LIDILEM
UNIVERSITÉ Grenoble Alpes

Des techniques complexes (et fragiles) pour des usages simples (et robustes) : le TAL peut-il rencontrer la didactique des langues ?

Olivier Kraif

17 novembre 2017

Introduction

- Le Traitement automatique des langues
  - Des techniques arrivées à maturité
  - Omniprésentes dans les usages quotidiens

Introduction – Des applications arrivées à maturité

Traduction automatique
Sites, informations touristiques, réseaux sociaux, affichage public, etc.

Dialogue et IHM
Synthèse et reconnaissance de la parole, interrogation de base de connaissances de type FAQ, commande vocale, reconnaissance de l'écriture manuscrite, etc.

Aide à la rédaction
Vérification orthographique et grammaticale, complétion prédictive, consultation de références – dictionnaires, encyclopédie, etc.

Recherche d'information
Indexation, catégorisation/classification de documents selon thème, genre, auteur, résumé automatique, fouille de texte, extraction de connaissances, etc.

etc. etc. etc.
Lexicographie, linguistique de corpus, terminologie, etc.

Introduction – applications didactiques ?

- Et dans le domaine de l'enseignement/apprentissage des langues ?
  - L'automatisation pourrait accompagner la recherche d'autonomie ?
    - *mobile learning*
    - *hybrid learning*
    - *blended learning*

Introduction – applications didactiques ?

- Approche « structurale » autonome vs approche communicative en classe ?

In a series of interviews with Spanish and Portuguese instructors at The Ohio State University, we found that instructors perceive the inability of students to handle appropriate linguistic forms as a main obstacle in reaching the communicative goals of meaning-based activities (Amaral, 2011). On the other hand, **the same instructors perceive form-based activities as problematic for use in the classroom** because they can reduce the pace of the lesson and take away time that could be dedicated to meaning-based, communicative activities

Amaral & Meurers, 2011

Introduction – applications didactiques ?

- Un domaine d'application spécifique
  - TAL pour l'apprentissage des langues, NLP for CALL
  - ICALL : Intelligent Computer Assisted Language Learning

Introduction – applications didactiques ?

- Principaux domaines d'application (Antoniadis & Desmet, 2017 [nous traduisons])
  1) Génération de ressources (références) : fouille de corpus, corpus bilingues, interface dictionnaire/corpus
  2) Aide à la lecture (annotations)
  3) Génération d'exercices et de tests
  4) Détection d'erreur et génération de feed-back, évaluation automatique
  5) Aide à la rédaction
  6) Sélection automatique de textes
  7) Adaptation de l'environnement en fonction du modèle d'apprenant

Introduction – applications didactiques ?

- Applications « grand public » difficiles à didactiser :
  - sorties aléatoires (traduction automatique, vérification orthographique, synthèse, etc.)
  - statut secondaire de la forme
  - au mieux, aides ponctuelles pour apprenants avancés

Introduction – applications didactiques ?

- Exemple

  Dans tout le monde, il y a plus que plusieurs langues étrangeres. (corpus Frida)

  → La version Libre-Office de Grammalecte propose de remplacer "étrangeres" par "étrangers" (premier choix).

  - du bruit
  - du silence (niveaux phraséologique, compétence rédactionnelle ?)

Introduction – applications didactiques ?

- De nombreuses recherches, beaucoup de prototypes
- Dans la réalité, des usages assez rares

Introduction

Pourquoi ?

Quels en sont les freins ?

Quelles pistes explorer?

Etude de cas

Etude de cas

- Quelques systèmes «en production»
  - Spanish for Business Professional (Hagen, 1999)
  - Robo-Sensei (Nagata, 2002)
  - E-tutor (Heift, 2003)
  - Alfalex (Verlinde, Selva, Binon, 2003)
  - Tagarella (Amaral, Meurers, 2011)

Etude de cas

- Spanish for Business Professional
  - 12 unités contextualisés avec audio / image
  - Liens vers fiches grammaticales
  - Mots du textes reliés avec entrée de dictionnaire (Lemmatisation)
  - Activités avec *input* contraint : dictée, traduction. Analyse automatique des réponses.

Etude de cas : E-Tutor

- E-tutor (allemand)
  - Activités avec *input* contraint : lacunaires, construction de phrases avec liste de mots, traduction de phrase, dictée. Analyse automatique des réponses et feedback automatique.
  - Construction d'un modèle d'apprenant (compétences grammaticales)

