Le Web, une mega base de données lexicale multilingue

Traduction automatique : du rêve…

- Débuts de la Traduction Automatique (1950) : espoirs naïfs
- De nos jours : déchiffrer le contenu global

Exemple : news dans une langue inconnue…

- Grec > Français

Traduction automatique : du rêve…

- Vastes progrès : prise en compte d'expressions figées

vol à main armée > armed robbery

vol à la roulotte > stealing from parked vehicles

vol à la tire > pick-pocketing

vol à voile > gliding

vol régulier > scheduled flight

…

A la réalité…

- Un exemple d'erreur (Systran)

caisse claire > clear case

(au lieu de snare drum)

Difficultés : Polysémie

- Ambigüité lexicale

caisse (usage BANQUE) > fund

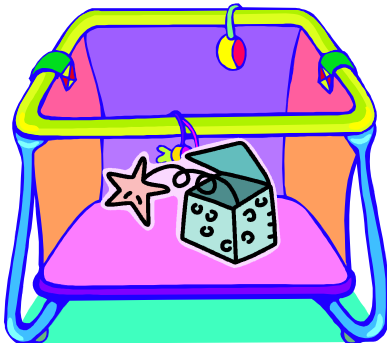
caisse (usage TAMBOUR) > drum

…

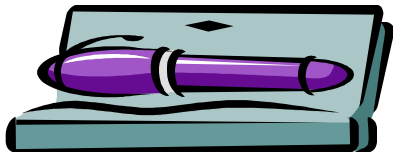
8 traductions en anglais recensées!

Déclin des recherches en TA

- Bar-Hillel, 1960
  - ces exemples sont hors de portée des machines car il demandent une connaissance du monde. Autre exemple :



The box is in the **pen** (la boîte est dans l'**enclos**)
The **pen** is in the box (le **stylo** est dans la boîte)



Difficultés : Idiomatie

- Traductions non littérales

caisse claire > snare drum

(snare : piège / drum : tambour)

Pluie forte > heavy rain

(heavy : lourd)

Solution

⇨ Accès à des bases de données d'équivalences bilingues d'unités lexicales complexes :

Sens d'un mot ambigu en « contexte »

Caisse claire : MUSIQUE

Caisse centrale : BANQUE

Difficultés

- Recensement manuel impossible : des milliers d'Unités Lexicales Complexes (ULC) par langue.

⇨ Méthodes d'acquisition automatique

Méthodes traditionnelles

Corpus parallèles

Ensemble de textes alignés avec leur traduction au niveau du paragraphe, de la phrase, des expressions ou des mots

(Véronis, 2000)

Corpus comparables

Corpus de langues différentes traitant du même domaine mais non parallèles

(Morin et al., 2004)

Corpus parallèles

Un corpus de textes et le corpus de leurs traductions

source		cible
texte s_1	$\longleftrightarrow$	texte c_1
texte s_2	$\longleftrightarrow$	texte c_2
texte s_3	$\longleftrightarrow$	texte c_3
...		...
texte s_n	$\longleftrightarrow$	texte c_n

(Zweigenbaum, 2006)

Corpus comparables

Deux corpus de textes de même domaine, genre, etc.

source
texte s_1
texte s_2
texte s_3
...
texte s_n

cible
texte c_a
texte c_b
...
texte c_m

(Zweigenbaum, 2006)

Difficultés des méthodes traditionnelles

- Corpus parallèles : ressources rares
- Corpus comparables : domaines restreints (médecine, juridique, etc.)
- Utilisation du Web

Objectifs

- Base de données de traductions (Web) :

  caisse claire > *snare drum*

- Français – Anglais

- Mise en place d'une méthodologie : exploiter les facettes multilingues du Web

Intérêts de l'utilisation du Web (1)

- Gigantesque base lexicale :

100 milliards de mots indexés par *Google* pour la seule langue anglaise (Kilgarriff et Grefenstette, 2003)

	BNC	WWW (août 2008)
medical treatment	414	38 000 000
prostate cancer	39	46 100 000
deep breath	732	20 900 000
acrylic paint	30	3 920 000
perfect balance	38	8 590 000
electromagnetic radiation	39	5 980 000
powerful force	71	5 510 000
concrete pipe	10	1 280 000
upholstery fabric	6	3 020 000
vital organ	46	627 000

Intérêts de l'utilisation du Web (2)

- Une base lexicale multilingue

- Exemple : langues sur Wikipédia

http://meta.wikimedia.org/wiki/Liste_des_Wikipédias

Intérêts de l'utilisation du Web (3)

- Une base lexicale évolutive
- Exemple : étude de créations lexicales (Sajous et Tanguy, 2006)

Intérêts de l'utilisation du Web (3)

- Termes techniques :

  Aquamarquage, hémaglutination, immunofixation

- Créations récentes :

  Pacser (se), surencadrement, intermédiation

- Langue populaire :

  Baisage, poilade

- Quelques créations transparentes :

  Pêchable, japonisation, europhobie, googler

Méthodologie

Corpus de pages Web

Appareil
Guitare
Etc.

