

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, ...?

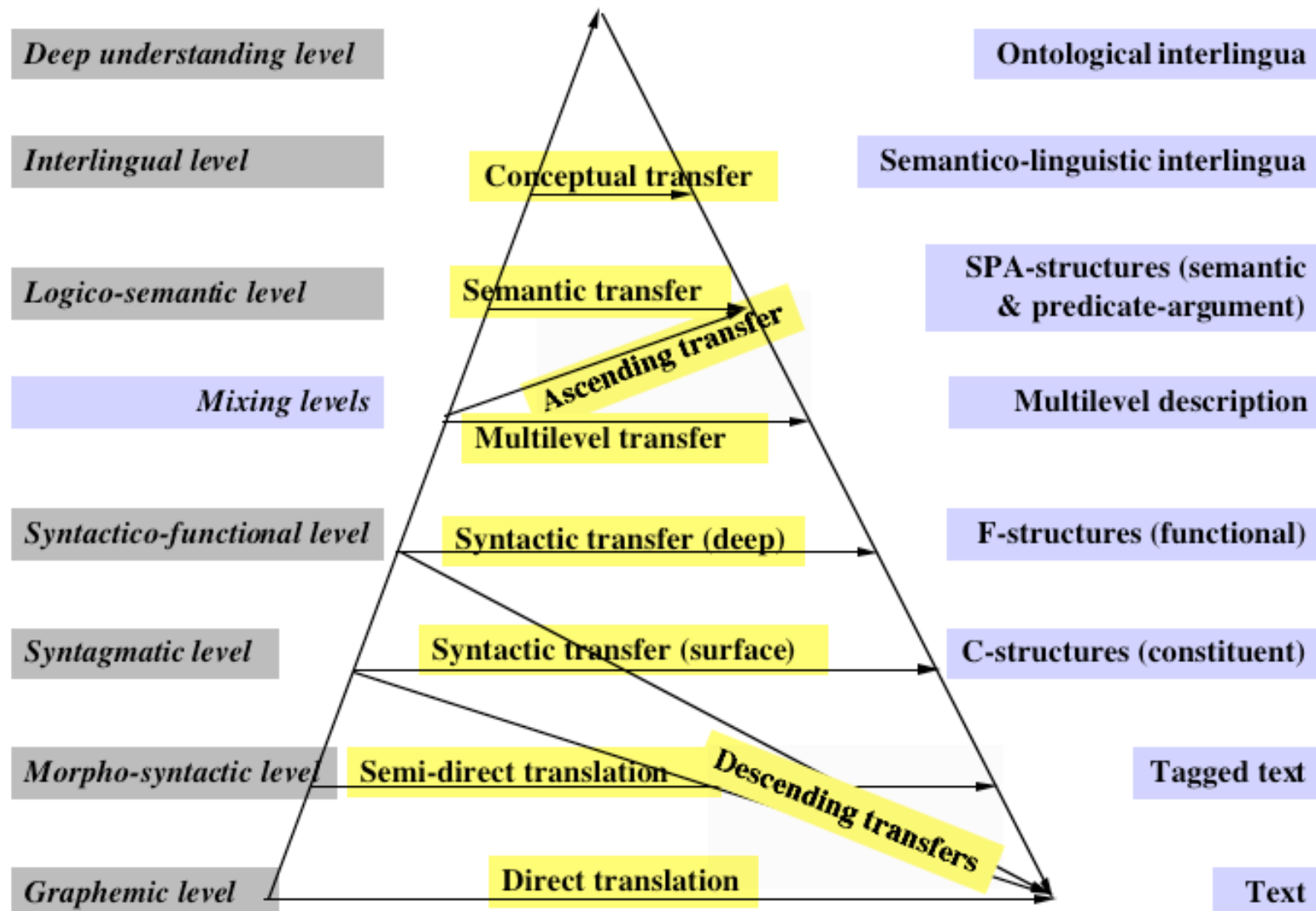
Some commercial systems

- 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), Linguattec/Lingenio (PC-translator, based on LMT), WordMagic, Compendium (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