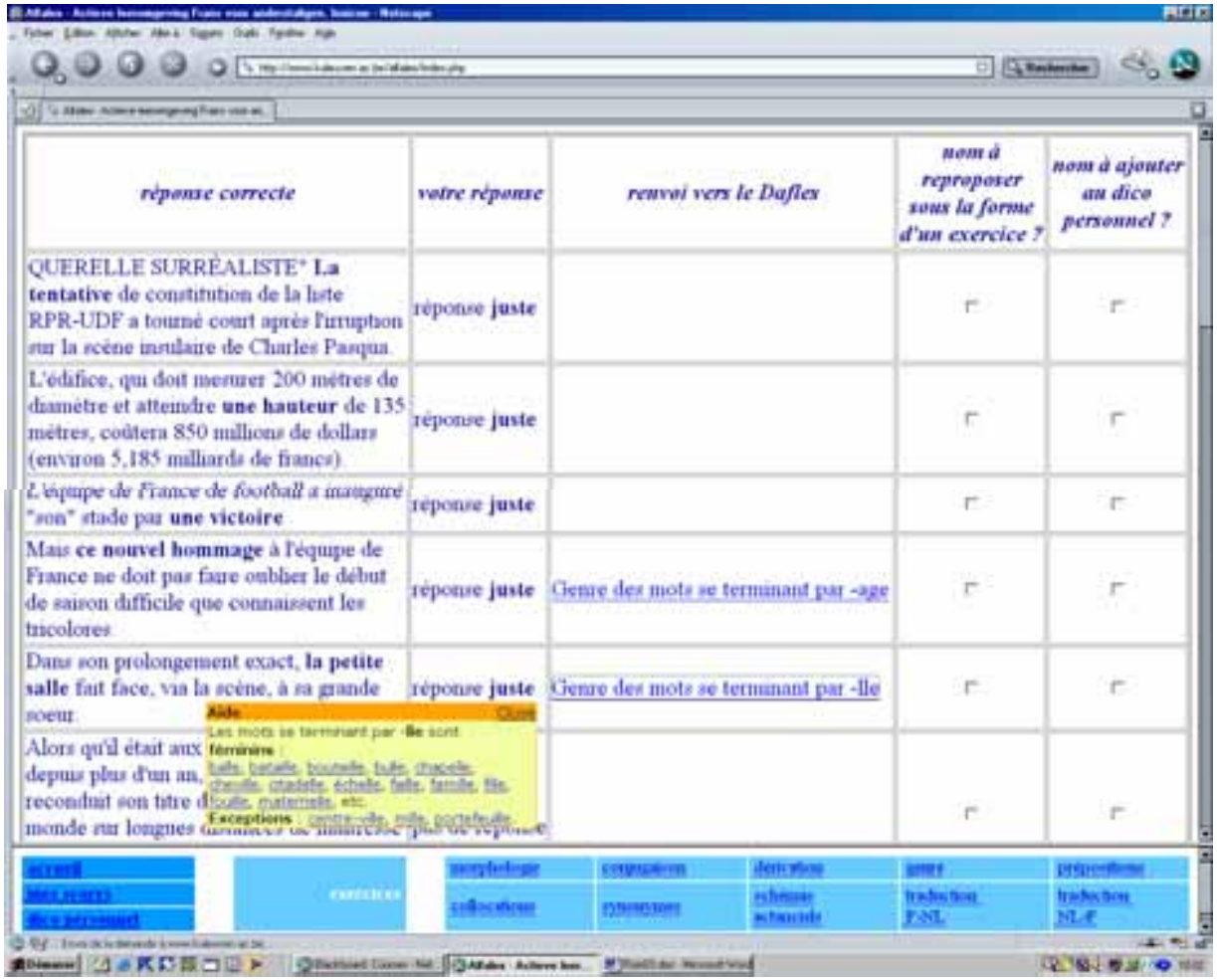
Etude de cas Robo-Sensei

- Robo-Sensei (japonais)
  - Activités contextualisées avec images, et références au manuel.
  - Activités avec *input* contraint : traduction. Analyse automatique des réponses et feedback automatique.

Etude de cas : Alfalex

- Alfalex (français)
  - Activités autour du lexique : morphologie flexionnelle et dérivationnelle, collocations, structure actantielles, traduction.
  - Génération automatique d'activité sur la base d'un dictionnaire informatisé intégrant informations morphosyntaxiques, collocationnelles, sémantiques et d'un corpus annoté (étiqueté, lemmatisé).
  - Activités avec *input* contraint : exercices lacunaires et traduction. Analyse automatique des réponses et feedback automatique.

Etude de cas Alfalex



Etude de cas : Portugais

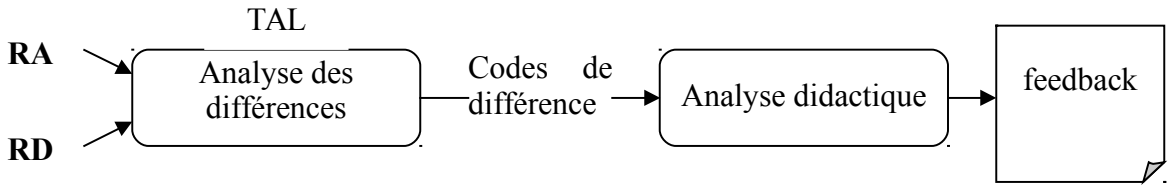
- Tagarella (portugais)
  - Activités contextualisées par documents (écrit, image) : compréhension écrite/orale, description d'image, reformulation, lacunaires et vocabulaire.
  - Réponses ouvertes courtes
  - Analyse automatique des réponses et génération du feedback en fonction du modèle d'apprenant, du modèle d'activité et de l'analyse linguistique

Etude de cas : le prototype ExoGen

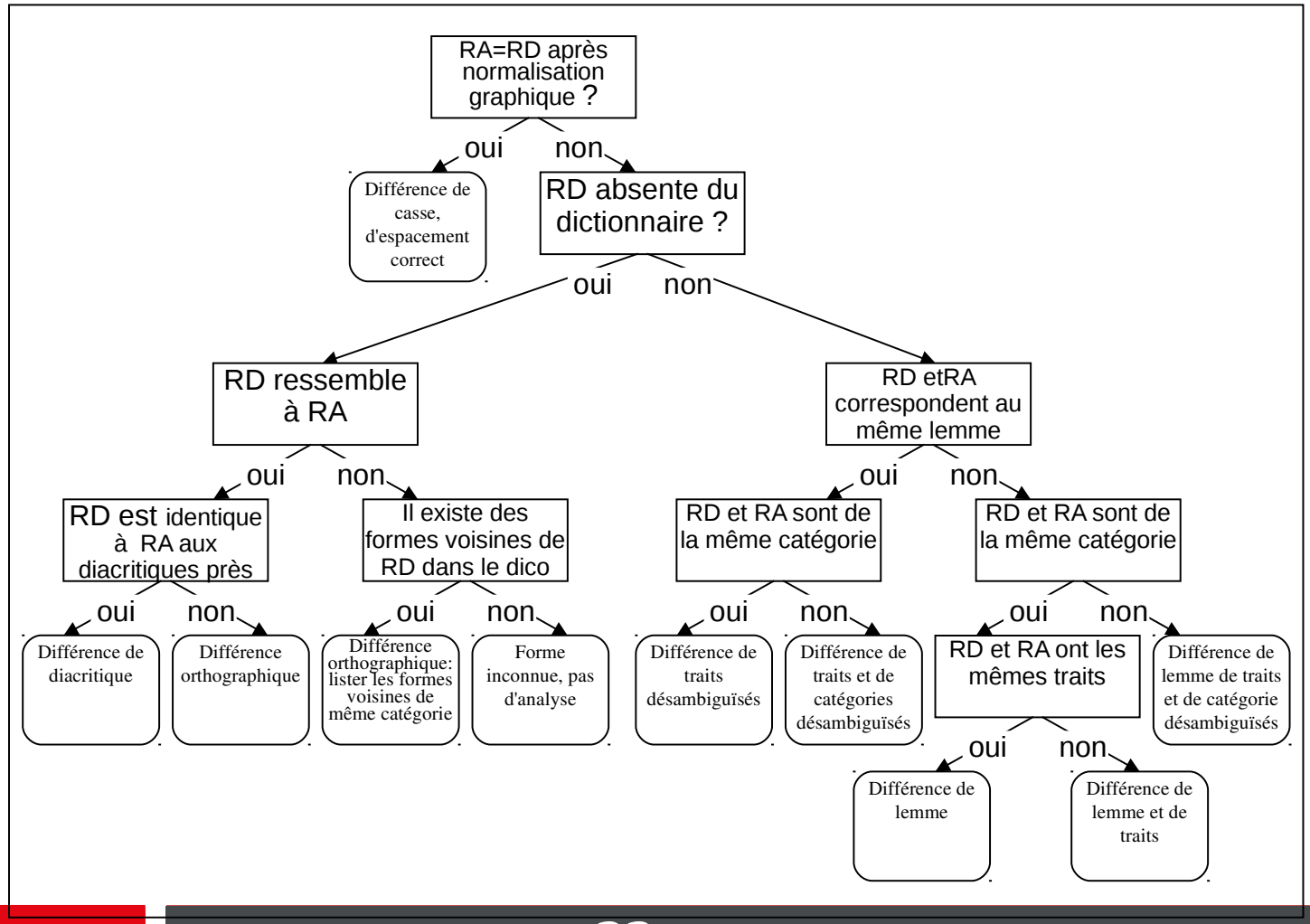
- Un prototype développé par nous (Kraif, Ponton,2007)
  - constat : dans de nombreux QROC, traitement binaire de la réponse (ajout d'un espace = erreur)
  - pourtant l'écart entre réponse attendue (RA) et réponse donnée (RD) est multidimensionnel
    - écart graphique (majuscule, espace, oe, accent, etc.)
    - écart orthographique
    - écart lexical
    - écart morphosyntaxique (genre, nombre, etc.)
    - écart sémantique, etc.