Unités lexicales complexes

Guitare électrique
Appareil ménager
Appareil circulatoire
Etc.

Dictionnaire bilingue

Polysémie?
appareil
(11 traductions)
vs. guitare
(1 traduction)

Traduction connue?
(circulatoire : pas dans le dictionnaire)

Décision *(3 phases d'acquisition sur le Web)*

Acquisition de ressources lexicales à partir du Web

Collocations

- Associations lexicales fréquentes

pluie diluvienne
café fort

- Relation de dépendance syntaxique

pluie diluvienne (modifieur)

Collocations

- Concept + caractérisation

pluie + diluvienne (INTENSITE)

Locutions (Polguère, 2003)

- Associations lexicales figées

fruit de mer
vs. café très fort

- Référent unique

pomme de terre > potatoe

Unité Lexicale Complexe (ULC)

- Une frontière floue (Tutin et Grossmann, 2002)

café noir?

- Caractérisation
- Non totalement figé
- Variété de café

- Continuum

Acquisition automatique

- Collecte de pages Web

appareil –appareils
appareils –appareil
appareil +appareils

- Repérer les ULC associées

Acquisition automatique

Appareil - Wikipédia
2 nov 2008 ... La notion d'**appareil** (du latin appărāre préparer) réfère à un assemblage cohérent d'organes actifs ou structurants conférant au tout sa ...
fr.wikipedia.org/wiki/**Appareil** - 21k - En cache - Pages similaires

Appareil Photo Numérique [illegible]chat **Appareils** Photo pas cher - Kelkoo
[illegible]chercher des **appareils** [illegible] numériques pas chers ? Avant d'acheter, comparer et trouver au meilleur prix une sélection d'**appareil** photo numérique parmi ...
hifiphotovideo.kelkoo.fr/c-124901-**appareil**-photo-numerique.html - 124k -
En cache - Pages similaires

Guide d'achat **appar**[illegible]**il** photo numérique février 2009
Trouv[illegible] l'**appareil** photo a[illegible]pté à ses besoins et à son budget est aujourd'hui un vrai parcours du com[illegible]ttant ! Pour s'y [illegible]etrouver dans la jungle de ces ...
www.clubic.com/article-74663-1-meilleur-**appareil**-photo-numerique-guide-comparatif.html - 52k -
En cache - Pages similaires

Achat **Appareil** photo numérique au meilleur prix, comparer et ...
Appareil photo numérique au meilleur prix, comparatif et vente de tous les **Appareil** photo numérique. Test de tous les **Appareil** photo numérique Achat ...
www.pixmania.com/**appareil**-photo-numerique/frfr1_1_pm.html - 83k -
En cache - Pages similaires

Revue **Appareil**
Revue **Appareil**. revue soutenue par la MSH Paris Nord. MSH Paris Nord · Revue **Appareil** ... Présentation. A propos d' **Appareil** · Comité de lecture · Rédaction ...
revues.mshparisnord.org/**appareil**/ - 15k - En cache - Pages similaires
de RSS Flux - Les 2 versions

Comparer les prix : **Appareil** photo numérique
Liste de prix pour comparer et acheter des **Appareils** photo numériques - Achat **Appareil** photo numérique en ligne - Classement par marque et nom.
www.i-comparateur.com/comparer-prix-x10c0052b0.htm - 115k - En cache - Pages similaires

Critère 1 - Morpho-syntaxique

- Etiquetage morpho-syntaxique (mot = catégorie grammaticale)

word	pos	lemma
The	DT	the
TreeTagger	NP	TreeTagger
is	VBZ	be
easy	JJ	easy
to	TO	to
use	VB	use
.	SENT	.

