
Introduction au TALN et à l'ingénierie linguistique

Isabelle Tellier

ILPGA

1. Quelques notions de sciences du langage
2. Applications et enjeux du TAL/ingénierie linguistique
3. Les deux approches fondamentales

1. Quelques notions de sciences du langage

Propriétés générales

Le langage est le propre de l'homme

- tous les groupes humains découverts à ce jour pratiquent au moins une langue
- il en existe entre 4 000 et 5 000 de par le monde
- l'essentiel d'une langue est acquise vers 5 ans
- les langues parlées par les humains sont les langues naturelles
- langage = capacité de langue

Objet de la linguistique ou "sciences du langage"

- étudier rigoureusement les langues telles qu'elles se parlent
- recherche de points communs entre différentes langues, universaux

1. Quelques notions de sciences du langage

Universaux du LN

La double articulation

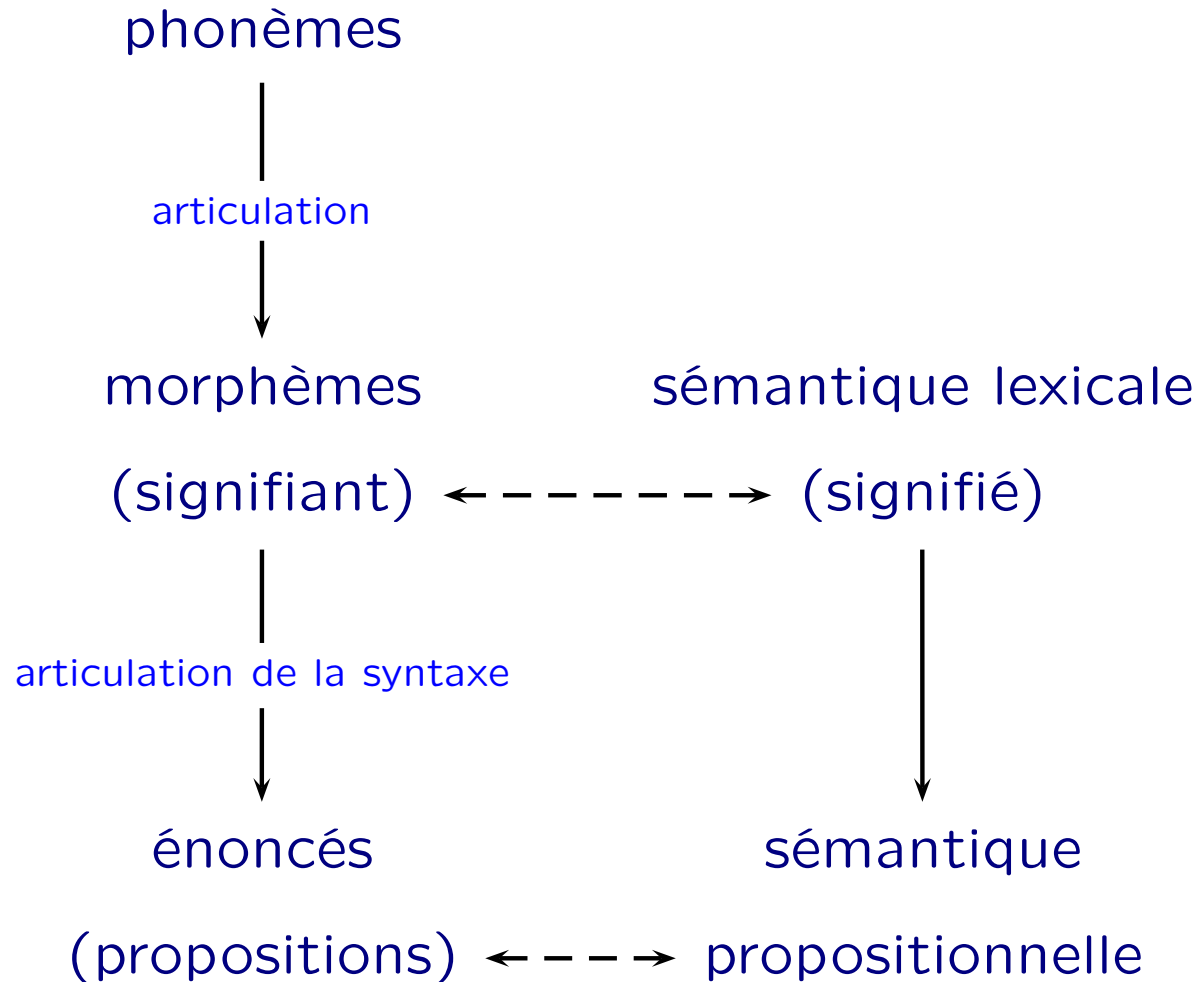
- les langues naturelles sont composées d'unités discrètes
 - phonèmes : classes d'équivalence de sons élémentaires distinctifs (non signifiants)
 - morphèmes : plus petites unités qui ont du sens (mots)
 - propositions : séquences d'unités dotés d'une valeur de vérité (phrases)
- pour passer d'un niveau à un autre : règles de combinaisons
- vrai pour toutes les langues naturelles (y compris langues des signes)
- aucun autre système naturel (communications animales) n'a l'ensemble de ces propriétés
- et en plus (par rapport aux langages de programmation) : grande expressivité du sens !