Etude de cas : le prototype ExoGen

- Le prototype ExoGen
  - Génération automatique d'exercices lacunaires à partir d'un corpus – modèle d'Alfalex
  - Critères de sélection variés : lexicaux, morphosyntaxique, etc. (corpus étiqueté)
  - Analyse automatique des réponses guidée par la réponse attendue



Etude de cas : le prototype ExoGen



Etude de cas : le prototype ExoGen

- Avec notre heuristique (moindre différence) très bonne précision (98 % de feedbacks corrects, sur 318 cas d'erreur tirées de productions d'apprenant).

Etude de cas : le prototype ExoGen

Exemple d'erreur	Description (obtenue automatiquement)
(…) avant de **retourner** *(arriver)* en Angleterre.	Forme grammaticalement correcte (verbe infinitif), mais on attendait une autre forme.
et beaucoup d'**échafaide** *(échafaudages)*.	Orthographe erronée ou mot inconnu du dictionnaire.
Je dois me **dépécher** *(dépêcher)*.	Orthographe erronée : problème d'accent.
(…) sommes bien amusées et c'est **vrai** *(juste)* de dire que nous avons dansé assez bien	Forme grammaticalement correcte (adjectif ou adverbe ou nom masculin singulier), mais on attendait une autre forme
C'était **désespéré** *(désespérant)* mais c'était la seule chance (…)	S'il s'agit du verbe *désespérer* : cas 1 [masculin singulier] : On attend un participe présent et non un participe passé.
Pour moi **l'** *(cette)* image crée une ambiance délassante	Forme grammaticalement correcte sur le plan de la catégorie (déterminant), mais on attendait une autre forme avec d'autres traits.
le Premier ministre reste toujours un **britannique** *(Britannique)*	Exact, mais il faut une majuscule à l'initiale.

Etude de cas : le prototype ExoGen

- Suivant le même principe on peut traiter des configurations complexes

  étiquetage + grammaires locales

→ analyse du problème de l'accord du participe passé en français

Discussion

Discussion

- Problème central : contraindre les réponses pour éviter les erreurs d'analyse

  « (…) in order to obtain tractable and reliable NLP supporting the analysis of both form and meaning, it is necessary to restrict the ill-formed and well-formed variation in learner input that an ICALL system needs to deal with »

  Amaral & Meurers, 2011

Discussion

Exercice : Evaluation d'un exercice lacunaire
http://www.didieraccord.com/Accord1/dossier4/html/base.html

Compléter les trous

1. - Je suis très ..... en ce moment.
   - Oh, tu as besoin de ..... Ça te dirait de venir ..... moi en Tunisie ?
   - Mais ce n'est pas trop ..... ?
   - Oh, non ! J'ai trouvé une ..... qui ..... des prix intéressants.
2. - Je pense qu'en Angleterre il ne fait jamais .....
   - Oh non, madame, je peux vous dire que c'est un ....., je ..... bien ce pays et il ne pleut pas toujours.
   - Oui, d'accord, mais je préfère choisir un pays plus ....., j'adore le soleil.
3. - Cet été, nous pourrions faire un ..... en camping-car, non ?
   - Tu sais, Anne, à mon avis, il est préférable de ..... une maison au bord de la .....

Discussion

Exercice : Evaluation d'un exercice lacunaire
http://www.didieraccord.com/Accord1/dossier4/html/base.html

Compléter les trous

1. - Je suis très fatiguée (malade, fatiguée, déprimée, pauvre, éreinté, lassé) en ce moment.
   - Oh, tu as besoin de vacances (repos, te reposer). Ça te dirait de venir avec (chez, sans) moi en Tunisie ?
   - Mais ce n'est pas trop cher (tard, onéreux) ?
   - Oh, non ! J'ai trouvé une agence de voyages (pension, auberge, offre, compagnie, promo, agence) qui propose (pratique, fait, a, offre, permet) des prix intéressants.
2. - Je pense qu'en Angleterre il ne fait jamais beau (chaud, du soleil)
   - Oh non, madame, je peux vous dire que c'est un mensonge (cliché, mythe, canular, préjugé, paradis, rêve, leurre), je connais (supporte, visiterais) bien ce pays et il ne pleut pas toujours.
   - Oui, d'accord, mais je préfère choisir un pays plus ensoleillé (chaud, chaleureux, touristique), j'adore le soleil.
3. - Cet été, nous pourrions faire un circuit (voyage, tour, trip, périple, week-end) en camping-car, non ?
   - Tu sais, Anne, à mon avis, il est préférable de louer (réserver, construire, squatter, trouver) une maison au bord de la mer (lagune, Méditerranée)

Discussion

- Contraindre tout en évitant les activités artificielles (traduction ? dictée ?)
- Privilégier une interaction qui a un sens → approche communicative
- Proposition de Amaral & Meurers : faire en sorte que l'input soit contrainte par l'activité elle-même :

  « It uses pictures, lists of words, contextualized listening passages, gap-filling, written cues in L2, or a combination of two or more of these techniques to constrain the space of potential answers that a learner might provide »