Critère 1 - Morpho-syntaxique

- Description de patrons morpho-syntaxiques

NOM-ADJECTIF
appareil digital

NOM-PREP-NOM
appareil de musculation
parc d'attraction

Critère II - Fréquences

- Fréquences au sein des pages Web collectées
- Fréquences sur tout le Web

  « appareil ménager » (91 500)

- Environ 10 000 ULC récoltées

Un exemple : « appareil »

appareil NOM2	d’	état, imagerie
NOM1 appareil	d’	catégorie
appareil NOM2	de	chauffage, contrôle, cuisson, mesure, poche, production, protection
appareil ADJECTIF	NOM-ADJECTIF	administratif, argentique, auditif, circulatoire, compact, critique, dentaire, digestif, électrique, électroménager, électronique, étatique, génital, gouvernemental, judiciaire, locomoteur, ménager, militaire, mobile, numérique, photo, photographique, politique, portable, productif, réflex, répressif, reproducteur, respiratoire, urinaire

Acquisition automatique de traductions :

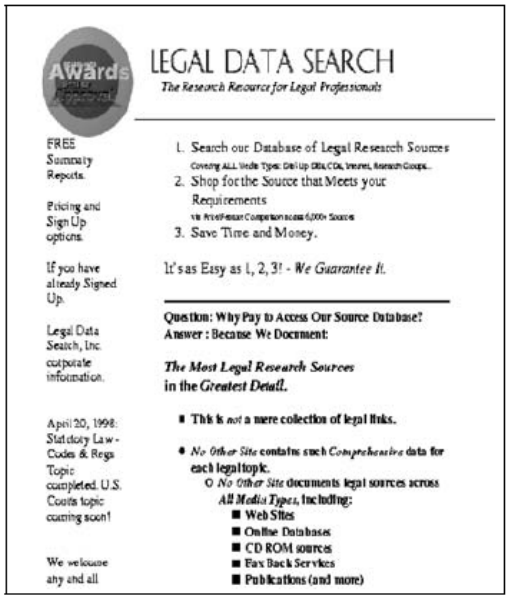
méthodes existantes

Méthodes existantes

- Utilisation des fréquences
- « Web parallèle »
- « Web comparable »

Le Web, un « corpus parallèle »

- Repérage de pages parallèles *(Ma & Liberman, 1999); (Nie, 1999); (Nie et al, 2000); (Resnik, 1999); (Resnik & Smith, 2002), etc.*



LEGAL DATA SEARCH

The Research Resource for Legal Professionals

FREE Summary Reports.

Pricing and Sign Up options.

If you have already Signed Up.

Legal Data Search, Inc. corporate information.

April 20, 1998: Statutory Law - Codes & Regs Topic completed. U.S. Courts topic coming soon!

We welcome any and all

1. Search our Database of Legal Research Sources
   Covering ALL Media Types: Dial Up DBs, CDs, Internet, Research Groups...
2. Shop for the Source that Meets your Requirements
   vs. Price/Feature Comparison across 6,000+ Sources
3. Save Time and Money.

It's as Easy as 1, 2, 3! - *We Guarantee It.*

Question: Why Pay to Access Our Source Database?
Answer : Because We Document:

The Most Legal Research Sources in the Greatest Detail.

- **This is *not* a mere collection of legal links.**
- ***No Other Site* contains such *Comprehensive* data for each legal topic.**
  - ***No Other Site* documents legal sources across *All Media Types*, including:**
    - **Web Sites**
    - **Online Databases**
    - **CD ROM sources**
    - **Fax Back Services**
    - **Publications (and more)**

LEGAL DATA SEARCH

El recurso de la investigación para los profesionales legales

SIDEBAR

Inglés Francés Español Alemán Italiano Portugués

Ejecute un informe LIBRE de Topic de la muestra " búsqueda " contra una vista limitada de la base de datos.

Tasación y muestra encima de opciones.

Si usted ha firmado ya para arriba.

Información corporativa Legal Data Search, Inc..

De marcha la 12 de 1998: Característica de búsqueda de encargo terminada. Ley

1. Busque nuestra base de datos de las fuentes legales de la investigación
   Cubre Todos los Tipos De Medio: Dial-Up DBs, Cds, Internet, Grupos De Investigación...
2. Departamento para la fuente que resuelve sus requisitos
   vs la comparación de Característica a través de las fuentes 6000+
3. Excepto tiempo y dinero.

Es tan fácil como 1, 2, 3! - *Lo garantizamos.*

Pregunta: Por qué paga tener acceso a nuestra base de datos de la fuente?
Respuesta: Porque Documentamos:

Las Fuentes Más legales De la Investigación en el detalle más grande.

