NRC·CNRC

Institute for Information Technology

Introduction à la traduction automatique par règles

Pierre Isabelle

Septembre 2010

National Research Council Canada
Conseil national de recherches Canada
Canada

NRC·CNRC
Institute for Information Technology

Les débuts de la traduction automatique

- 1948: Andrew Booth (Londres) présente un dictionnaire électronique
  - Avec rudiments d'analyse morphologique pour réduire le nombre d'entrées
- 1949: Mémorandum de Warren Weaver
  - Un fondateur de la théorie de l'information, avec C. Shannon (cf vidéo sem. dern.)
  - Propose d'appliquer les nouveaux calculateurs au problème de la traduction automatique
  - Associe le problème de la TA à celui de la cryptographie développée durant 2ème guerre
    - Préfigure la TA statistique moderne
    - Mais à l'époque on ne disposait ni des données requises (corpus parallèles en format électronique) ni des machines suffisamment puissantes
  - Propose aussi de recourir à des analyses linguistiques qui permettraient de découvrir une base universelle commune à toutes les langues
    - Préfigure la TA « à base de connaissances » qui tente le développement d'une « interlangue sémantique »

NRC·CNRC
Institute for Information Technology

Les débuts (suite)

- 1952: première conférence scientifique sur la TA au MIT
  - Montre un clivage entre les « empiristes » et les « perfectionnistes »
- 1954: première démo publique (Univ. Georgetown et IBM)
  - Approche assez simpliste: 49 phrases russes simples traduites en anglais au moyen de dictionnaire 250 mots et 6 règles;
  - Peu de généralité, mais suscite l'enthousiasme des médias et des sponsors
- Beaucoup de financement pour des travaux fondés sur des approches « empiriques »
  - Vite implanter une méthode simple, examiner les sorties, apporter des améliorations incrémentales
  - Contexte de la guerre froide: besoin de traductions russe → anglais
  - Morphologie + dictionnaire bilingue de (groupes de) mots + règles simples de réordonnancement
- Les espoirs des chercheurs et des sponsors sont élevés:
  - Cf. vidéo de la semaine dernière: « dans 5 ans les problèmes de la traduction technique

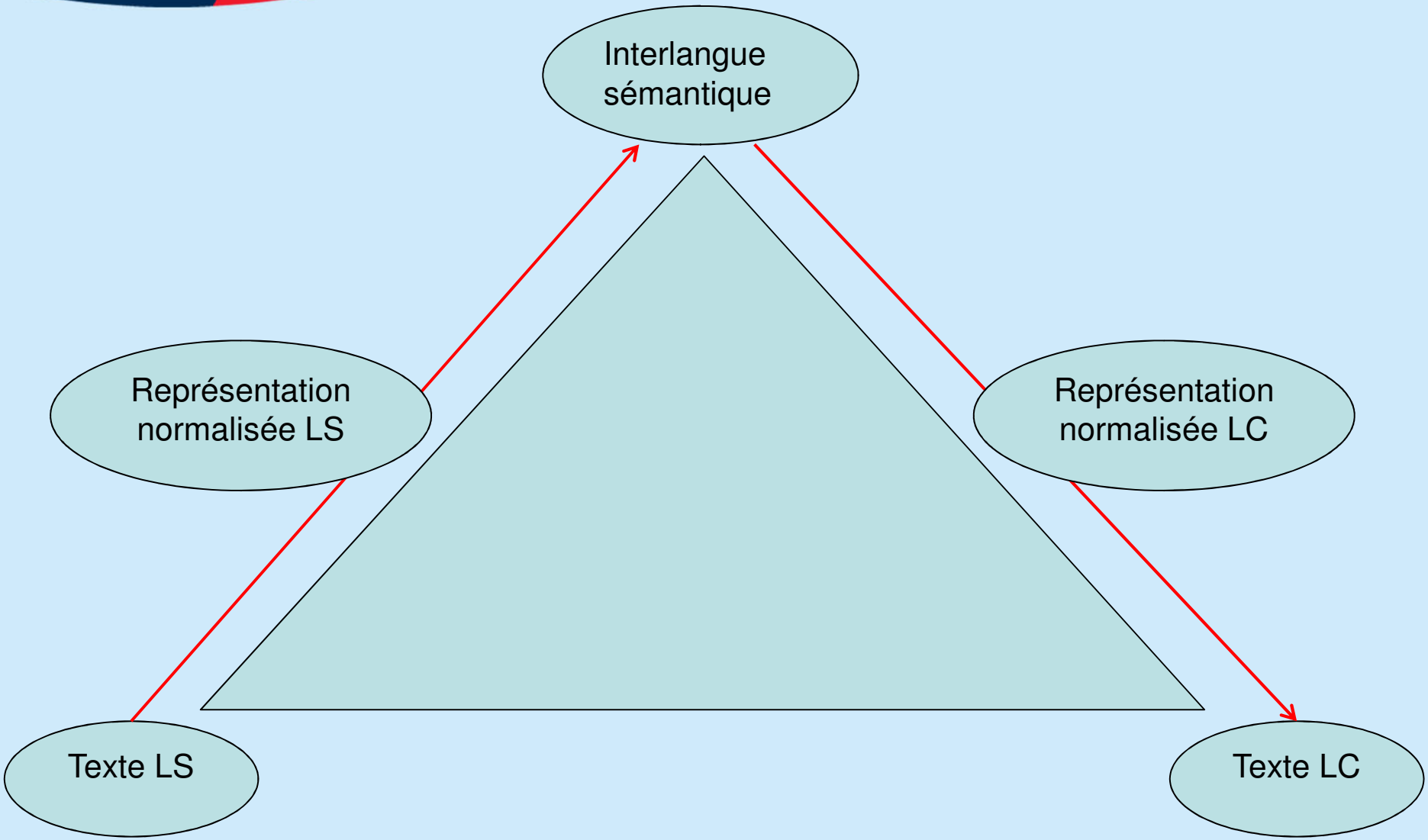
NRC·CNRC
Institute for Information Technology

