>
<th>$u$</th>
<th>$v$</th>
<th>$T_{g} = Nu$</th>
<th>$C = N \log_2 N$</th>
<th>$(N + C)u$</th>
<th>$1.5u \leq v \leq 2u$</th>
<th>$170u \leq T \leq 220u$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Suspicion</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N outputs on screen</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$N = 20$</td>
<td></td>
<td>$u$</td>
<td>$v$</td>
<td>$T_{g} = 20u$</td>
<td>$\approx 100$</td>
<td>$T \approx 200u$</td>
<td>$= [8..11]T_{g}$</td>
<td>$170u \leq T \leq 220u$</td>
</tr>
<tr>
<td>Worst (real case (C/2))</td>
<td></td>
<td>$u = 3-9s$</td>
<td>$v = 20s$</td>
<td>$T_{g} = 3mn$</td>
<td>$\approx 50—80—100$</td>
<td>$20—30—40mn$</td>
<td>$\approx 60—180 + 1000—1600—2000s$</td>
<td></td>
</tr>
<tr>
<td><strong>If C/2</strong></td>
<td></td>
<td>$u = 3-9s$</td>
<td>$v = 20s$</td>
<td>$T_{g} = 3mn$</td>
<td>$\approx 50$</td>
<td>$\approx 20mn$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grading without compar.</td>
<td></td>
<td></td>
<td></td>
<td>$T_{g} = 3mn$</td>
<td>$0$</td>
<td>$3mn$</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Grading CE</strong></td>
<td>Turns</td>
<td>Screens</td>
<td>$T_{g} = 3mn$</td>
<td>$0$</td>
<td>$13.5$ hours</td>
<td>$4—5$ days $\approx 36—40$ hours</td>
<td>$\approx 3$</td>
<td></td>
</tr>
</tbody>
</table>
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

- Fear: task-related measures too costly

- Possibilities
  - If goal = HQ translation
    - measure postedited 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
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 post-edition 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