- **Ésta no es una colección mera de conexiones legales.**
- ***Ninguno otro sitio* contiene tales *datos* comprensivos para cada asunto legal.**
  - ***Ninguno otro sitio* documenta fuentes legales a través *Todos los tipos de medios*, incluyendo:**
    - **Sitios Del Web**
    - **Bases de datos En línea**
    - **Fuentes "COPIA MÁS OSCURA" de la ROM**
    - **Servicios Posteriores Del Fax**
    - **Publicaciones (y más)**

Typologie

- Page parente : page qui contient au moins 2 liens vers des pages traduites (ou sites)

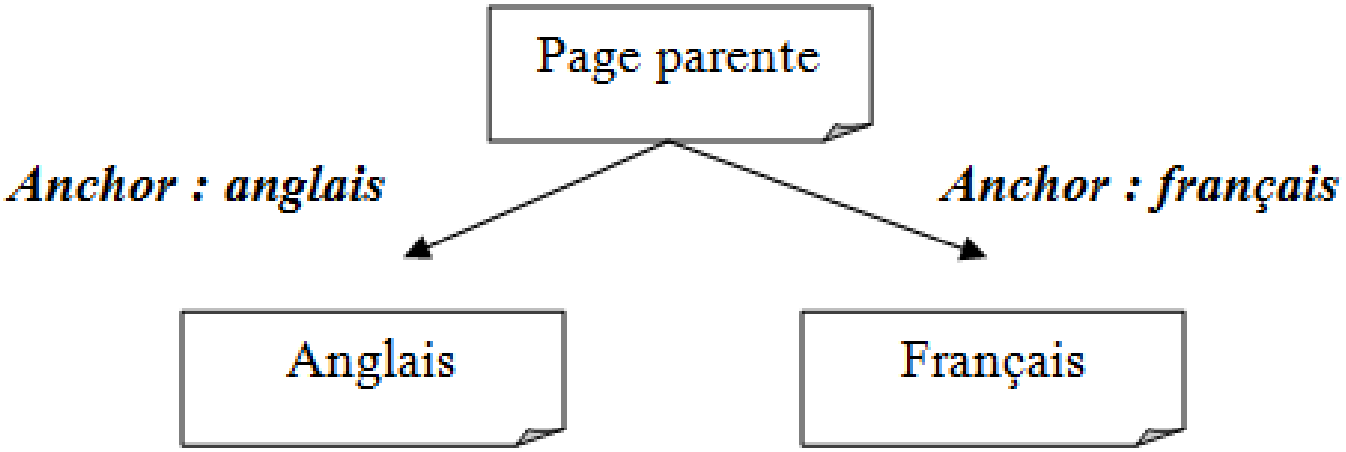
Typologie

- Page sœur : document monolingue qui contient un lien vers sa traduction

Stratégies (1)

- Repérage de liens hypertextes :

anchor : « language1 » AND anchor : « language2 »
(anchor : « english » OR anchor: "anglais") AND (anchor : « french » OR anchor : "français")



Stratégies (2)

- Repérage d'URL similaires (exemple : pt – portugais / en – anglais) :

http://www.ex.pt/index_pt.html
http://www.ex.pt/index_en.html

- Comparaison du contenu sémantique des sites candidats

Stratégies (3)

- Comparaison de la structure HTML :

<HTML> <TITLE>Emergency Exit</TITLE> <BODY> <H1>Emergency Exit</H1> If seated at an exit and ⋮	<HTML> <TITLE>Sortie de Secours</TITLE> <BODY> Si vous êtes assis à côté d'une . . . ⋮

Le Web, un « corpus partiellement parallèle »

- Documents « mixtes » (Nagata, 2001)

“緑内障が早期発見によって管理可能な病気となったため、黄斑変性(macular degeneration)が先進国の視力障害の主な原因となりつつある。”

Typologie des documents mixtes

- Paragraphes alignés : alternance de paragraphes traduits

Registration for Foreign Residents and Birth Registration
がいこくじんとうろく しゅっせいとど
外国人登録と 出生届け
The official name for registration for foreign residents in Japan, as determined by the Ministry of Justice, is “Alien Registration”.
...
Anyone staying in Japan for more than 90 days, children born in Japan, ...
90日以上日本に滞在するとき、子供が日本で生まれたとき、...
...