La désillusion

- Dès 1960, Y. Bar-Hillel mettait en doute la faisabilité de la TA: 1) entièrement automatique; 2) de haute qualité; 3) de textes généraux
  - Proposait un tandem personne/machine (pré-édition et/ou post-édition manuelle)
- Entretemps, premiers essais d'exploitation de systèmes de TA (IBM et Georgetown University); la qualité des traductions machine s'avère encore en général très mauvaise
- 1966: rapport du comité ALPAC mandaté par la NSF pour étudier les résultats et perspectives de la TA
  - Condamne les efforts « empiriques » à court terme: « *There is no immediate or predictable prospect of useful machine translation* ».
  - Recommande le développement d'outils moins ambitieux (e.g. dictionnaires électroniques)
  - Effet dévastateur sur les recherches « empiristes » en TA aux U.S.A. et ailleurs
  - Recommande la poursuite de recherches à plus long terme en linguistique informatique → favorise l'approche « perfectionniste ».

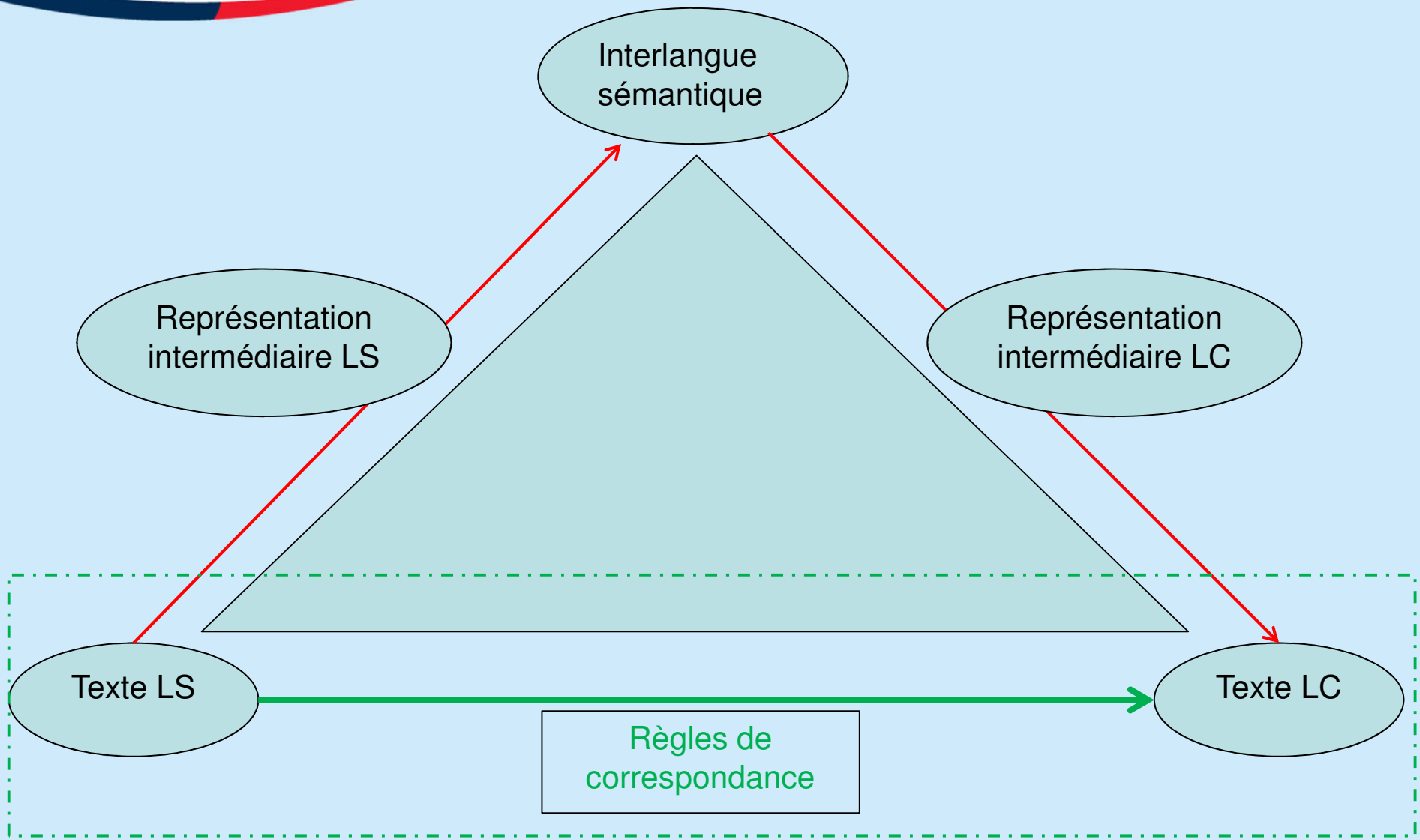
NRC·CNRC
Institute for Information Technology