Discussion


THE TAGARELA System
THE OHIO STATE UNIVERSITY
ICALL RESEARCH GROUP
Listening
Reading
Description
Fill-in-Blanks
Rephrasing
Vocabulary
Home
Logout
Módulos: 1 2 3 4 5 Atividades: 1
Leitura
Instrução
Leia o texto e responda às questões usando frases completas e o vocabulário apresentado no texto. Escreva os números por extenso.
Regiões do Brasil
NORTE
NORDESTE
CENTRO OESTE
SUDESTE
SUL
O Brasil está política e geograficamente dividido em cinco regiões. Os limites de cada região (Norte, Nordeste, Sudeste, Sul e Centro-Oeste) coincidem sempre com as fronteiras dos estados que as compõem.
A região Norte ocupa a maior parte do território brasileiro, com uma área que corresponde a 45,27% da área total do País. Formada por sete Estados, tem sua área quase totalmente dominada pela bacia do Rio Amazonas.
A região Nordeste pode ser considerada a mais heterogênea do País. Dividida em quatro grandes zonas - meio-norte, zona da mata, agreste e sertão -, ocupa 18,26% do território nacional e tem nove estados.
O Sudeste é formado por quatro Estados. Esta é a região de maior importância econômica do País, onde está concentrado também o maior índice populacional - 42,63% dos brasileiros.
Já o Sul, região mais fria do País, com ocorrências de geadas e neve, é a que apresenta menor área, ocupando 6,75% do território brasileiro e com apenas três Estados. Os rios que cortam sua área formam a bacia do Paraná em quase toda sua totalidade e são de grande importância para o País, principalmente pelo seu potencial hidrelétrico.
Finalmente, a região Centro-Oeste tem sua área dominada basicamente pelo Planalto Central Brasileiro e pode ser dividida em três porções: maciço goiano-mato-grossense, bacia de sedimentação do Paraná e as depressões. Ela é formada por quatro Estados e nela está a capital do Brasil.
Questões: 1 2 3 4 5 6 7
Questão 1
Próxima Questão (2)
Análise:
Quantas regiões tem o Brasil?
à á ã â é ê í ó õ ô ú ü ç
À Á Ã Â É Ê Í Ó Õ Ô Ú Ü Ç
Enviar
Report Errors & Suggestions

Discussion


THE TAGARELA SYSTEM
THE OHIO STATE UNIVERSITY
ICALL RESEARCH GROUP
Listening
Reading
Description
Fill-in-Blanks
Rephrasing
Vocabulary
Home
Logout
Módulos: 1 2 3 4 5 Atividades: 1
Descrição
Instrução
Descreva a foto usando as palavras apresentadas no exercício e uma das preposições abaixo.
em cima de - entre - embaixo de - ao lado de
Questões: 1 2 3 4
Questão 1
Próxima Questão (2)
Análise:
vaso - mesa
Enviar
Report Errors & Suggestions

Discussion

- Pas de génération automatique d'activité
- La définition adaptative des *feedbacks* est multidimensionnelle :
  - en fonction de l'activité
  - en fonction de l'apprenant
  - en fonction d'éventuelles ambiguïtés
- Paramétrage complexe pour le concepteur de l'activité !
- Système reposant sur une forte intégration didactique et TAL → **interdisciplinarité**

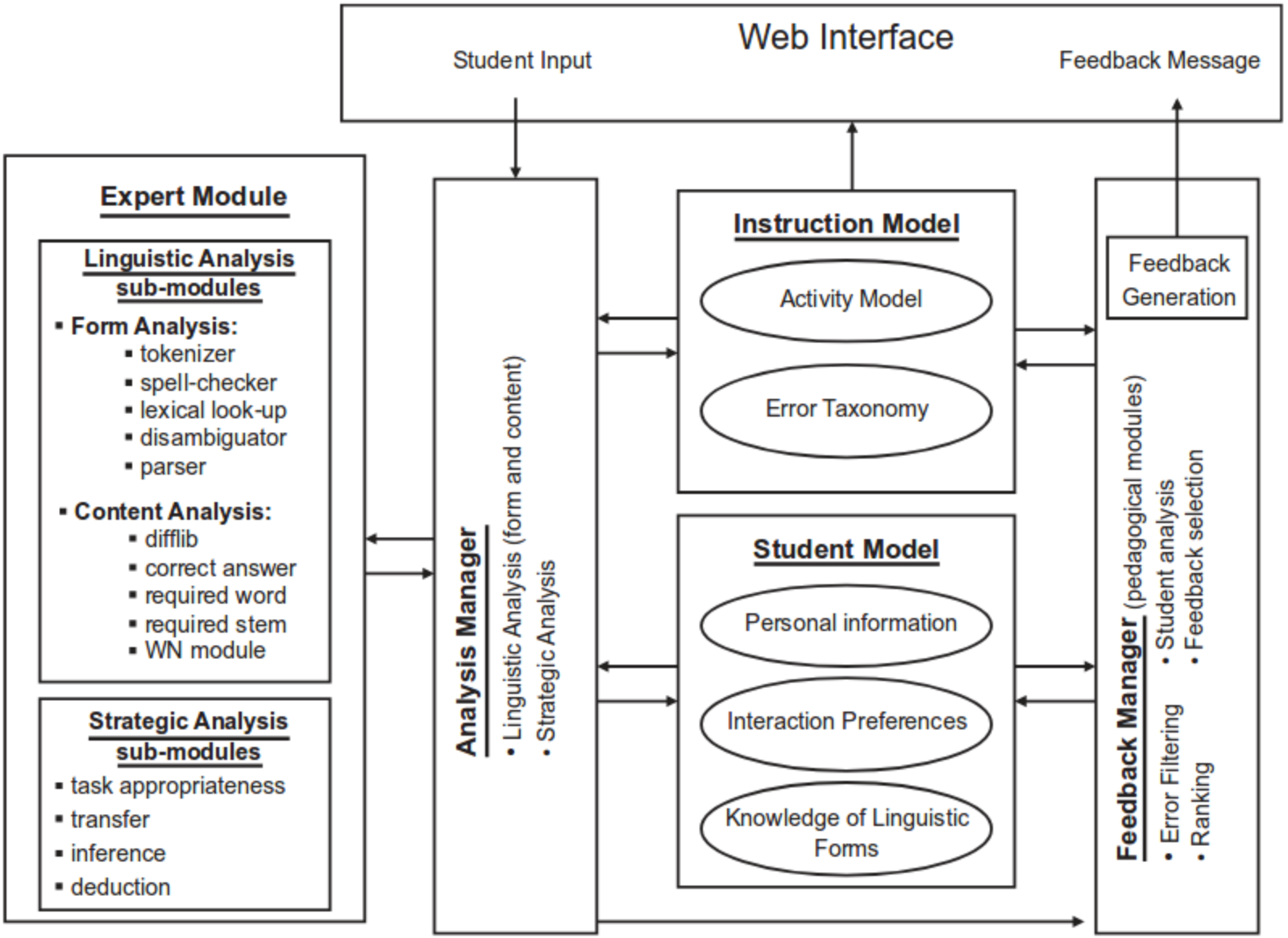


Fig. 4. TAGARELA Architecture

Discussion

- Système *ad hoc* complexe et coûteux à développer
- Perspective d'ExoGen : système auteur générique configurable par enseignant (*Hot Potatoes* avec du TAL!)