1. Quelques notions de sciences du langage

Les niveaux d'analyse

niveaux d'analyse

sémantique associée

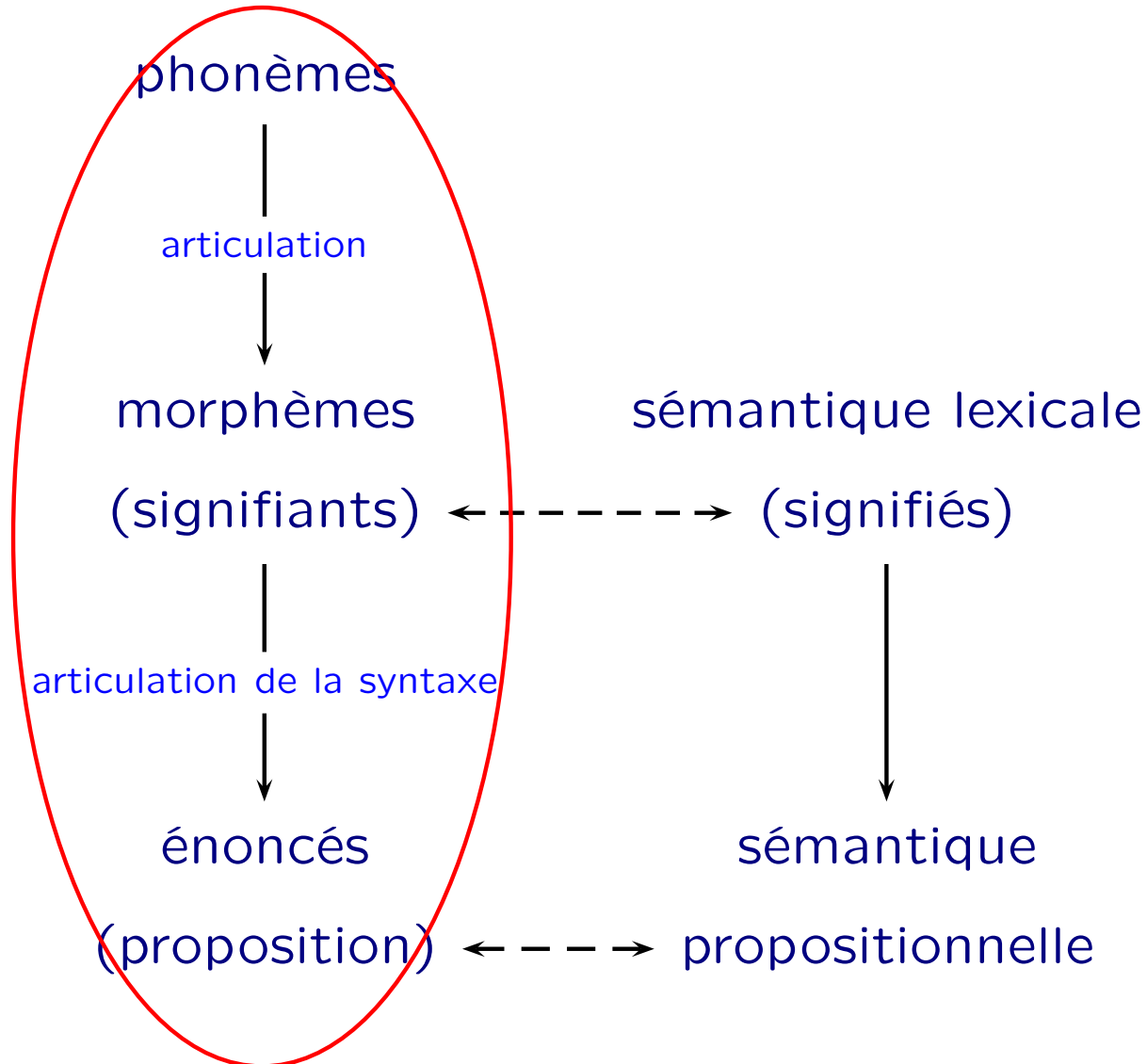


1. Quelques notions de sciences du langage

La double articulation

niveaux d'analyse

sémantique associée

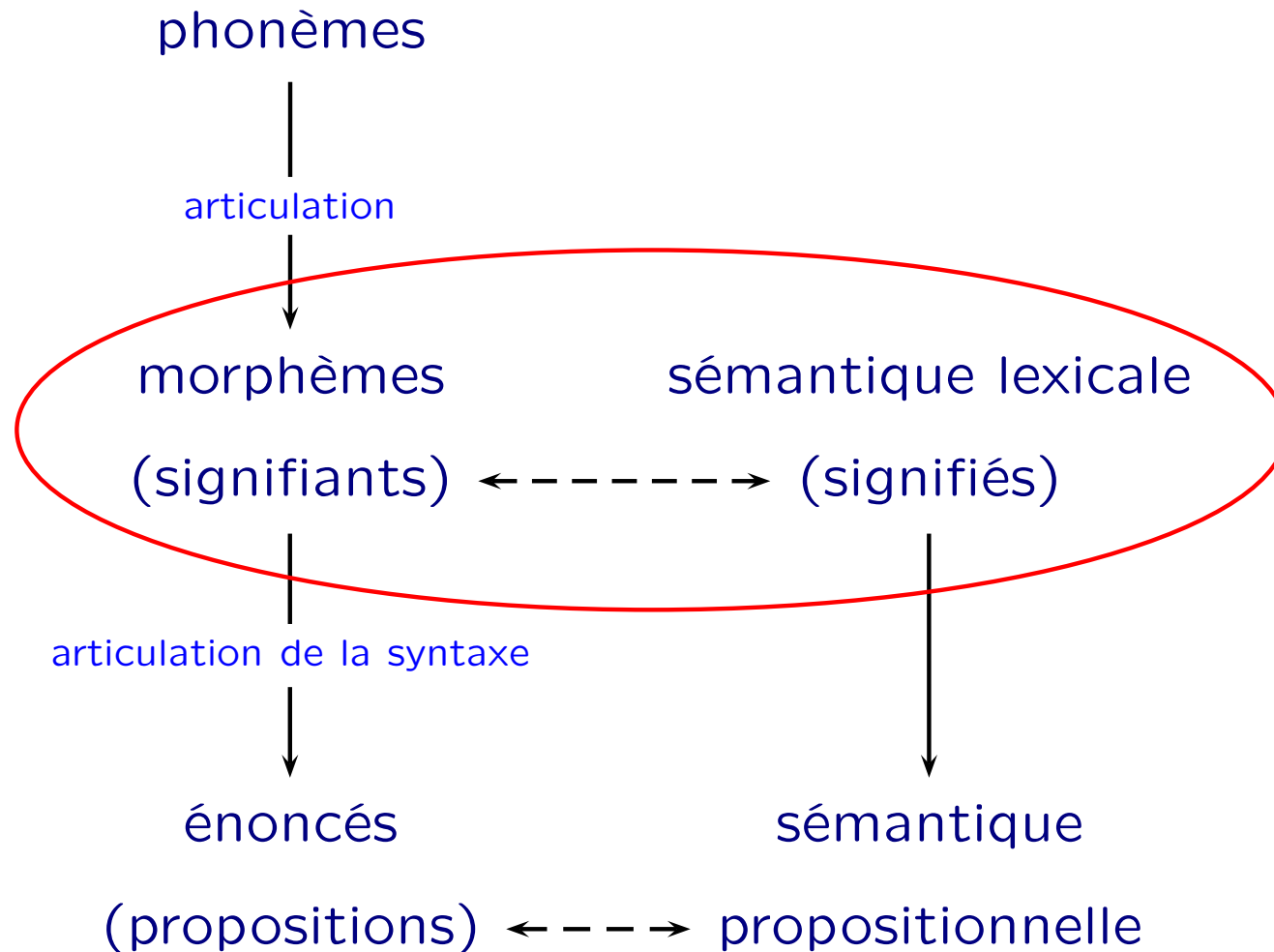


1. Quelques notions de sciences du langage

Le niveau des “protolangages”

niveaux d'analyse

sémantique associée

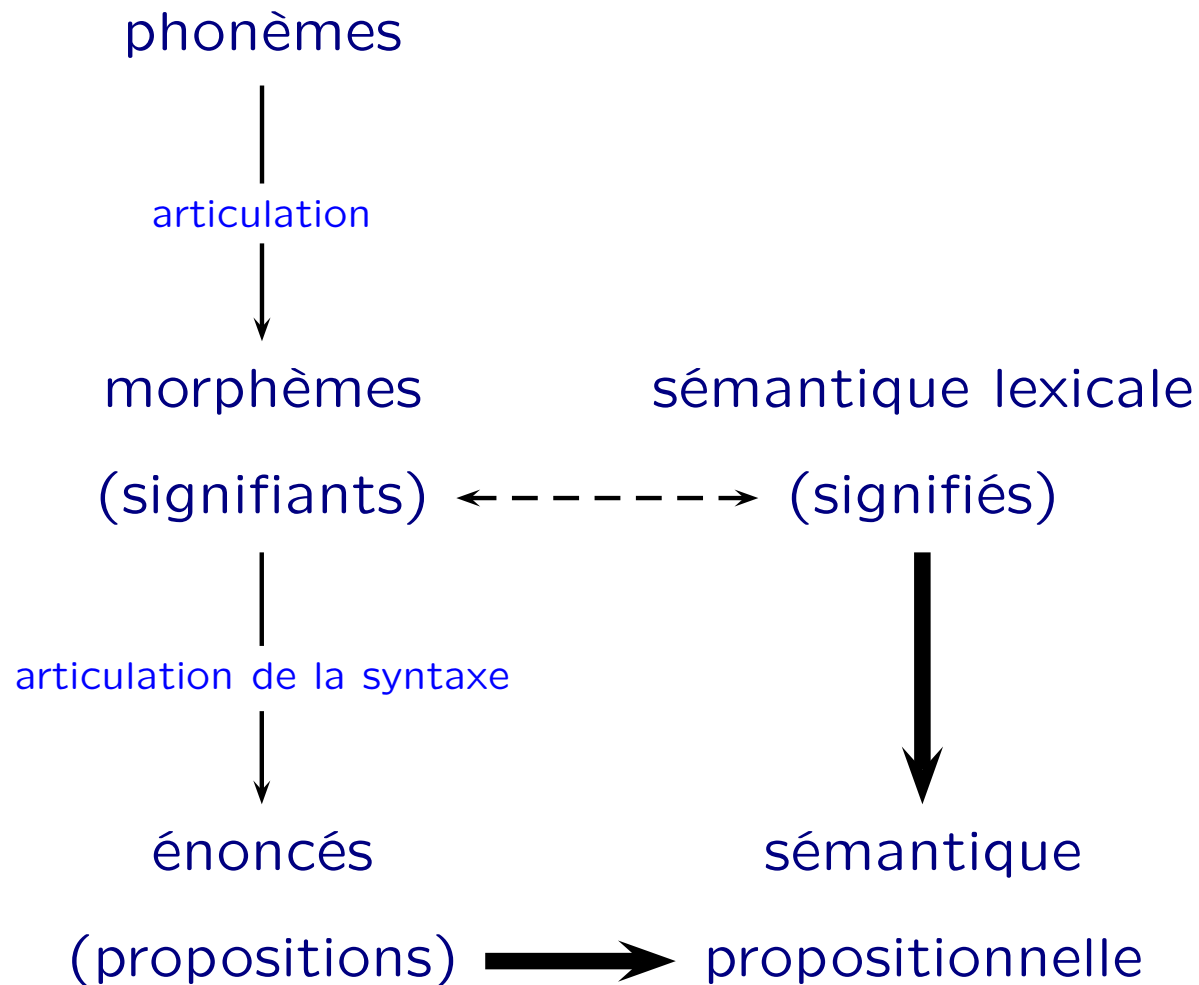


1. Quelques notions de sciences du langage

Le Principe de compositionnalité

niveaux d'analyse

sémantique associée



1. Quelques notions de sciences du langage

Objet du TAL

TAL “traditionnel”

- but : opérationnaliser le domaine
- les différentes données (phonèmes, mots, phrases, “sens”...) doivent être codées informatiquement
- les opérations de manipulations, combinaisons... de ces données doivent être codées sous la forme de programmes
- résultats : étiqueteurs, analyseurs lexicaux, syntaxiques, sémantiques...