Le triangle de Vauquois
Niveaux d'abstraction



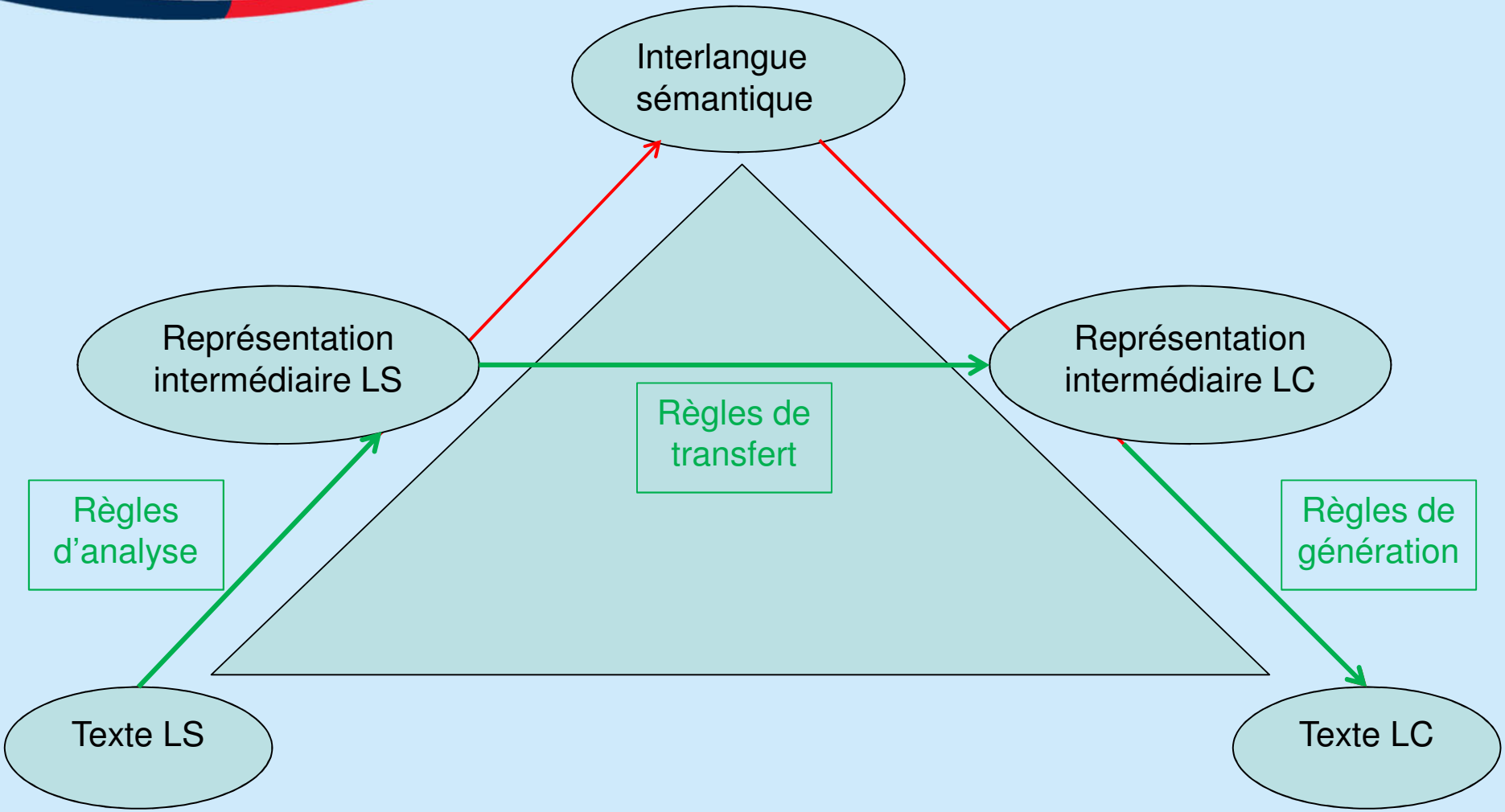
NRC·CNRC
Institute for Information Technology