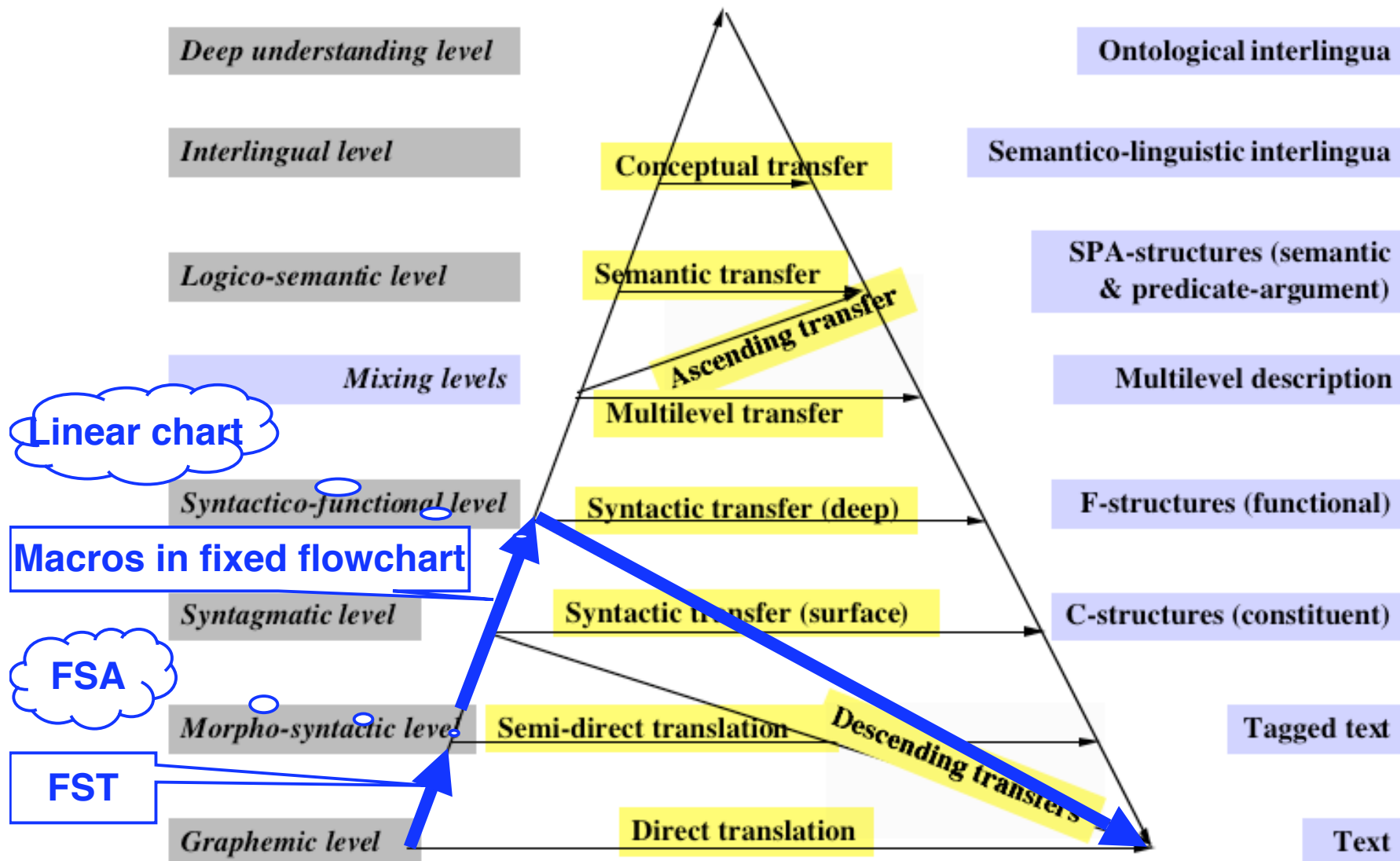
Vauquois's triangle



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

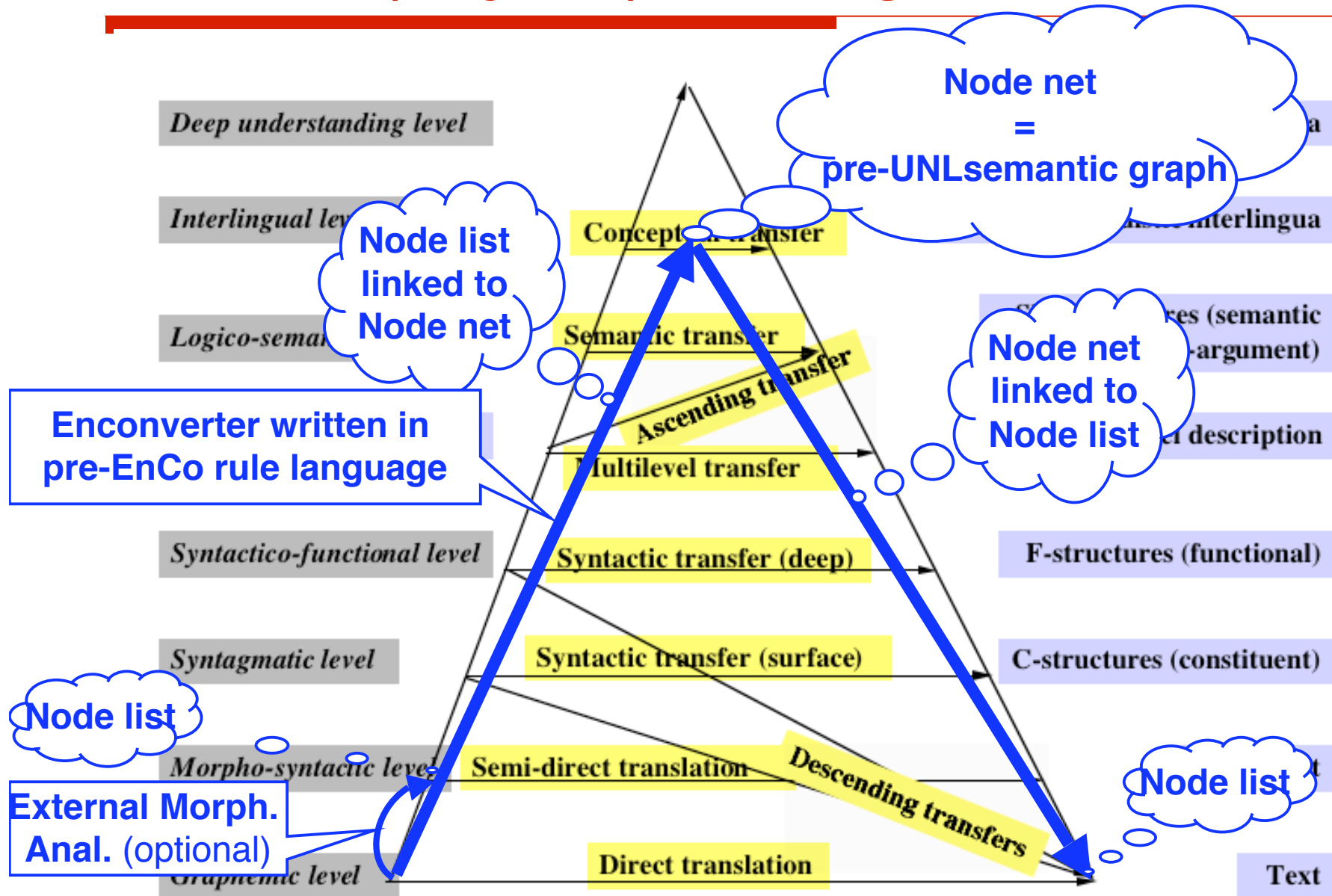
SYSTRAN v5 — diagram



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 ($\cup \cap \supseteq \subseteq$ 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羹汤(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(proper name)	Memphis
理查德(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。 → “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。 → “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.” もしもし。 → “It does.” さようなら。 → “Way if.”
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.”
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。 → “The air ticket was forgotten in the house.”
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 でないで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 “でないで”
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 っとそれは 600 yen.	Phatic “えーっと”
以前は野球をするのが好きでしたが今で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"
<u>いいえ</u> 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 <u>もうしばらく</u> 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" Deictic word "こちら"
<u>あちら</u> 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"