Ingénierie linguistique

- but : réaliser une tâche complexe faisant intervenir des données linguistiques (RI, classification de textes, traduction, résumé...)
- mise en oeuvre d’une chaîne de traitements réalisant cette tâche

1. Quelques notions de sciences du langage
2. Applications et enjeux du TAL/ingénierie linguistique
3. Les deux approches fondamentales

2. Applications et enjeux du TAL

Interfaces hommes/machines, mythe de l'IA

Pour accéder à un service

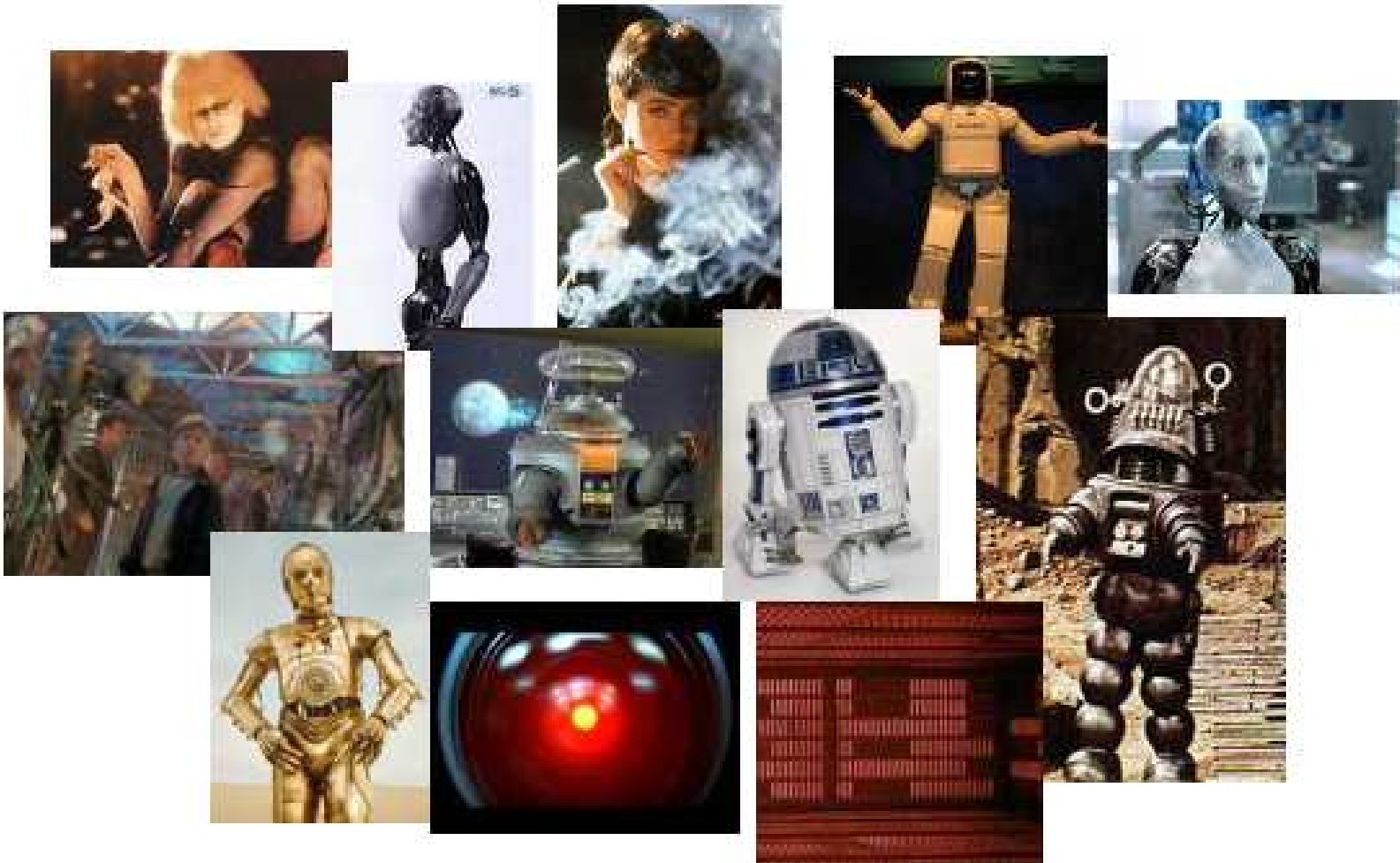
- consultation de BD (annuaires, transport, hébergement, spectacle...)
 - commande vocale (dont aide aux handicapés)
 - dictée/synthèse vocale, traduction instantanée, réalité augmentée
 - utilisation/apprentissage d'une langue étrangère
 - assistant personnel (l'application Siri des iPhones)
- ⇒ dialogues hommes/machines, systèmes question/réponse

Pour le plaisir

- machines intelligentes, robots de compagnie
- jeux (Watson, vainqueur de Jeopardy)

Test de Turing...

2. Applications et enjeux du TAL Robots mythiques du cinéma



2. Applications et enjeux du TAL

Faciliter la production et l'accès à l'information

Aide à la rédaction/lecture

- correction orthographique/grammaticale
- traduction automatique de textes, aide à la traduction
- résumé automatique

Gestion de textes/documents électroniques

- extraction de mots clés, indexation automatique
- classement assisté (mails), analyse d'opinions, recommandation
- extraction des entités nommées (pour synthèse ou anonymisation)
- recherche et extraction d'information, fouille de textes :
 - dans des documents scientifiques (biologie, droit...)
 - dans des documents techniques (documentation, brevets)
 - sur Internet (mails, blogs, forum, réseaux sociaux)

2. Applications et enjeux du TALN

Outils disponibles en ligne

1. Add Documents

Add a file from your computer

Charger le texte ou une autre sélection

Upload File

Add your own text

Add Text

Add a URL

matou

fein

chaton

individu

miriflore

chien

animal familier

chat tigré

chien

fauve

Sep 0c No Dec Jan Feb Mar Apr May Jun Ju Aug

— Cainko — Royal — Bayro — Le Pe

Sallience (%)

0 25 50 75 100

0 5 10 15 20

Display Options

Display edge weight

Display vertex name

Document Text:

Iraqi Vice President Taha Vassin Ramadan

nounced today, Sunday, at Iraq refuses to back wn from its decision to ip cooperating with iarmament inspectors fore its demands are it.

q i Vice president Taha ssin Ramadan nounced today, ursday, that Iraq rejects operating with the

Create Graph...

Create From File...