Le triangle de Vauquois
Approche directe



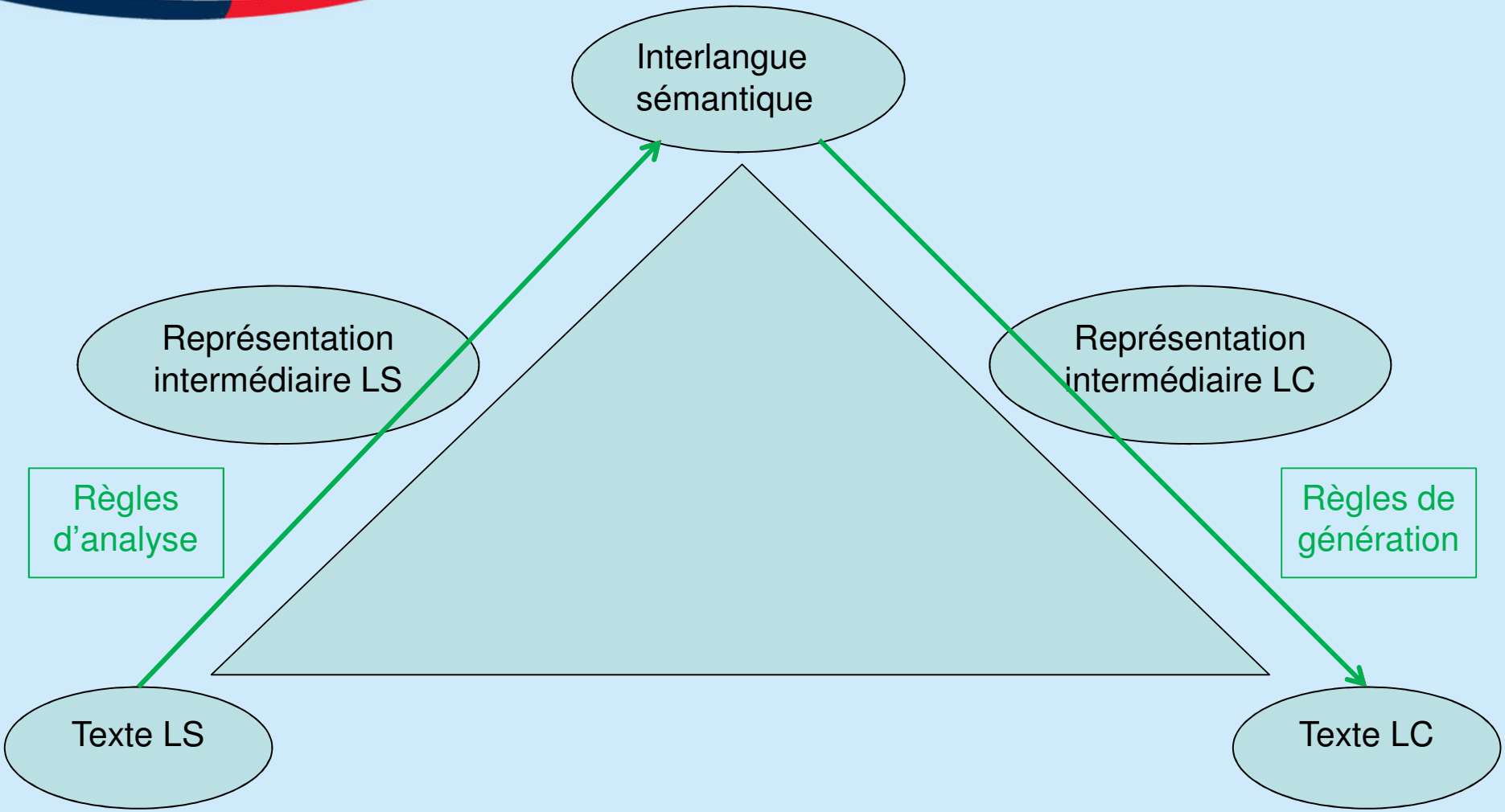
NRC·CNRC
Institute for Information Technology