  → pas encore à notre portée

Quelques pistes de recherche

Les corpus

- Les corpus d'apprenants
  - Développement d'outils pour l'évaluation / l'annotation
    - Tack et al. (2017) : modèle adaptatif de la compétence lexicale
    - Bestgen Y. (2017) : indices phraséologiques et lexicaux pour l'évaluation de textes en EFL
    - Hamel et al. (2017) : MyAnnotator, outil pour l'annotation dédié à la mise en œuvre de *feedback* de correction d'écrit
  - Le TAL permet d'enrichir ces données pour une meilleure exploitation

Corpus et DDL

- Et le Data Driven Learning (Apprentissage Sur Corpus, Boulton) ?
  - Beaucoup de chemin parcouru depuis Tim Johns
  - Très motivant pour certains publics (Chambers, 2005), qu'il s'agisse de consultation **directe** ou **indirecte**.
  - Adapté dans une perspective de langue de spécialité et d'étude de phénomènes discursifs (Chambers 2010)
  - La méta-analyse de Boulton et Cobb (2017) montre qu'il y a une convergence des études empiriques sur l'efficacité de l'usage des corpus

Corpus et DDL

- Une méta-analyse à grande échelle
  - 64 études empiriques (pré-test/post-test ou groupe de contrôle vs groupe expérimental). 3000 participants.
  - en retenant seulement les études les plus robustes, 70 % ont montré un effet largement positif, 25 % un effet moyen
  - adapté aussi bien pour les niveaux intermédiaires que avancés, pour la langue générale que pour des objectifs spécifiques, pour l'usage sur concordancier aussi bien que sur support imprimé, pour l'apprentissage aussi bien que comme matériel de référence. Spécialement adapté pour le vcabulaire et la lexico-syntaxe.

Corpus et DDL

- Contrairement aux idées reçues, *Data Driven Learning* (DDL) pas limitée à l'étude du lexique et des collocations, pour des apprenants avancés, dans une approche indirecte :

  « There is a corpus revolution underway in both applied linguistics and language instruction (e.g., McCarthy, 2004) and what we have found here suggests that even learners can participate. » (Boulton & Cobb, 2017, p. 41)

Corpus et DDL

- Or le TAL peut apporter une vraie plus-value dans la perspective du DDL
  - Expression de recherche
  - Enrichissement et classement des sorties
- Avec des fonctions génériques (déjà existantes)

Corpus et TAL : étiquetage

- Etiquetage / lemmatisation
  - recherche par lemme (et non forme fléchie) : possible dans la plupart outils :
    - application *standalone* : TXM, AntConc
    - en ligne : BYU COCA/BNC, Frantext, AnaText, AntConc, ...
  - association forme → dictionnaire (aide à la lecture) ou conjugueur
  - recherche de patterns :
    - p.ex. BYU : Verb + Pron + to + Verb : *_v* *_p* to *_v*

Corpus et TAL : étiquetage

- Identification de pattern avec BYU/Coca

when I got to the Wake and Amanda Sam	asked me to buy	her a drink . She drank brandy . A slow si
Michelle Monaghan Table THE REQUEST THE OUT A friend	asks you to help	her move to a new place on your only da
other night after you 'd gone to bed and	wants you to interview	her when the trial 's over ? " " What time
could have him trapped . I just called to	remind you to be	here early because Daddy is pretty upset
even write his own kiss-and-make-up notes but always	asked me to help	him on the sly . This time I could see it
was . You 'll never believe it . He	asked me to marry	him when we were fifteen . His was my fi
and what she could see-and he thought she was	doing it to spite	him . Daniel took to spending more time
1996 he decided to run for state representative and	asked me to be	his campaign manager . No one thought
" " What happened next ? " " He	told me to cook	his supper , " said Hermione . She was ca
When at last the fish stopped , my grandfather	told me to reel	In fast . The dancing rod tip bent out of s

→ Représentation analogique des étiquettes

Corpus et TAL : étiquetage

- Autre exemple : dans la nouvelle *Boule de suif,* de Maupassant, on distingue deux usages du « on » indéfini :
  - désigne les habitants de Dieppe en général, par oppositions aux troupes allemande
  - désigne les occupants de la diligence, par opposition à Boule de suif

Corpus et TAL : étiquetage

- Ces usages du « on » peuvent être contrastés en fonction du temps du verbe qui suit :
  - Imparfait : les habitants de la ville → temps de la description du cadre narratif

    … officier prussien mangeait à table. Il était parfois bien élevé, et ,par politesse, plaignait la France , disait sa répugnance en prenant part à cette guerre . **On** lui était reconnaissant de ce sentiment ; puis **on** pouvait, un jour ou l' autre, avoir besoin de sa protection.
  - Passé simple : le groupe des voyageurs → temps de la narration

    On ne pouvait manger les provisions de cette fille sans lui parler. Donc **on** causa, avec réserve d'abord, puis, comme elle se tenait fort bien , **on** s' abandonna davantage .
- Le deuxième « on » se rapproche d'un « ils ».
- Recherche de pattern avec AnaText :
  - on .*ait_VER vs on .*a_VER

Corpus et Tal : étiquetage

AnaText 2.3

Texte source (utf-8)

Texte étiqueté (utf-8)