Traduire avec SYSTRAN BOX

Texte URL

Français Anglais

Le petit chat noir dort

le black kitten

Word to search for: (au)

Search libraries

Display Options: (show more options) (2) (Change)

Key: "S" = Show Synset (semantic) relations, "W" = Show Word (lexical) relations

Noun

- S: (n) cat, (n) chatte (domestic mammal usually having thick soft fur and no ability to roar; domestic cats) **felis**
- S: (n) cat, cat, hombre, homo (an informal term for a youth or man) "is cat you?" "the cat's only dog is he."
- S: (n) cat (a spiteful woman gossip) "what a cat she is!"
 - direct hyperonym / (other) hyperonym / sister term
 - S: (n) gossip, gossiper, gossipmonger, gossipmonger, gossipmonger, gossipmonger (a person of divulging personal information about others)
 - S: (n) woman, adult female (an adult female person (as opposed to a man)) "We women kept he derisively related form."
- S: (n) lar, kat, cat, am, cat, Arabian cat, African cat (the leaves of the shrub Catha adults which are chew make tea; has the effect of a stimulant) "in Yemen cat is used daily by 85% of adults"
 - direct hyperonym / subordinated hyperonym / sister term
- S: (n) cat (a whip with nine knotted ends) "British authors feared the cat"
 - direct hyperonym / subordinated hyperonym / sister term
 - derisively related form
- S: (n) Camptiler, cat (a large tracked vehicle that is propelled by two endless metal belts; frequently used to construct and farm work)
 - direct hyperonym / subordinated hyperonym / sister term
 - derisive usage
- S: (n) big cat, cat (any of several large cats typically able to rear and living in the wild)
- S: (n) computerized tomography, computed tomography, CT, computerized axial tomography, computed ax method of examining body organs by scanning them with X rays and using a computer to construct a series of cross-sections

2. Applications et enjeux du TAL

Enjeux

Domaines concernés par le TAL, enjeux

- politique, économique : surveillance et traduction des échanges (de la guerre froide à PRISM), veille, “industries de la langue”
- juridique : traçage de sources et de plagiats, jurisprudence
- culturel : numériser et exploiter le patrimoine littéraire, préserver la diversité des langues (CE), développer les échanges
- pédagogique : ens. des langues, e-learning, correction automatique
- cognitif : aider à accéder, trier et synthétiser l’information
- sociologique : Web 2.0, analyse d’opinions, tendances, e-réputation
- social : faciliter la mobilité, lutter contre les handicaps
- sociétal : enrichir le débat citoyen (analyse de discours politiques)
- philosophique : les machines peuvent-elles parler, penser ?
- scientifique, technologique : pluridisciplinarité, innovation

2. Applications et enjeux du TAL

Enjeux

Interfaces avec d'autres disciplines

- sciences humaines : sciences du langage, psychologie/pédagogie (acquisition, rééducation après perte...), sciences cognitives, sociologie (analyse des réseaux sociaux via les messages échangés)
- sciences de l'ingénieur : traitement de la parole, automatique, robotique
- informatique : théorie des langages, logique, IA, ingénierie des connaissances (Web sémantique), apprentissage automatique, fouille de données

1. Quelques notions de sciences du langage
2. Applications et enjeux du TAL/ingénierie linguistique
3. Les deux approches fondamentales

3. Les deux approches fondamentales

Approche “symbolique”

- inspirée par un modèle cognitif fonctionnel de l’esprit humain
- influence chomskienne de la primauté de la syntaxe
- formalisation de la compétence par l’intuition
- conception calculatoire du sens (principe de compositionnalité)

Approche “statistique”

- fondée sur le traitement outillé de données attestées
- influence de la linguistique de corpus
- observation rigoureuse de la performance
- conception distributionnelle du sens (dépend du contexte)

3. Les deux approches fondamentales

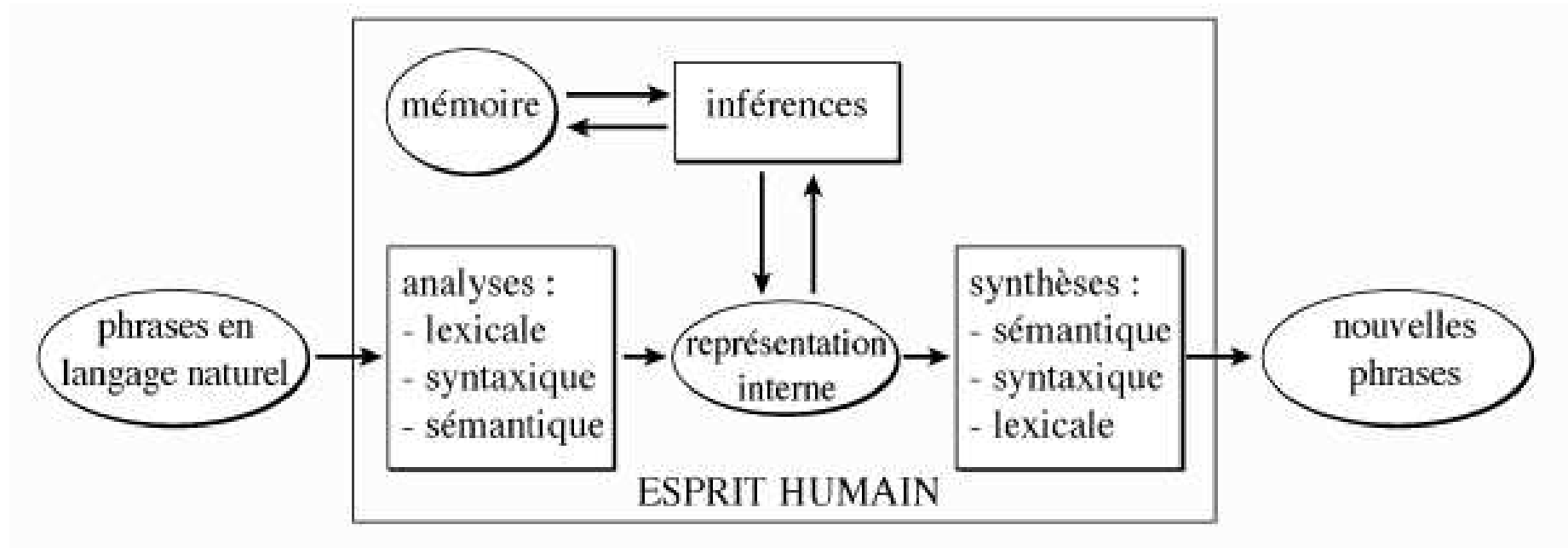
L'approche symbolique

Modélisation symbolique

- “canal historique” du TALN et de l’IA (1950-1990)
- outils utilisés : grammaires formelles, formalismes logiques, combinatoire, mathématiques discrètes
- méthode : écriture à la main de règles ou (rarement) apprentissage symbolique
- inférence déductive, modélisée par la logique
- requiert (souvent) une expertise linguistique
- ce qui est modélisé : le passage d’un niveau d’analyse à un autre,
- traduit l’approche cognitiviste de l’esprit humain