Le triangle de Vauquois
Approche par transfert



NRC·CNRC
Institute for Information Technology

Le triangle de Vauquois
Approche par interlangue



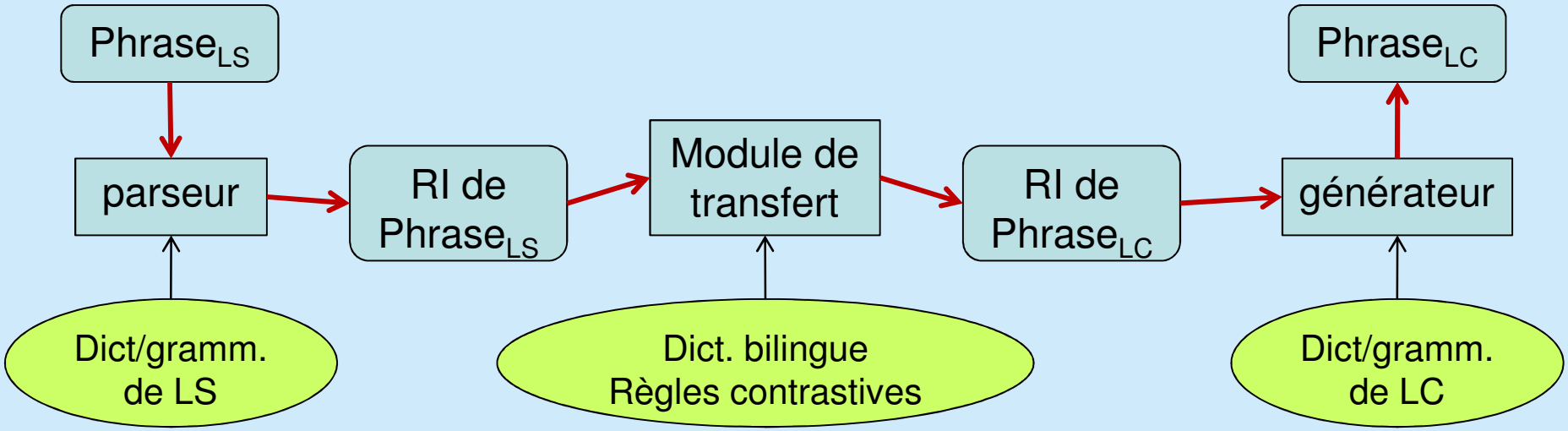
NRC·CNRC
Institute for Information Technology