- Statistiques générales
- Formes spécifiques
- Tranches de fréquence
- Morphologie verbale
- Noms lemmatisés
- Verbes lemmatisés
- Adjectifs lemmatisés
- Adverbes lemmatisés
- Tous les lemmes
- Toutes les formes
- Concordance
- Cooccurrences

	le coeur des commerçants du pays . Quelques-uns avaient de gros intérêts engagés au Havre que l' armée française occupait , et ils voulurent tenter de gagner ce port en allant par terre à Dieppe où ils s' embarqueraient .		l' influence des officiers allemands dont on avait fait la connaissance , et une autorisation de départ fut obtenue du général en chef . Donc , une grande diligence à quatre chevaux ayant été retenue pour ce voyage , et
9.78%		On employa	
26.87%	pas bien , dit le comte ; comment n' ai -je pas songé à apporter des provisions ? “ Chacun se faisait le même reproche . Cependant , Cornudet avait une gourde pleine de rhum ; il en offrit :	on refusa	froidement . Loiseau seul en accepta deux gouttes , et , lorsqu' il rendit la gourde , il remercia : “ C' est bon tout de même , ça réchauffe , et ça trompe l' appétit . “ L' alcool
30.56%	Elles acceptèrent toutes les deux instantanément , et , sans lever les yeux , se mirent à manger très vite après avoir balbutié des remerciements . Cornudet ne refusa pas non plus les offres de sa	on forma	avec les religieuses une sorte de table en développant des journaux sur les genoux . Les bouches s' ouvraient et se fermaient sans cesse , avalaient , mastiquaient , engloutissaient

Corpus et Tal : analyse syntaxique

- Le recours à l'analyse syntaxique permet d'aller plus loin : extraction des cooccurrents syntaxiques.
  - Sketch Engine : Word Sketches (Kilgarriff, Tugwell, 2002)
  - Lexicoscope : Lexicogrammes (Kraif, Diwersy, 2012)

LIDILEM

UNIVERSITÉ Grenoble Alpes

Corpus et Tal : analyse syntaxique

- lexicogramme de « cap » (presse)

Lexicogramme *Graphiques*

Show 25 entries Search:

l1	l2	f.deprels	f	f1	f2	N	f.disp	am.log.likelihood	r.log.likelihood
cap_NOUN	franchir_VERB	~OBJ ~SUBJ ~DEEPOBJ ~SUBJ_PASSIVE ~VMOD_POSIT2 NMOD_POSIT1 NMOD ~VMOD	248	9852	10838	93405750	3	2186,4975	1
cap_NOUN	changement_NOUN	~NMOD_POSIT1 NMOD_POSIT1	161	9852	24035	93405750	3	1023,3582	2
cap_NOUN	fixer_VERB	~OBJ ~SUBJ ~DEEPOBJ ~SUBJ_PASSIVE NMOD_POSIT1 ~VMOD_POSIT1 ~OBJ_COORD	112	9852	21448	93405750	3	656,3949	3
cap_NOUN	changer_VERB	~VMOD_POSIT1 ~SUBJ ~OBJ	114	9852	44913	93405750	3	508,1684	4
cap_NOUN	nègre_NOUN	NMOD	39	9852	729	93405750	3	410,2801	5

Corpus et Tal : analyse syntaxique

- Cooccurrence de surface (fenêtre fixe) vs cooccurrence syntaxique
  - Moins de bruit : cooccurrence non fortuite
  - Moins de silence : cooccurrence à longue portée

"(...) unlike surface cooccurrence, it does not set an arbitrary distance limit, but at the same time introduces less “noise” than textual cooccurrence" (Evert, 2007) .

Corpus et Tal : analyse syntaxique

Requêtes

- **Requête : <l=cap,c=NOUN,#1> && <l=franchir,c=VERB,#co> :: (.*,1,co)**

Show 10 entries Search:

Identifiant	Contexte gauche	Pivot	Contexte droit
s1019309	Le	cap	des 100 000 migrants secourus devrait être aisément franchi d'ici la fin de l'été.
s1026986	Lors de son exercice 2013-2014, clos fin mai, ses ventes annuelles ont franchi le	cap	des 20 milliards d'euros et son bénéfice a, comme son chiffre d'affaires, bondi de 10 % pour atteindre 2 milliards d'euros.
s1033561	Un réel motif d'inquiétude, alors que la fécondité se maintient à un niveau élevé et que le	cap	du million d'habitants devrait être franchi d'ici dix ou quinze ans.
s103386	L'affaire des matchs truqués dans le foot italien, instruite au palais de justice de Crémone (Lombardie), a franchi un nouveau	cap	.
s105740	Le	cap	des 7 milliards d'individus sur Terre doit être franchi en octobre.
s107282	Le	cap	d'un milliard de SMS devrait avoir été franchi, hier, à l'occasion du rituel des voeux.

Showing 1 to 10 of 211 entries First Previous Next Last

Corpus et Tal : analyse syntaxique

- Exemple de scénarisation (M1 didactique du FLE, C. Cavalla)

Exemples autour de « cap »

Partie A

Lisez ces phrases et répondez aux questions de la « partie B »..

1) Après un bon début de saison, l'entraineur demande à ses joueurs de **maintenir le cap** afin d'être parmi les premiers dans le classement.
2) Malgré ses difficultés, Lucie a réussi à **passer le cap** et elle est maintenant admise en 3e.
3) Cette épreuve représente **un cap** de plus à **franchir** pour Jules.
4) Le commandant de bord nous demande de **maintenir le cap** dans cette direction.
5) Aujourd'hui, j'ai décidé de **franchir un cap** important dans ma vie.
6) En plus de franchir cet obstacle, Lucie a **passé un cap**.

Partie B

1) Selon vous, "cap" a t-il toujours le même sens dans ces expressions ?

2) Quels sont les collocations avec cap les plus couramment utilisées ? Utilisez tous les outils à votre disposition.
[Ils ont des dictionnaires monolingues papier et en ligne.]

Reliez les collocations suivantes à la signification de "cap" correspondante en vous aidant de la partie A] de l'exercice :

a) "maintenir le cap"	1) Une étape
b) "passer un cap"	2) Une direction
c) "franchir un cap"	

Imaginez une phrase avec chacune de ces expressions.