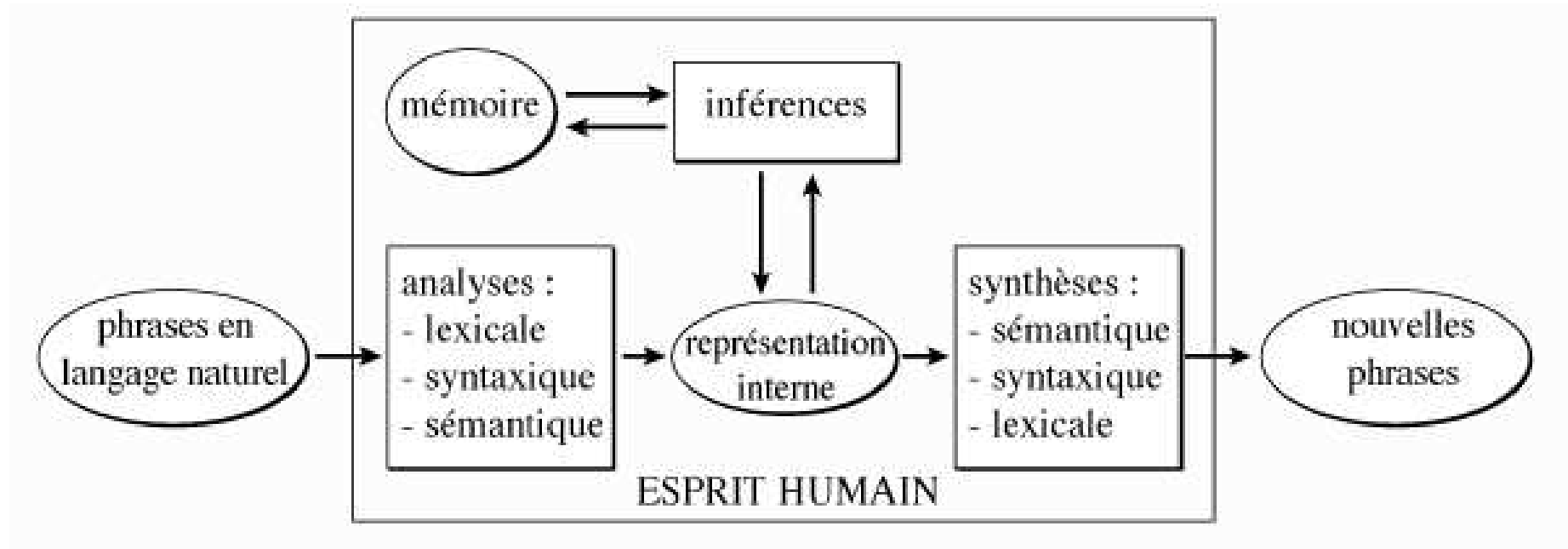
3. Les deux approches fondamentales

Modélisation symbolique de la traduction



3. Les deux approches fondamentales

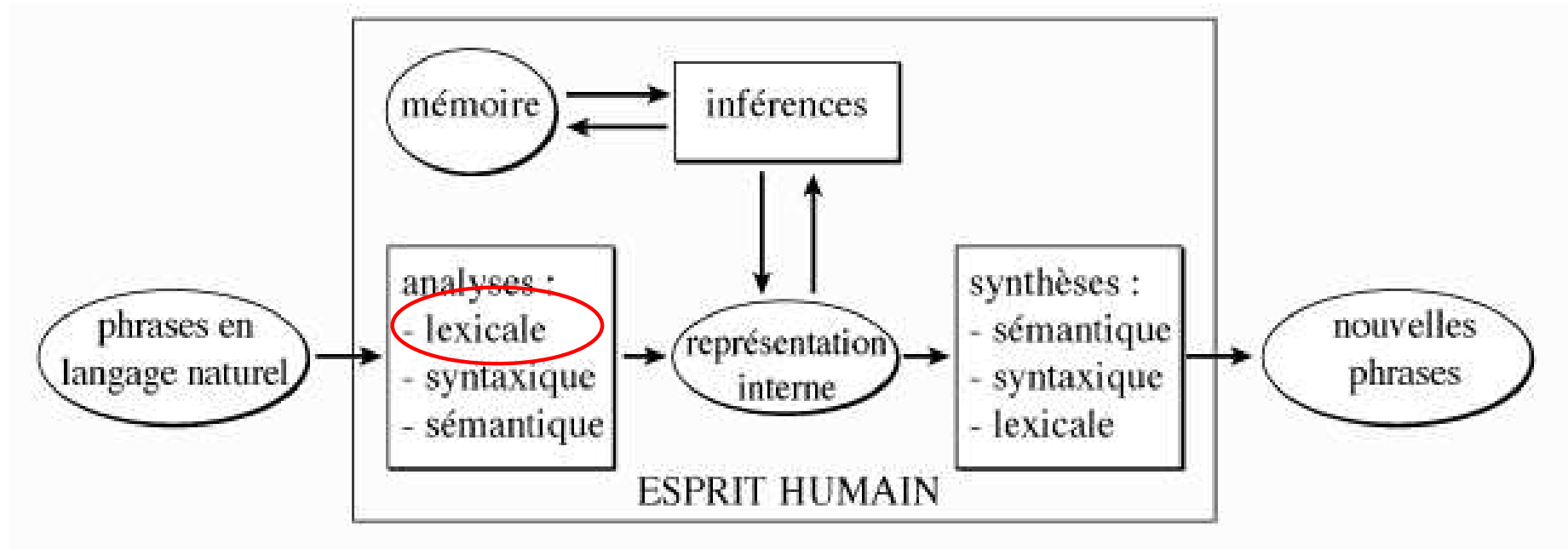
Modélisation symbolique de la traduction



un chat noir dort

3. Les deux approches fondamentales

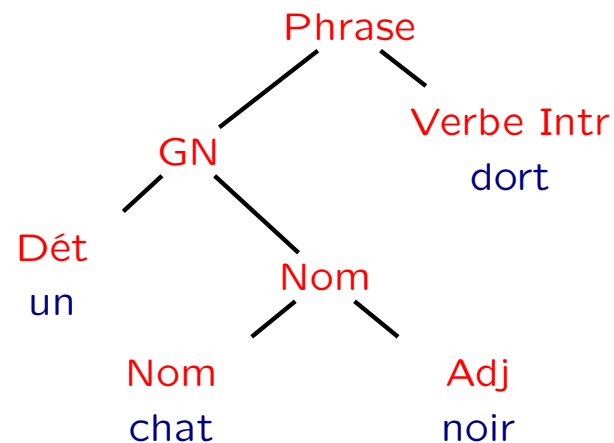
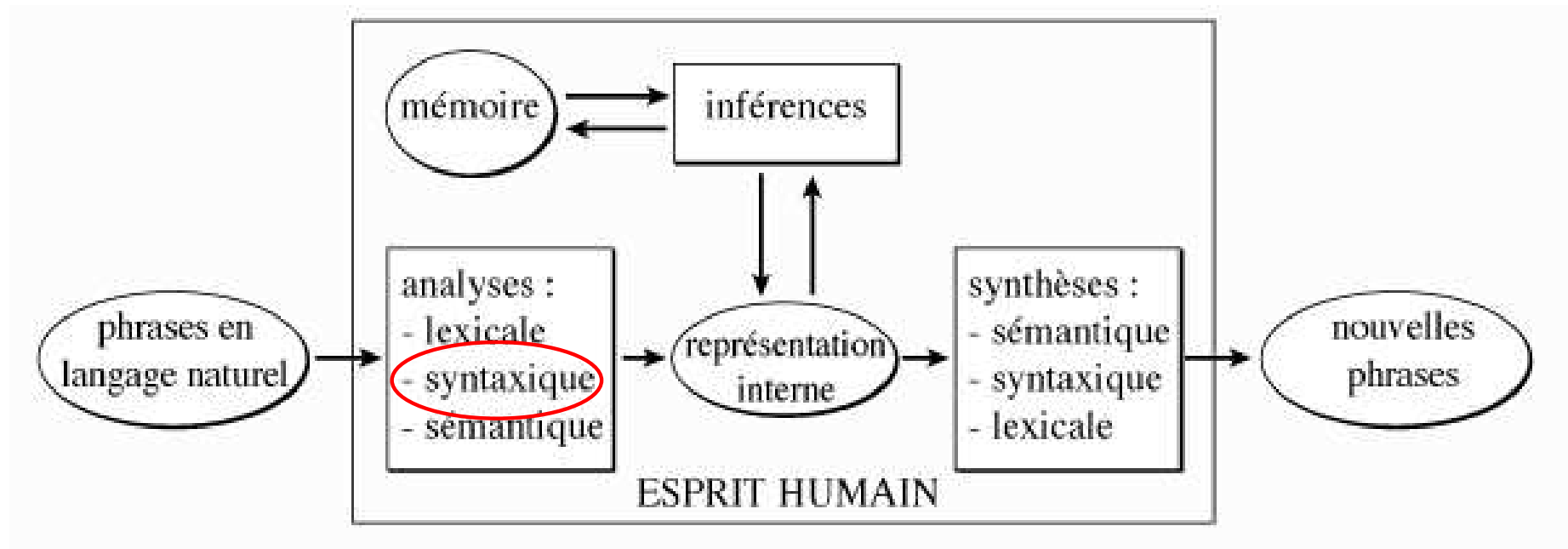
Modélisation symbolique de la traduction