Le triangle de Vauquois Discussion

- Approche directe: pas d'analyse préalable des phrases à traduire
  - Au départ = approche simpliste « dictionnaire seulement »
  - Se complexifie et absorbe certains éléments des approches indirectes
  - Au début des années 70 donne lieu à plusieurs efforts de commercialisation, dont certains des systèmes les plus connus aujourd'hui (*Systran* et *Pro-MT*)
  - Toutes les règles dépendent du couple de langue particulier; pour traduire entre *n* langues on a besoin de *n*(n-1)* modules de règles (CE: 23*22 = 506!)
- Approches indirectes
  - Analyse LS indépendante de LC (*n* modules) ; génération de LC indépendante de LS (*n* modules)
  - Approche par interlangue: pas d'autres modules requis, mais l'analyse et la génération sont d'une complexité problématique
  - Approche par transfert: l'analyse et la génération sont beaucoup moins abstraites car elles s'appuient sur un module de transfert spécifique à chaque couple LS-LC

NRC·CNRC
Institute for Information Technology

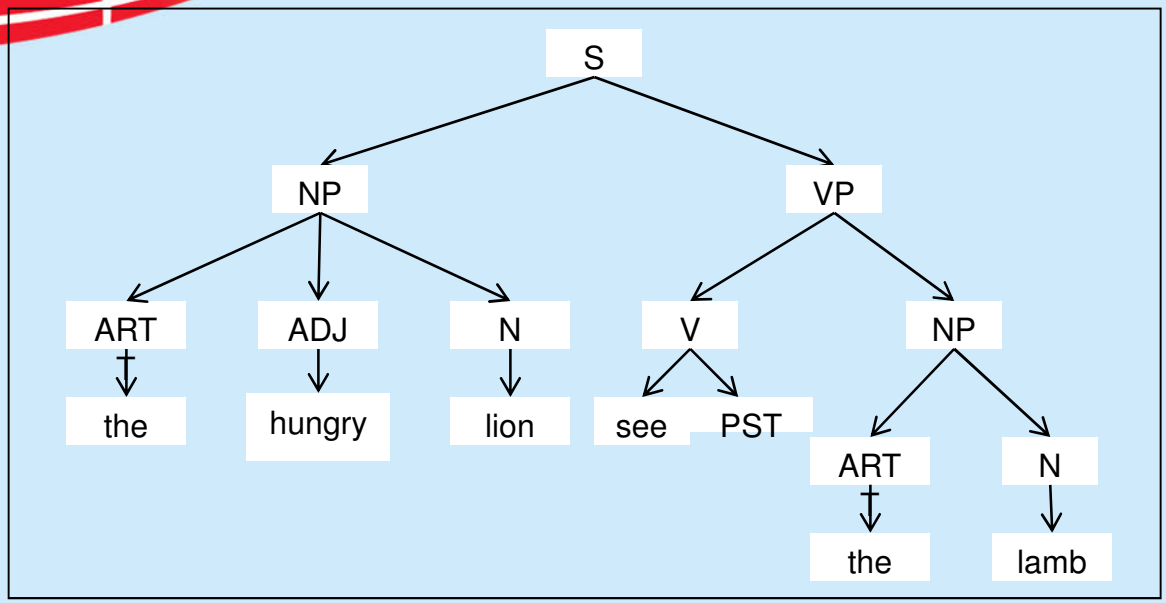
L'approche par transfert
Architecture générale



- Modèle proposé par V. Yngve, MIT, 1957; vraiment implanté seulement après 1970
- Équivalences LS-LC formulées à un niveau + abstrait: représentations intermédiaires
  - Nature syntaxique ou syntaxico-sémantique
- 3 étapes: analyse, transfert, génération
- Séparation entre les données linguistiques et les algorithmes :
  - Dictionnaire et grammaires = règles
  - Parseur/générateur = programmes qui appliquent ces règles sur texte à traduire