Corpus et Tal : analyse syntaxique

- Travail à la fois sur les collocations, et sur la polysémie de la base
- Les apprenants peuvent utiliser directement le Lexicoscope comme ressource (pas de formalisme à connaître)

Corpus et Tal : analyse syntaxique

- Extraction automatique d'expressions polylexicales (ALR, cf. Kraif, Tutin, Diwersy, 2014, concgram cf. Cheng, Greaves, Warren, 2006)

Corpus et Tal : analyse syntaxique

- Extraction automatique d'expressions polylexicales (ALR, cf. Kraif, Tutin, Diwersy, 2014, concgram cf. Cheng, Greaves, Warren, 2006)

Corpus et Tal : analyse syntaxique

Collocatif changement_NOUN

- Forme canonique : changement de cap (p.ex. : changement de cap) (166)
  - Forme canonique : $UN changement de cap (p.ex. : un changement de cap) (61)
    - Forme canonique : marquer $UN changement de cap (p.ex. : marquant un changement de cap) (3)
    - Forme canonique : $UN changement de cap qui (p.ex. : un changement de cap qui) (5)
  - Forme canonique : ce changement de cap (p.ex. : ce changement de cap) (24)
    - Forme canonique : justifier ce changement de cap (p.ex. : justifier ce changement de cap) (3)
  - Forme canonique : réclamer changement de cap (p.ex. : réclame changement de cap) (4)
- Forme canonique : impliquer changement cap (p.ex. : implique changement cap) (3)

Relation ~OBJ

Collocatif franchir_VERB

- Forme canonique : franchir cap (p.ex. : franchi cap) (183)
  - Forme canonique : avoir franchir $UN cap (p.ex. : a franchi un cap) (30)
    - Forme canonique : avoir franchir $UN nouveau cap (p.ex. : ont franchi un nouveau cap) (6)
    - Forme canonique : il avoir franchir $UN cap (p.ex. : ils ont franchi un cap) (5)

Corpus et Tal : analyse syntaxique

- Entrée vers la phraséologie au sens étendu
  - pas seulement collocation
  - notion de genre : « le cap a été franchi » exclusivement dans la presse
  - étude des marqueurs discursifs (routines sémantico-rhétoriques, cf. Tutin et Kraif, 2015)

Corpus et Tal : problèmes

- Deux problèmes se posent :
  - langage de requête :
    - POS TAG : *_v* *_p* to *_v* (BYU/Coca)
    - Dépendances : <c=VERB,#1>&&<c=PRON,#2>&&<l=to,c=PREP,#3>&&<c=VERB,#4>::(EMBED_INFINIT,4,1) (OBJ_POST,1,2) (PREPD,4,3) (SUBJ_PRE_NFINIT,4,2)
  - que faire des erreurs d'étiquetage ?

Corpus et Tal : les pistes de la RI

- Des solutions issues de la Recherche d'Information
  - recherche simple sans formalisme
  - pas d'annotation dans les résultats (le moteur TAL reste "sous le capot")
  - affinage de la recherche : reformulation de requête (lemmatisation, généralisation) / retour de pertinence (requêtes similaires)
  - tri des résultats par pertinence

Corpus et Tal : la piste analogique

- Placer l'**analogie** au centre du processus de recherche
  - requêtes basées sur l'exemple :
    - GrETEL et Poly-GrETEL (Augustinus et al., 2012, 2016)
    - Lexicoscope (Kraif, 2016)

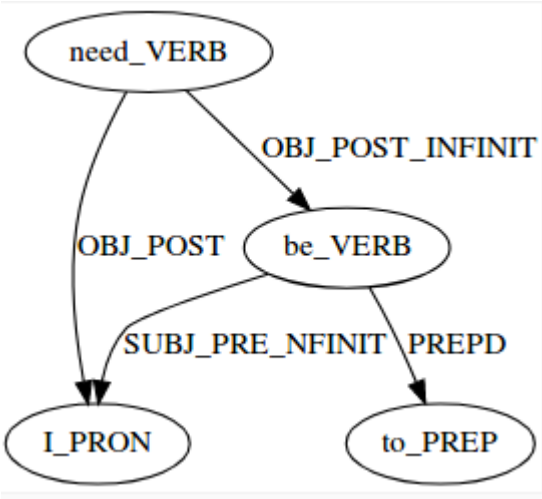
      → requêtage de corpus arborés

      → formulation des requêtes à partir d'exemples

Corpus et Tal : requêtes analogues

- Exemple : Recherche de la construction *to need* + Pronom + *to* + Verb
  - input :

    "need me to be"
  - trouvé :



<l=need,c=VERB,#1>&&<l=I,c=PRON,#2>&&<l=to,c=PREP,#3>&&
<l=be,c=VERB,#4>::(OBJ_POST,1,2) (OBJ_POST_INFINIT,1,4)
(PREPD,4,3) (SUBJ_PRE_NFINIT,4,2)

Corpus et Tal : requêtes analogues

- Généralisation de la requête :

<l=need,c=VERB,#1>&&<l=I,c=PRON,#2>&&<l=to,c=PREP,#3>&&<l=be,c=VERB,#4>::
(OBJ_POST,1,2) (OBJ_POST_INFINIT,1,4) (PREPD,4,3)
(SUBJ_PRE_NFINIT,4,2)

Corpus et Tal : requêtes analogues

- Généralisation de la requête :

<l=need,c=VERB,#1>&&<l=I,c=PRON,#2>&&<l=to,c=PREP,#3>&&<l=be,c=VERB,#4>::
(OBJ_POST,1,2) (OBJ_POST_INFINIT,1,4) (PREPD,4,3)
(SUBJ_PRE_NFINIT,4,2)

Corpus et Tal : requêtes analogues

- Résultats

Identifiant	Contexte gauche	Pivot	Contexte droit
s1047	"I	need	someone like you to bring a bit of chaos into my life." He stood up with the bowl.
s1218	I'm going to give you an Internet link and I	need	you to check it daily.
s1379	Jane; the Whore of Babylon's alternate self is being held in a night club, a place of dancing, I	need	you to enter the front of the building and do not be distracted by the strange things you see, an army will be waiting in there.
s1488	Perhaps he just	needed	someone to aim the words at.\n
s164	"I	need	you to get an army together, to invade heaven, a rather satanic army then, and here would be a good place to start, many of these people will be dead in a few years, and believe me they're going to be pissed."\n
s1878	"I don't	need	anyone to help me upstairs, but I'd like you to come with me and I'm not going back to bed.
s1976	He	needed	somebody to talk to, not about his failure perhaps, but somebody with whom he could exchange words to show him that the universe had not combined in enmity against him.
s2076	She	needs	me to write her life.
s2195	I realise you've got a parent to attend to, but it doesn't	need	two of you to sit with him and hold his hand, and in any case I imagine by this time he's settled down for the night?
s2315	He	needed	me to get it out of him.