Dét	Nom	Adj	Verbe Intr
un	chat	noir	dort

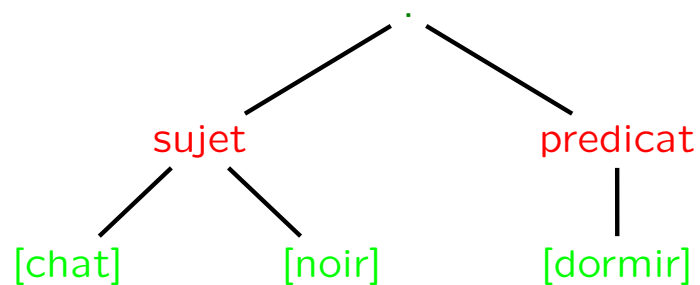
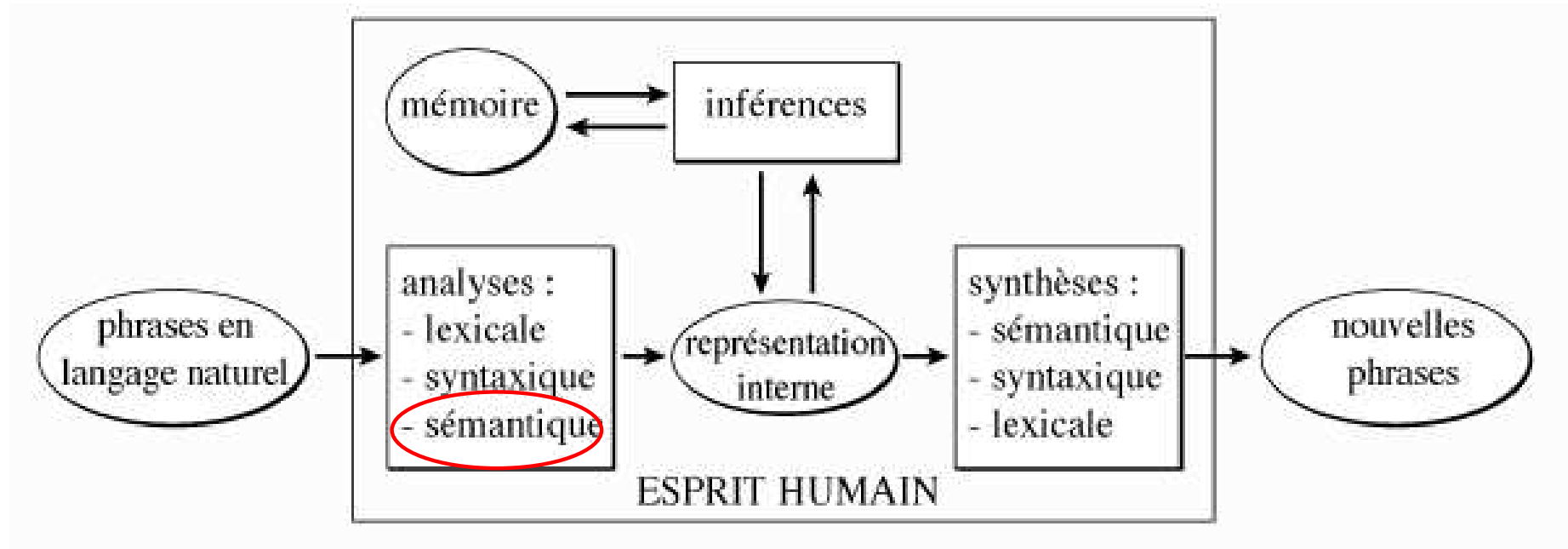
3. Les deux approches fondamentales

Modélisation symbolique de la traduction



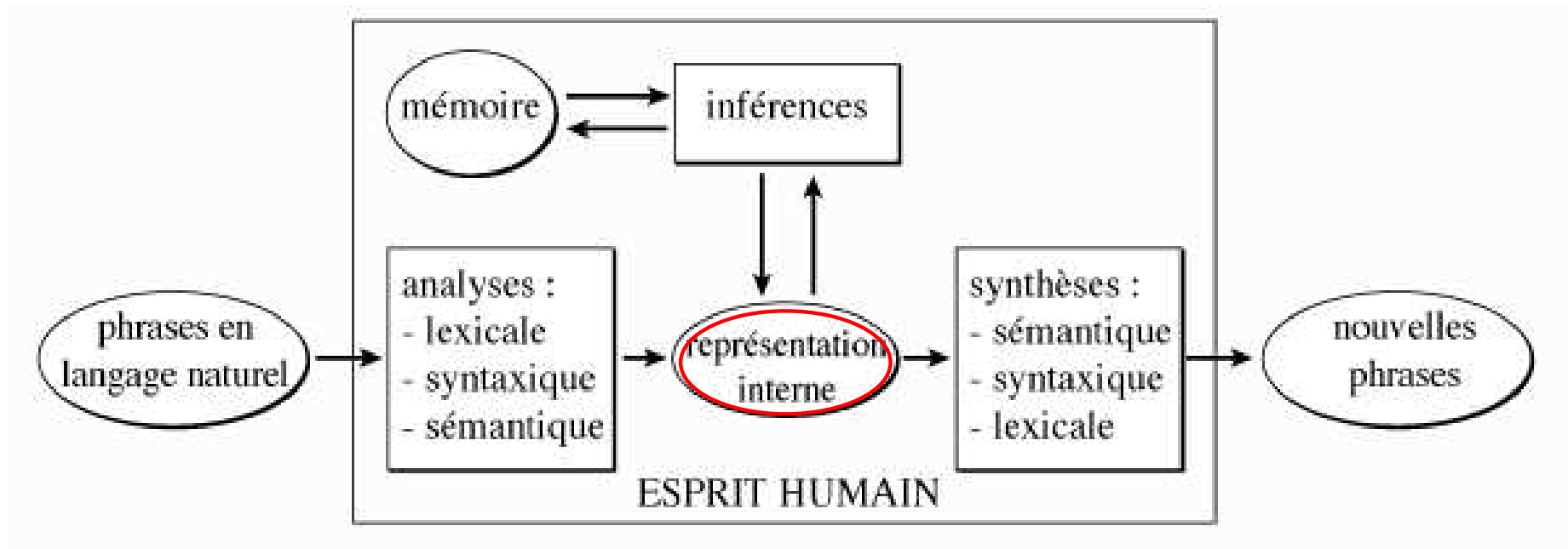
3. Les deux approches fondamentales

Modélisation symbolique de la traduction



3. Les deux approches fondamentales

Modélisation symbolique de la traduction

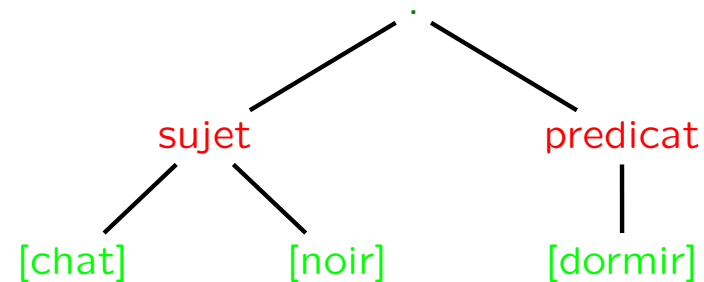
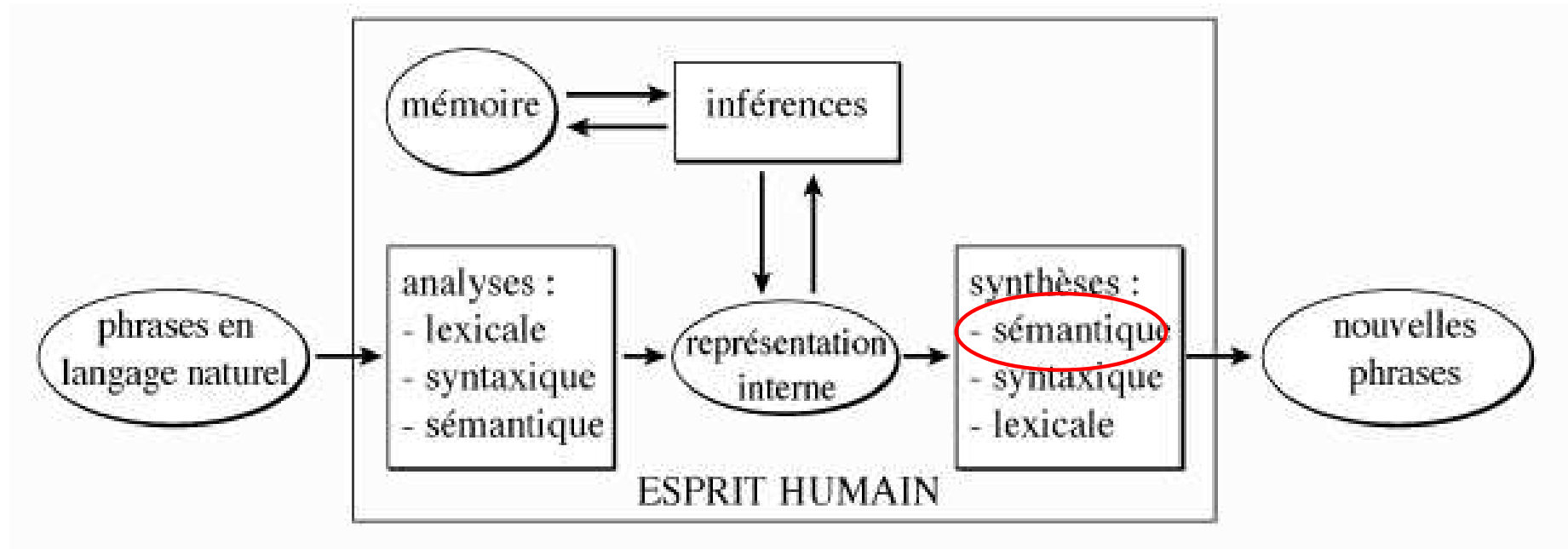