Typologie des documents mixtes

- Tables : paires d'équivalences de termes (glossaires bilingues, etc.)

instrument	bass drum	grosse caisse
instrument	bassoon	basson
instrument	bugle	clairon
instrument	cello	violoncelle
instrument	double bass	contrebasse
instrument	electric guitar	guitare électrique
instrument	English cor	cor anglais
instrument	French horn	cor français
instrument	Gong	gong
instrument	guitar	guitare
instrument	harmonica	harmonica
instrument	harp	harpe
instrument	Jew harp	guimbarde
instrument	kettledrum	timbale
instrument	mandolin	mandoline
instrument	Oboe	hautbois
instrument	organ	orgue
instrument	panpipe	flûte de Pan
instrument	Piano	piano
instrument	piccolo	piccolo
instrument	snare drum	caisse claire

Typologie des documents mixtes

- Texte plein : traductions ponctuelles dans un document monolingue

*Further support was guaranteed […], the Saudi Fund, France's Central Fund for Economic Cooperation (**Caisse Centrale de Coopération Economique**–CCCE).*

Informations quantitatives sur le Web

- Fréquences pour l'aide au choix lexical
  (Grefenstette, 1999), (Cao & Li, 2002), (Léon & Millon, 2005)

- Hypothèse : la fréquence reflète l'usage (dans une certaine mesure)

Un exemple…

Appareil numérique

> Digital **aeroplane**? Digital **camera**? …

Difficulté

- Pas de test d'équivalence langue source/langue cible. Exemple : *cours de formation*

cours > rate (FINANCE)…
formation > group (COLLECTIVITE)…

Traduction existante, mais non équivalente (« tarif de groupe »)

Hypothèse

Les contextes lexicaux en langue source et en langue cible sont proches entre une unité lexicale et sa traduction correcte, mais différents lorsque la traduction est erronée.

Cours de formation

Cours de formation
cours de formation offert par le Conseil canadien de la sécurité ... **Cours de formation** de garde-enfants. La conduite préventive. La sécurité routière ...
www.**safety-council.org**/CCS/formation/coursd.htm - En cache

Truc a la con talc : formation pour homme
COURS DE FORMATION OFFERT AUX HOMMES (merci à Mu pour cette info ; ... OBJECTIF PEDAGOGIQUE : **Cours de formation** permettant aux hommes d'éveiller cet ...
humour-blague.com/page/formation.php - En cache

Formation musicale - cours de solfege : Allegromusique
Allegro musique, c'est le plaisir de la musique - Allegro Musique propose des cours de musique à domicile pour tous ... des **cours de Formation** Musicale ...
www.**allegromusique.fr**/solfege.htm - En cache

Training course

Training course
Poclain ... **Training course**. Commercial-Marketing. Quality. Human Resources ... tell us... Engineering. **Training course**. Commercial-Marketing. Quality ...
www.**poclain-hydraulics.com**/Default.aspx?tabid=214 - 115k - En cache

Training course of golf with Hammamet Tunisia
training course,**Training course** of golf with Hammamet,Tunisia,lessons, ... raftered player, we have a formula of **training course** which corresponds to you...
traininggolf.canalblog.com - En cache

Motorcycle **Training Course** - GTA to Ottawa Ontario - Traduire
The Canada Safety Council's Gearing Up Program, developed in co-operation with the Federal Government, is endorsed by all levels of government and the Insurance ...
www.**motorcyclecourse.com** - En cache

Group rate

Services : Hotel Ruby Foo's, Montréal - TripAdvisor
Hotel Ruby Foo's, Montréal, Québec: Our group stayed at Ruby Foos, **Group rate** was done by... - Retrouvez sur TripAdvisor 25 critiques impartiales sur Hotel Ruby Foo's.
tripadvisor.fr/ShowUserReviews-g155032-d183226-r8496483-Hotel_Ruby_...

Activités Durango - Sorties Durango - TripAdvisor
Sorties Durango : Consultez TripAdvisor, le meilleur site de critiques impartiales, ... HotelPlanner.com Group Hotel Rates Did you lock in a **group rate** yet? ...
www.**tripadvisor.fr**/Attractions-g33397-Activities-Durango_Colorado.html - 66k - En cache

Les Groupes
Plus qu'un hotel, la Grande Cascade est le maison de la famillle Pierre. ... Dancing on request. Board - Half board. 30 rooms. Half board **Group rate**: €39 ...
www.**grandecascade.fr**/pagesang/groupes.htm - En cache

Le Web, un corpus « comparable »

- Comparaison des contextes cibles et sources, à partir de requêtes sur le Web

- *Exemple :*

Appareil numérique
canon, photographie, nikon, informatique, produits, accessoires, digital, mémoire, kodak, pc, etc.

Digital camera
photography, film, computer, kodak, technology, olympus, canon, right, zoom, sony, etc.