<u>いいえ</u> 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 <u>hangs</u> in the vicinity for 30 another minutes <u>in the</u> 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 (*0).	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 <u>えで</u> <u>データ</u> で上がって下さい。	Cosmetics which have (*0) <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 <u>にを</u> 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 のでご注意ください。	Please < 払わなければなり > note (*0) <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 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 $N \approx 20$)
- 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	$Tg = Nu$	$C = N \log_2 N$	$(N + C)u$	$1.5u \leq v \leq 2u$
$N = 20$	u	v	$Tg = 20u$	≈ 100	$T \approx 200u$ $= [8..11] Tg$	$170u \leq T \leq 220u$
Worst (real) case (C/2)	$u = 3-9 \text{ s}$	$v = 20 \text{ s}$	$Tg = 3 \text{ mn}$	$\approx 50-80-100$	$20-30-40 \text{ mn}$	$\approx 60-180 +$ $1000-1600-2000 \text{ s}$
If C/2	$u = 3-9 \text{ s}$	$v = 20 \text{ s}$	$Tg = 3 \text{ mn}$	≈ 50	$\approx 20 \text{ mn}$	
Grading without compar.			$Tg = 3 \text{ mn}$	0	3 mn	
Grading CE	Turns	Screens			Total time	Gain
Cao WJ	5400	270	$Tg = 3 \text{ mn}$	0	13.5 hours	
IWSLT05 figures	$u = 3 \text{ s}$		$T = 9-10 \text{ mn}$	$\approx 20 ?$	$4-5 \text{ days} \approx$ $36-40 \text{ hours}$	≈ 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 \approx 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 \approx 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