représentation sémantique dans un langage pivot :

$$\exists x[[chat(x) \wedge noir(x)] \wedge dormir(x)]$$

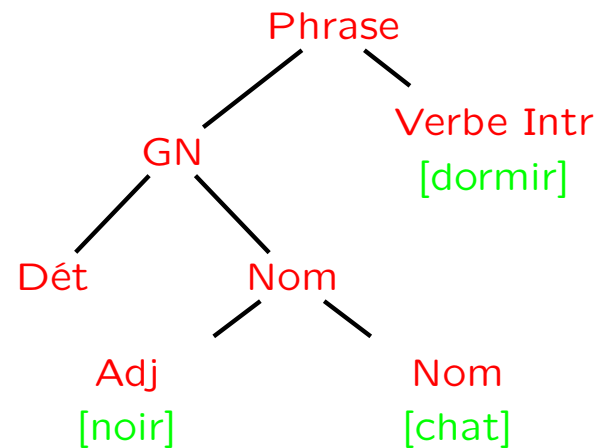
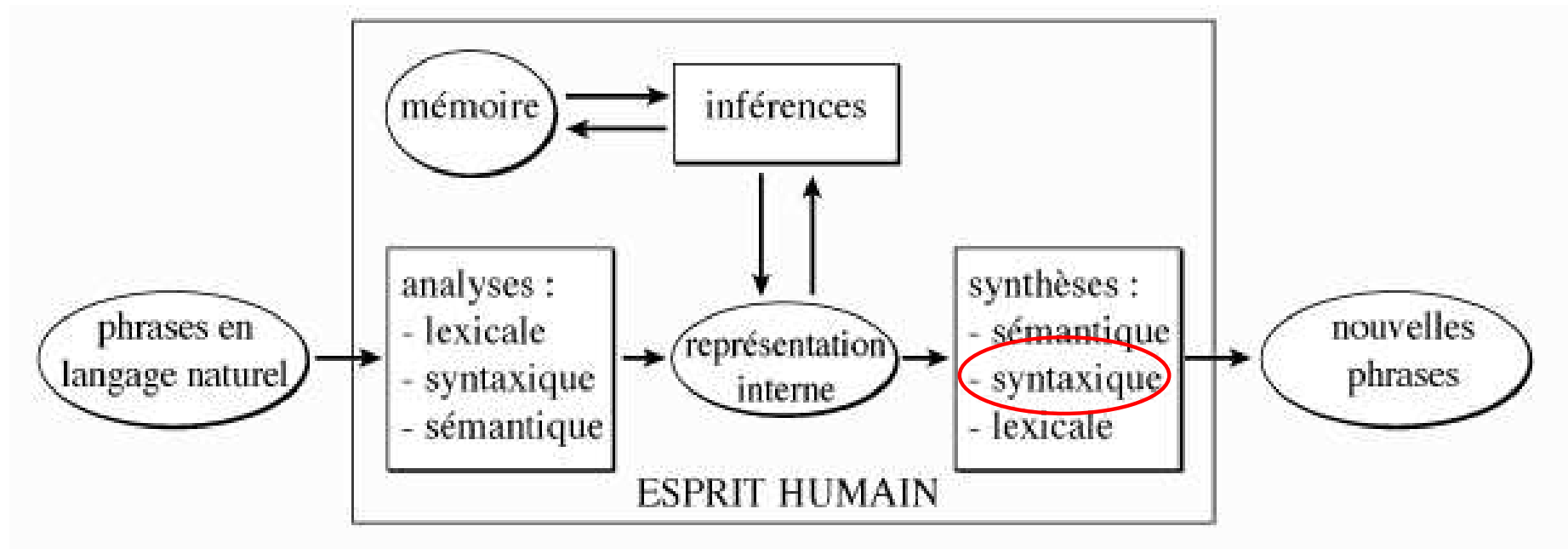
3. Les deux approches fondamentales

Modélisation symbolique de la traduction



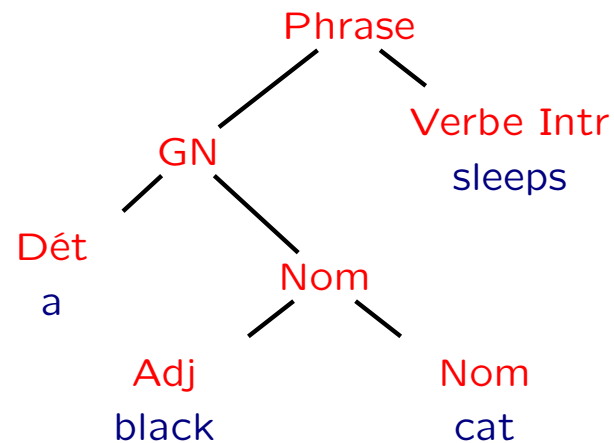
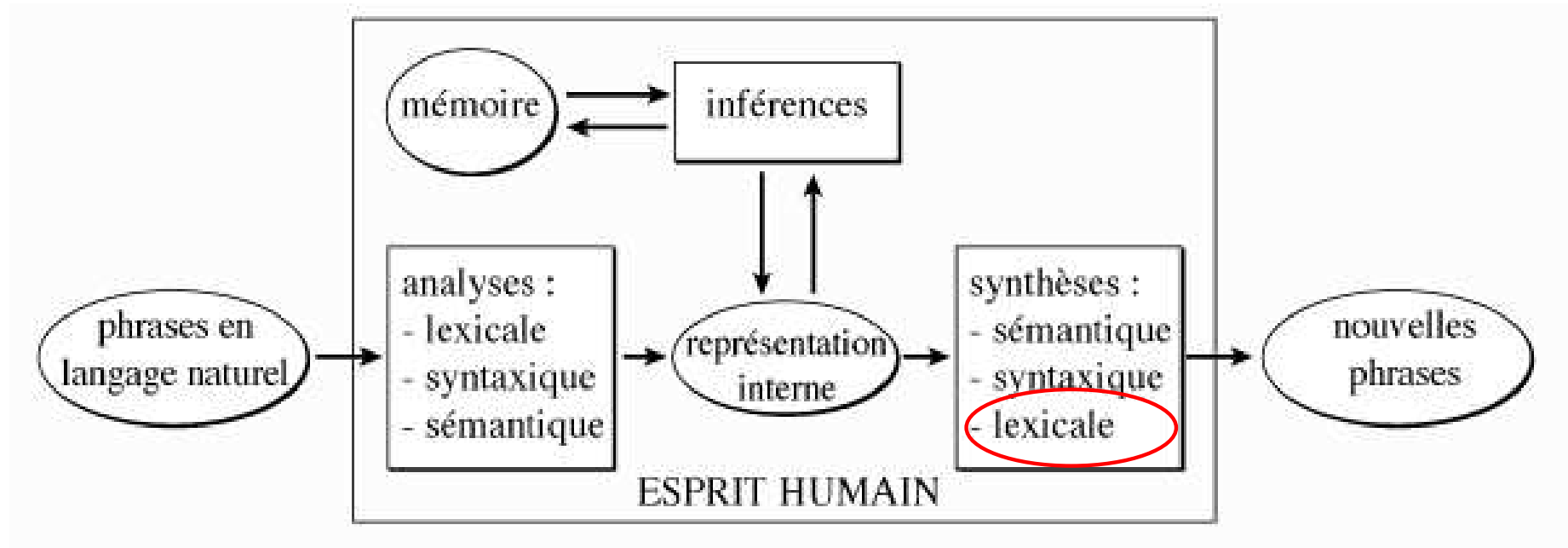
3. Les deux approches fondamentales