NRC·CNRC Institute for Information Technology

L'approche par transfert
Module d'analyse vers une RI

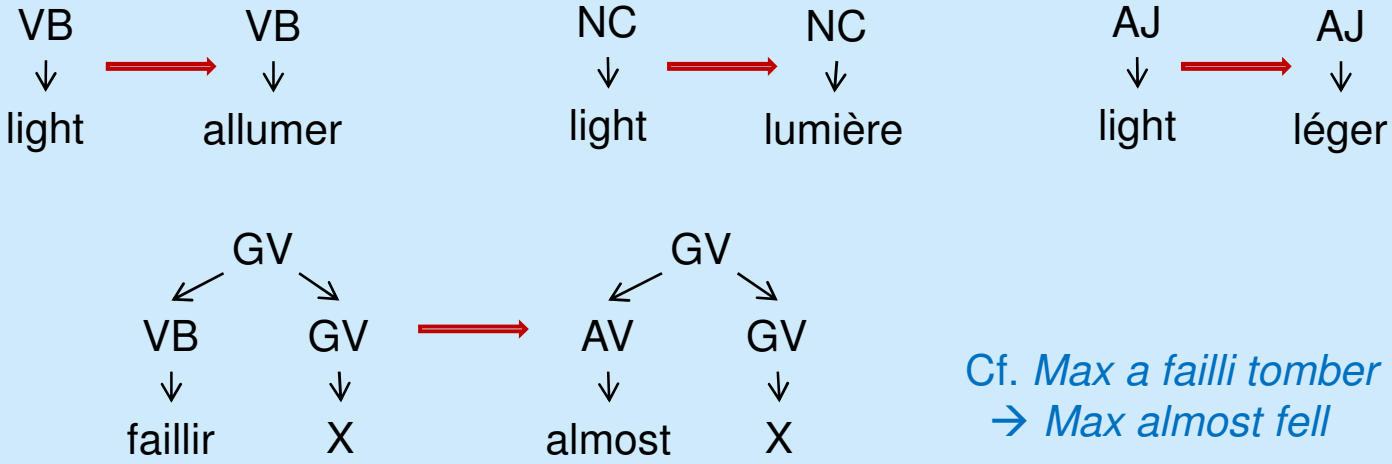


- Typiquement: le module d'analyse effectue pour chaque phrase du texte LS:
  - segmentation en mots
  - analyse morphologique
  - parsing → arbre de structure syntaxique à la Chosmky
- Recours à des méta-langages spécialisés facilitant l'écriture de règles de grammaire *formelles* i.e. exploitables par une machine bête

  Symbole entrée X → Symbole sortie Y

NRC·CNRC Institute for Information Technology

L'approche par transfert
Module de transfert



Cf. *Max a failli tomber*
→ *Max almost fell*

- Module de transfert convertit arbre LS en un arbre équivalent de LC
- Transfert lexical (illustré ci-dessus)
- Transfert structural
  - *Prépositions* en français mais *postpositions* en japonais
  - Dépend du niveau d'abstraction des représentations intermédiaires

NRC·CNRC
Institute for Information Technology

L'approche par transfert
Évolution

- Paradigme dominant dans les recherches entre 1970 et 1995
- Tendance de la recherche: structure intermédiaire *de plus en plus abstraite*
- Représentations syntaxiques *profondes*:
  - Factoriser le traitement des différentes versions grammaticales d'une même phrase
  - P. ex. règles de *transformation* entre phrases actives et les passives correspondantes)
- Représentations *syntaxico-sémantiques*:
  - Efforts de normalisation sémantique
  - Faire converger des éléments sémantiquement équivalents comme les synonymes (cf. modèle Sens-Texte de I. Melcuk)
  - À l'inverse, tenter de lever des ambiguïtés lexicales
    - Restrictions de sélection

NRC·CNRC
Institute for Information Technology

L'approche par transfert
Capacités et limites

- Essai de réponse à plusieurs des difficultés mentionnées la semaine dernière. Les équivalences traductionnelles reposent sur:
  - Segmentation et normalisation morphologique préalables; les règles se généralisent aux différentes variantes d'un même mot
  - La mise en correspondance LS-LC peut référencer le contexte grammatical: résolution des ambiguités catégorielles, identification des expressions idiomatiques et des collocations, traitement beaucoup de divergences structurales et de l'ordre des mots
- Limites:
  - Le parsing introduit de nouvelles difficultés :
    - On doit supposer (souvent à tort) que le texte LS est grammaticalement correct
    - On doit faire face aux ambiguïtés grammaticales (cf. *Time flies like an arrow*)
  - Approche sémantique
  - Contexte se limite à la structure grammaticale des phrases individuelles
    - Ne permet de résoudre qu'une petite partie des ambiguités