Corpus et Tal : requêtes analogues… et résultats similaires

- Thèse d'Ilaine Wang (2017) :
  - recherche par similarité syntaxique pour l'apprentissage des langues
    - recherche de constructions grammaticales – similarité syntaxique / lexicale / dissimilarité
    - catégorisation automatique des résultats par similarité
    - tri des résultats par pertinence

Corpus et Tal : requêtes analogues… et résultats similaires

- Exemple de requête (*past perfect continuous*) :
  - I ’d been working hard all day
- Résultats par ordre de pertinence
  1) I’ve been watching you
  2) I’ve been looking for you everywhere
  3) I’ve been searching for ages
  4) He had been correct about the pain

Corpus et Tal : requêtes analogues… et résultats similaires

- La notion de pertinence (sémantique, lexicale, grammaticale) décrit un *continuum*
  - Pas de jugement binaire : correct/incorrect
  - Le rapport à l'erreur d'annotation change : l'utilisateur sait qu'il y a du bruit
  - Il n'a pas d'accès direct aux annotations

Conclusion

Conclusion

- Les applications TAL en didactique sont
  - complexes (pas d'outils auteurs clés en main – storyline, hotpot, etc.)
  - fragiles (du bruit dans les résultats)

    → coûteuses à mettre en œuvre (pluridisciplinarité forte)

Conclusion

- L'analogie : une piste encore peu exploitée et pourtant...
  - au centre de l'acquisition du langage
    - paradigmes flexionnels : mangeons/mangez, faisons/*faisez
  - au centre des descriptions linguistiques :
    - test de commutation → établissements de rapports d'analogie
    - Corpus Pattern Analysis (Hanks)
  - au centre du *Data Driven Learning*
    - série d'exemples analogues (lexique, grammaire, etc.)
    - feedbacks analogues (ex. Boulton PERL 2017, requête Coca)

Conclusion

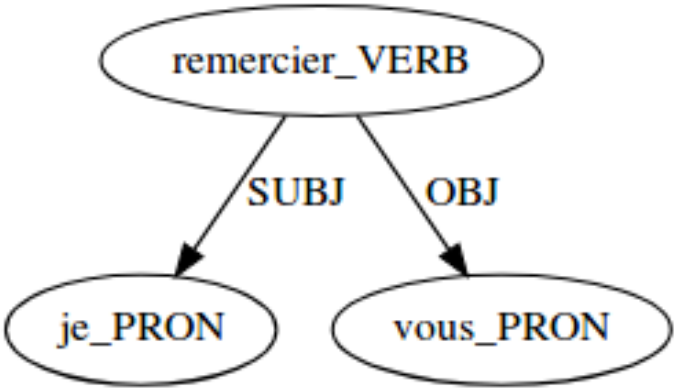
- L'analogie : une piste prometteuse
  - pour la recherche d'exemples (expressions, constructions, traductions, énoncé oraux...)
  - pour la présentation des résultats (catégorisation, colorisation, etc.)
- De nombreuses techniques de TAL s'appuient des mesures de similarités (espaces vectoriels)

je vous remercie

Concordances

Cooccurrences

p.ex. "considération" ou "prendre en considération"



Editer comme requête avancée

Nombre d'occurrences :

*4419 – Bonjour madame, fit M. Jo, **je vous remercie**.*
*4454 – **Je vous remercie**, dit M. Jo, je ne demande pas mieux.\n*
*9777 **Je vous remercie**, elle dit, mais non.*
*14504 – **Je vous remercie** de nous faire confiance.*
*79 – **Je vous remercie** de cette invitation, dit enfin le voyageur en montrant le plus grand trouble.*

Références

Cheng, Winnie, Chris Greaves & Martin Warren, 2006, From n-gram to skipgram to concgram. *International Journal of Corpus Linguistics*, Vol .11, No. 4, pp. 411-433.

Evert, Stefan(2007). Corpora and collocations. in A. Lüdeling and M. Kytö (eds.), *Corpus Linguistics. An International Handbook*, article 58. Mouton de Gruyter, Berlin.

Kilgariff A.,Tugwell D. (2001) WORD SKETCH: Extraction and Display of Significant Collocations for Lexicography, *Proc ACL workshop on COLLOCATION Computational Extraction Analysis and Exploitation*, Toulouse July 2001.

Références

Antoniadis, G., Desmet, P. (2016) NLP for learning and teaching : challenges and opportunities, Traitement Automatique des Langues, Vol.57 N° 3, p. 7-13

Amaral, L. A., Meurer, D. (2011) On using intelligent computer-assisted language learning in real-life foreign language teaching and learning, ReCALL 23(1): 4–24

Bestgen, Yves (2017), Validation interne et externe d'indices phraséologiques pour l'évaluation automatique de textes rédigés en anglais langue étrangère,

Chambers, A. (2010), « L'apprentissage de l'écriture en langue seconde à l'aide d'un corpus spécialisé », Revue française de linguistique appliquée 2010/2 (Vol. XV), p. 9-20.

Boulton, A, Cobb, T. (2017) Corpus Use in Language Learning: A Meta-Analysis, Language Learning, University of Michigan, pp. 1–46

Johns, T. (2002), 'Data-driven learning: the perpetual challenge.' In: B. Kettemann & G. Marko (Eds.), Teaching and Learning by Doing Corpus Analysis. Amsterdam: Rodopi. 107-117.

Hagen, L. K. (1999) Spanish for Business Professionals. Project Web Page. http://www.uhd.edu/academic/research/sbp/

Hamel, M.-J., Slavkov N., Inkpen D., Xiao D. (2016) MyAnnotator : A Tool for Technology- Mediated Written Corrective Feedback, Traitement Automatique des Langues, Vol.57 N° 3, p. 119-142

Heift, T. (2005) Corrective Feedback and Learner Uptake in CALL. ReCALL, 17(1): 32–46.

Meurers, D. (2015) "Learner corpora and natural language processing", in Granger, S., Gilquin, G., & Meunier, F. (eds.), The Cambridge Handbook of Learner Corpus Research, Cambridge: Cambridge University Press, p. 37 - 566

Tack, Anaïs, François T., Ligozat A.-L., Fairon C. (2016) Modèles adaptatifs pour prédire automatiquement la compétence lexicale d'un apprenant de français langue étrangère, JEP-TALN-RECITAL 2016, Vol. 2.

Selva, T., Verlinde, S., Binon, J. (2004) ALFALEX, un environnement d'aide à l'apprentissage lexical du français langue étrangère. Technologies de l'Information et de la Connaissance dans l'Enseignement Supérieur et de l'Industrie, Oct 2004, Compiègne, France. Université de Technologie de Compiègne, pp. 515-522

Wang, I. (2017) Syntactic Similarity Measures in Annotated Corpora for Language Learning : application to Korean Grammar, Thèse de doctorat, Sous la dir. de Sylvain Kahane et Isabelle Tellier, Université Paris 10