Acquisition automatique de traductions :

une méthodologie mixte et modulaire

Méthodologie

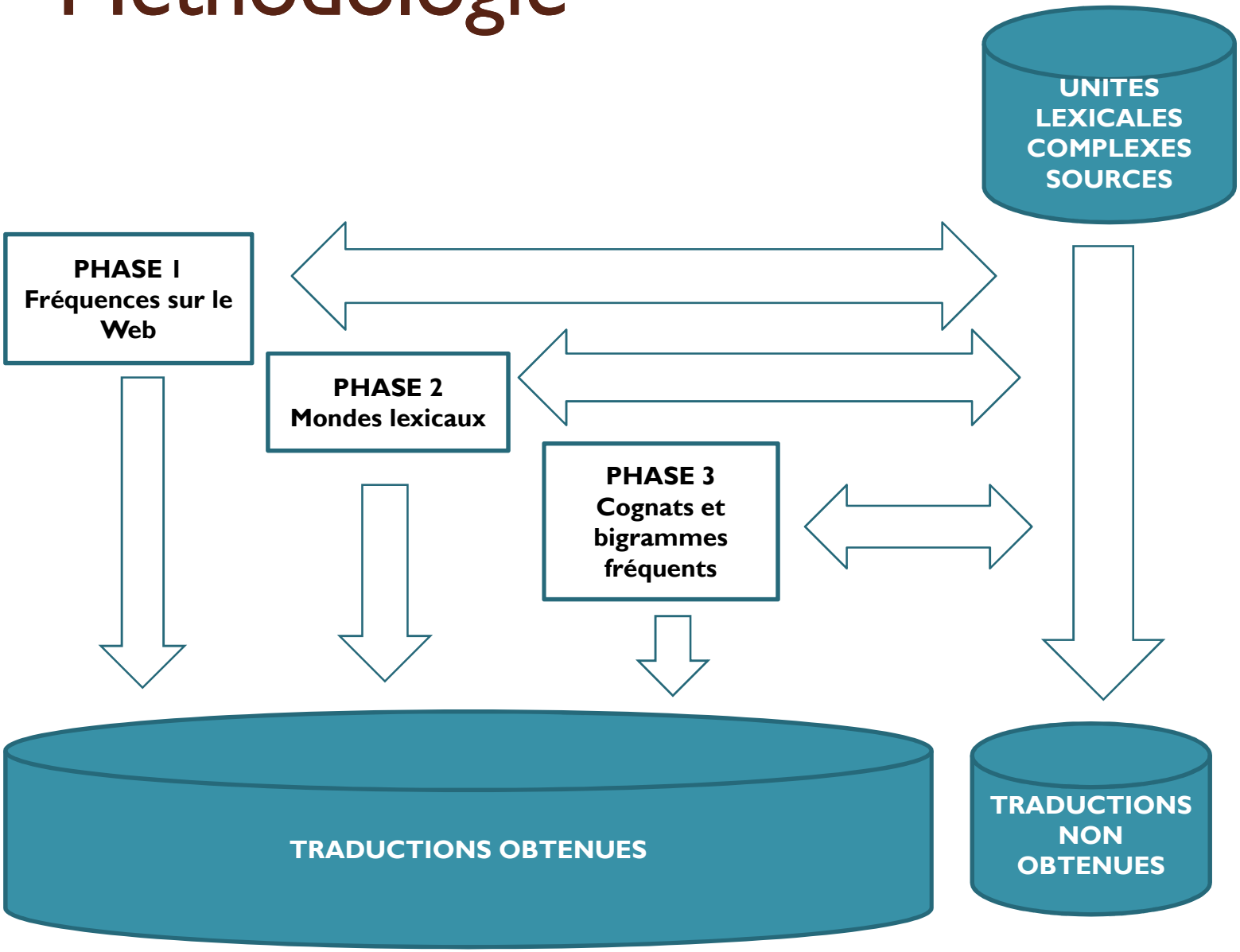
- Traitement modulaire
- « Mixte » de stratégies
- Système linéaire

Méthodologie

⇨ Echantillon aléatoire de 1075 ULC sources

(parmi les 10 000 collectées)

Méthodologie



Phase I

Traductions compositionnelles non polysémiques

Exemple prototypique

guitare électrique

- Cumul des deux traductions
- Nombre de traductions

guitare (1 traduction) / électrique (1 traduction)
vs. appareil (11 traductions)