NRC·CNRC
Institute for Information Technology

L'approche par transfert Applications

- Un succès retentissant: système MÉTÉO:
  - Développé par le groupe TAUM de l'Université de Montréal
  - Sous-langue des bulletins météorologiques
  - En exploitation dès 1977
  - 30 millions de mots par an
- Sérieuses difficultés de mise à l'échelle pour des applications plus ambitieuses
- Certains systèmes généraux ont été et continuent d'être commercialisés:
  - METAL et ses descendants
  - LMT (IBM)
  - Etc.
- Malheureusement, ces systèmes n'ont jamais réussi à s'imposer sur le marché devant les systèmes plus « directs » comme Systran et PROMT

NRC·CNRC
Institute for Information Technology

Approches par interlangue et à base de connaissances

- En principe deux notions distinctes mais qui ont convergé
- Issu de la communauté intelligence artificielle (IA) et non de la linguistique (informatique)
- Emphase sur la nécessité de *comprendre* le texte LS, de *raisonner* à partir non seulement du texte mais aussi de *connaissances non-linguistiques* (sens commun, stéréotypes sociaux, savoir spécialisé)
- Exemples de la semaine dernière:

  Monkeys like bananas. They are good for them.
- Un autre exemple emprunté à Martin Kay:
  - En suisse française, on *valide* son ticket en montant sur le train
  - En suisse allemande, on *invalide* (« entwerten ») son ticket exactement dans les mêmes circonstances
  - Une approche purement linguistique : on traduirait un mot par son contraire!
  - On doit passer par l'intermédiaire d'une correspondance entre les expressions linguistiques et une conceptualisation non linguistique des situations

NRC·CNRC
Institute for Information Technology

Approches par interlangue et à base de connaissances

- Les chercheurs de la communauté IA ont mis au point différent formalismes pour représenter des connaissances non linguistiques et raisonner sur ces représentations
  - Divers formalismes basés la logique du premier ordre et diverses extensions
  - Le formalisme *Conceptual Dependency* de Schank qui analyse les verbes en un petit nombre d'atomes sémantiques censés capter des aspects du raisonnement
- Idéalement ces formalismes devraient permettre une *interlangue sémantique*
  - Neutres entre les langues différentes
  - Sans ambiguité: tout est rendu explicite
- On a aussi étudié l'organisation du savoir non-linguistique en structures complexes liées aux comportements humains
  - Ex. les « scripts » de (R. Schank); une visite au resto se déroule selon un scénario typique: entrée, accueil par une hotesse, désignation d'une table, présentation de la carte, etc.
  - Sert à prédire la traduction de « check » dans *The waiter brought the check*

= l'addition ≠ le chèque

NRC·CNRC
Institute for Information Technology

Approches à base de connaissances : Capacités et limites

- Recherches très intéressantes
- On n'a jamais vraiment dépassé la taille de systèmes jouets
- Se butent à deux problèmes extrêmement difficiles:
  - On ne sait pas comment définir une interlangue neutre; cette notion a-t-elle-même un sens?
    - En pratique les tentatives d'interlangue réalisées reviennent plus ou moins à « ENGLISH WITH CAPITAL LETTERS ».
    - Si l'interlangue n'est pas neutre, fait-on mieux qu'*une double traduction*?
  - Nécessite en un savoir de taille gigantesque dont la nature demeure mal comprise (e.g. le « gros bon sens »)
    - On se heurte au goulement d'étranglement de l'acquisition des connaissances!

NRC·CNRC
Institute for Information Technology

CONCLUSIONS

- Suite à l'échec des approches simplistes « directes » les chercheurs en TA se sont tournés vers des modèles indirects spécifiés par des systèmes de règles
- Les approches par transfert visaient à formuler des équivalences langue-à-langue à un niveau d'abstraction qui permet de tenir compte du contexte grammatical et de contraintes sémantiques « locales »
  - Succès limité, surtout pour des sous-langues simple comme la météo
  - L'automatisation du parsing est un problème qui n'est que partiellement résolu
  - Trop d'ambiguïtés surgissent et subsistent
- Les approches par interlangue et à base de connaissances exploraient des solutions « perfectionnistes » qui seraient en principe capable de résoudre les ambiguïtés sur la base d'un contexte global, linguistique et extra-linguistique
  - A permis de faire progresser nos connaissances
  - Mais a échoué sur le récif du problème d'acquisition du savoir
- Vers 1990, le constat est fait et la scène est mise pour un changement de paradigme fondé sur l'apprentissage machine