Modélisation symbolique de la traduction



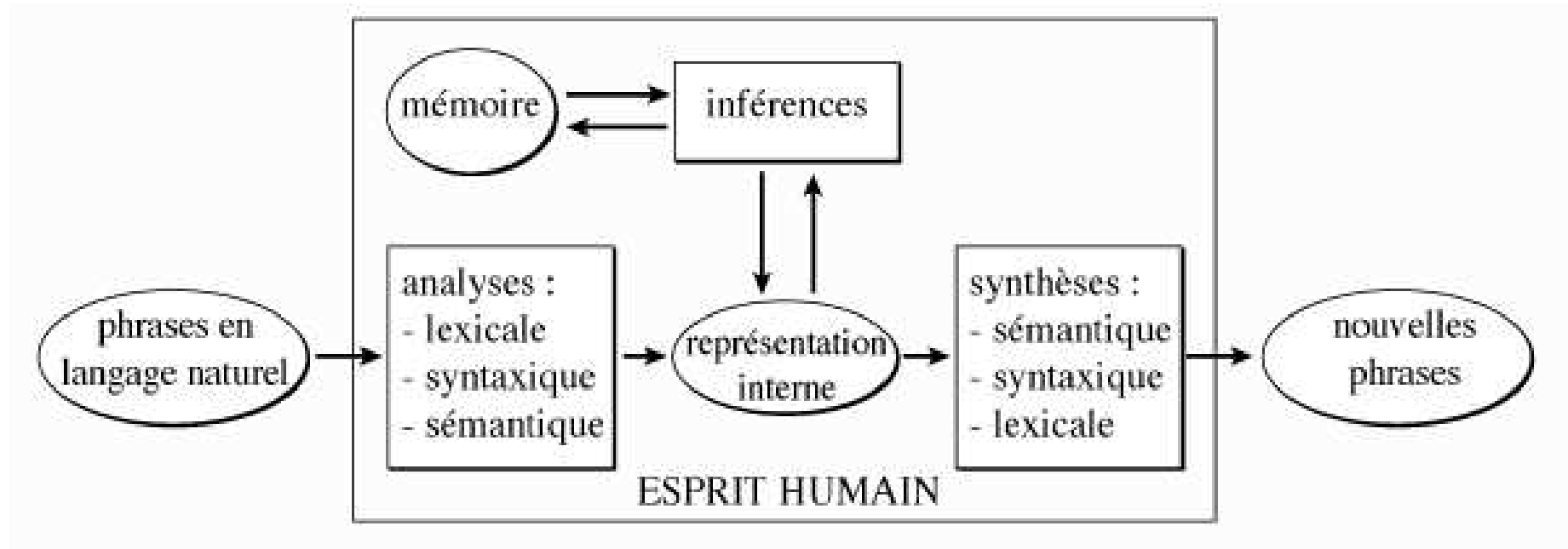
3. Les deux approches fondamentales

Modélisation symbolique de la traduction



3. Les deux approches fondamentales

Modélisation symbolique de la traduction



a black cat sleeps

3. Les deux approches fondamentales

L'approche symbolique

Intérêts

- approche top-down : traitements bien maîtrisés conceptuellement et compréhensibles
- bonne précision : ce qui est traité est bien traité
- déductions logiques puissantes

Inconvénients :

- aucun niveau d'analyse d'aucune langue n'est parfaitement modélisé
- mauvaise couverture : beaucoup de cas non traités
- mise à jour des systèmes laborieuse dès qu'on change de langue, de registre, de thème, de style... (travail de Sisyphe!)
- frame problem : impossible de rendre explicites toutes les connaissances sur le monde

3. Les deux approches fondamentales

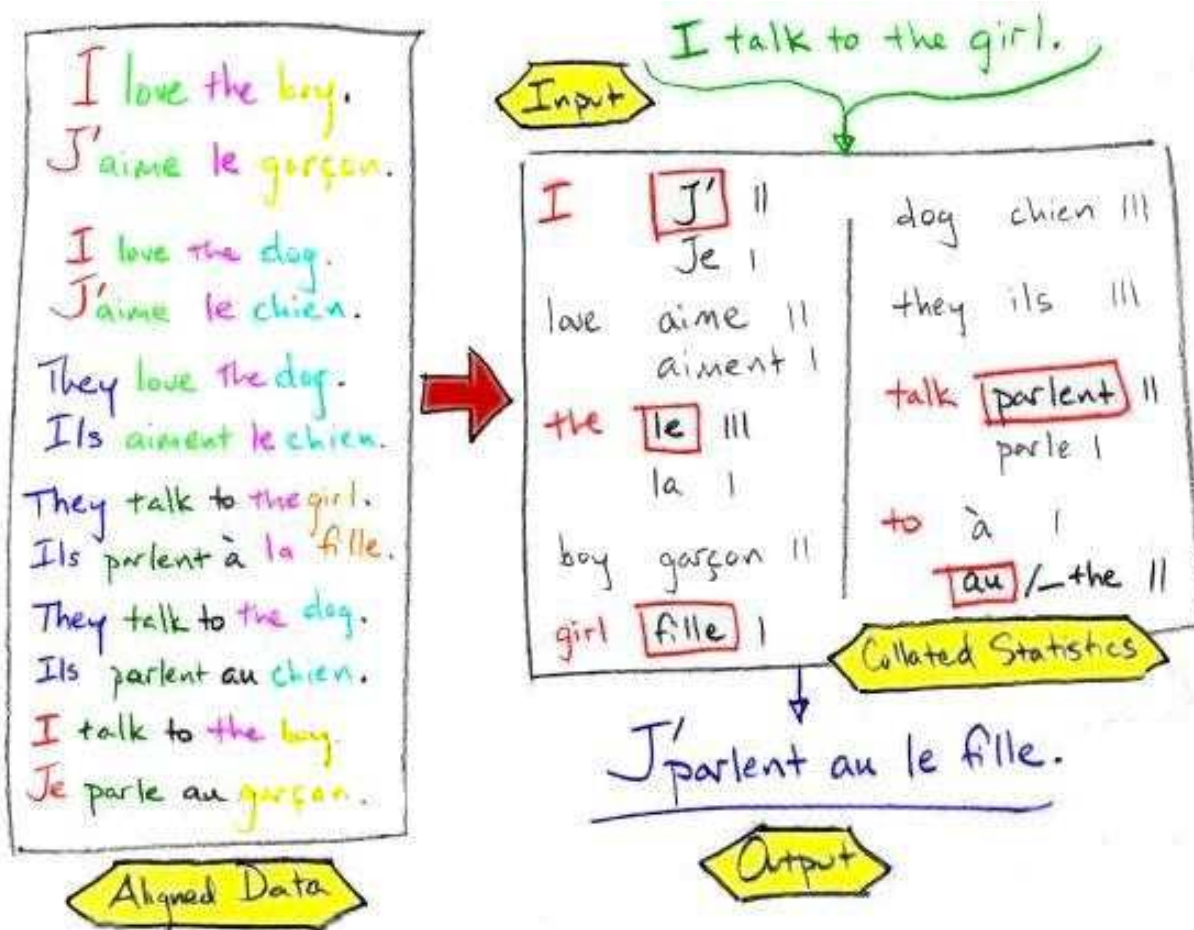
L'approche statistique

Modélisation fondée sur les données

- évolution contemporaine du TALN et de l'IA (depuis 1990)
- rendue possible par l'apparition de machines capables de stocker et traiter de grandes masses de données
- le Web fournit cette grande quantité de textes (Big Data !)
- techniques : analyse numérique, probabilités et statistiques, inférence inductive, mathématiques du continu
- méthodes : transformation des textes en vecteurs, repérage de co-occurrences/corrélations, apprentissage automatique statistique
- ce qui est modélisé : la réalisation d'une tâche à partir d'exemples
- la linguistique devient superflue ? (“every time I fire a linguist, the performance of our speech recognition system goes up”)

3. Les deux approches fondamentales

L'approche statistique de la traduction



3. Les deux approches fondamentales

L'approche statistique

Intérêts :

- approche bottom-up : fondée sur des occurrences attestées plus que sur l'intuition, opérationnalise la "linguistique de corpus"
- les mêmes algorithmes sont applicables quelle que soit la langue (capacité d'apprentissage prime sur érudition)
- plus grande souplesse et adaptabilité, mise à jour facilitée à condition de ré-entraîner
- bonne couverture : plus on a de données, mieux ça marche (Google Translate)

Difficultés, inconvénients :

- difficulté de disposer de bon corpus annotés
- sémantique reste difficile d'accès
- effet boîte noire : interprétation des résultats (erreurs) difficile

Bilan et perspectives

- le TAL est partout !
- nombreux outils et ressources utiles :
 - programmes "grands publics" : moteurs de recherche, classification des mails, traduction, résumé, réponse à des questions...
 - programmes "pour professionnels" : reconnaissance des entités nommées pour la veille, analyse d'opinion, recommandation pour le marketing
- en recherche : deux écoles encore partiellement étanches
- nombreuses passerelles émergentes entre les deux, par ex. : utiliser une ressource symbolique pour enrichir des exemples soumis à un programme statistique...
- il reste beaucoup de travail à faire pour des linguistes-informaticiens