Méthodologie

- Combinaison des traductions de chaque élément

guitare > guitar
électrique > electric

- Règles de transformation

NOM-ADJECTIF > ADJECTIF-NOM
guitare électrique > electric guitar

Méthodologie

- Règles de transformation

NOM1-de(d')-NOM2 > NOM2-NOM1

acte de résistance > resistance act

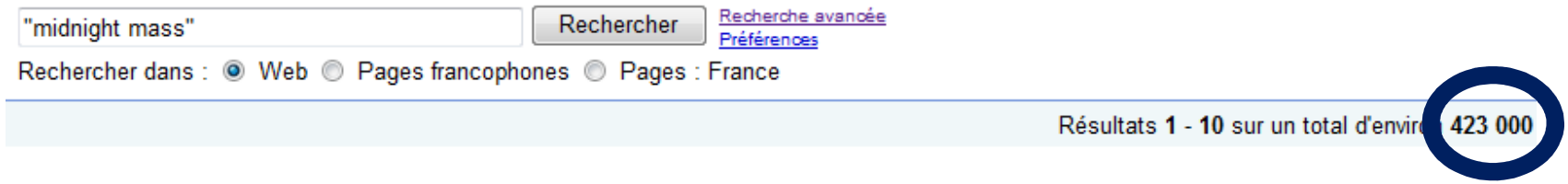
NOM1-de(d')-NOM2 > NOM1-of-NOM2

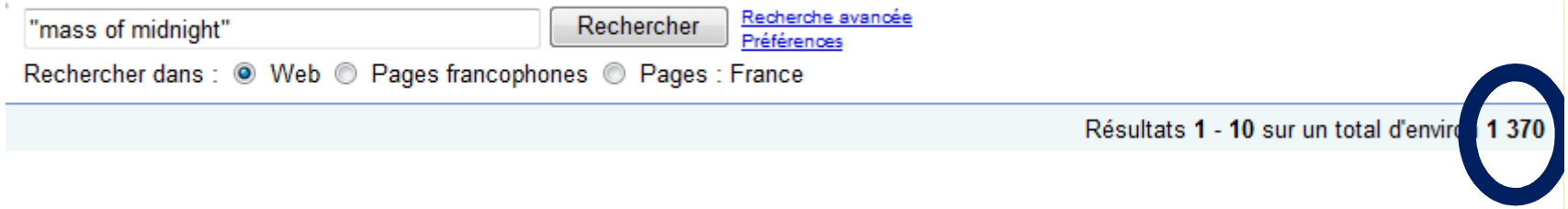
acte de résistance > act of resistance

Informations quantitatives

- Aide à la validation (Grefenstette, 1999; Léon & Millon, 2005; etc.)

messe de minuit





Test de validité

- Si valide : traduction stockée

drame musical > musical drama

averse de neige > snow shower

- Si invalide : module suivant

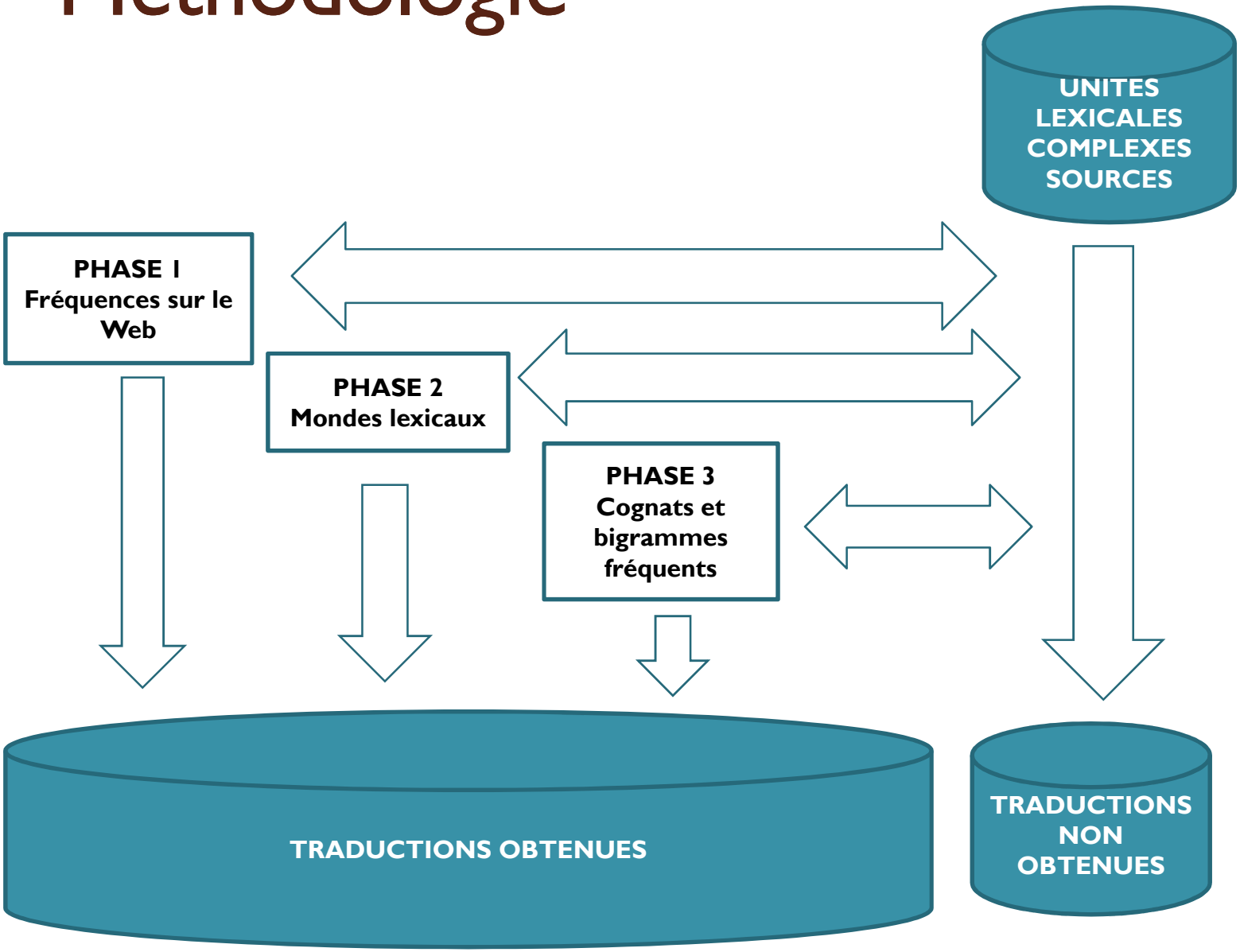
*bonheur perdu > stray hapiness**

Limite

- Pas de test d'équivalence lorsque les mots sont polysémiques

- Si polysémie : contextes lexicaux sur le Web

Méthodologie



Phase 2

Traductions compositionnelles polysémiques

Exemple prototypique

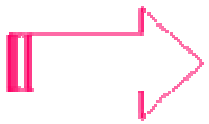
appareil ménager

- Cumul des deux traductions
- Au moins 2 traductions possibles pour au moins un des constituants (appareil : 11 traductions)

Méthodologie

- Filtres des traductions candidates

appareil	ménager
appliance	household
aeroplane	domestic
device	
camera	
…	



Traductions candidates
household appliance
household aeroplane
…
domestic appliance
domestic aeroplane
…

I- Filtre du « Web parallèle »

- Pages linguistiquement mixtes (Nagata, 2001)

- La caisse claire (snare drum) est constituée d'un fût (shell) , de deux peaux (heads), d'un timbre (snare wires), de deux cercles (hoops), d'un déclencheur (throw-off), de coquilles (lugs) et de vis (screws). Les caisse claires que l'on rencontre le plus souvent sont fabriquées en érable ou en aluminium, mais il existe une grande variété de matériaux utilisés pour leur fabrication: laiton, bronze, fonte, fer, acier, cuivre, érable, bouleau, bambou, cerisier, tilleul, des bois exotiques comme le bubinga, le jarrah, le sheoak, des matières synthétiques comme l'acrylique, la fibre de verre ou le carbone, des matériaux composite et même du verre ! Les dernières nouveautés en terme de construction de caisse claire sont les fûts hybride en bois et métal (DW, edge), ou bois et acrylique (Spaun, hybrid).

I- Filtre du « Web parallèle »

- Requêtes des couples de traduction :

  «ULC» «TRADUCTION CANDIDATE»
  « caisse centrale » « central fund » (328)
  « caisse centrale » « central case » (3)…

- Hypothèse : co-occurrence du couple lorsque la traduction est correcte

II- Construction de mondes lexicaux

- Acquisition automatique de résumés (API Yahoo)

«appareil numérique» -«appareils numériques»
«appareils numériques» -«appareil numérique»
«appareil numérique» +«appareils numériques»

- « Mondes lexicaux » (Véronis, 2003, 2004)

Exemple (français)

APPAREIL MILITAIRE	
NOMS	pays (123), guerre (121), avion (120), sécurité (108), membre (103), source (84), existence (77), société (61), réaction (61), vol (60), monde (46), technologie (45), esprit (44), conflit (44), libération (43), transport (42), aviation (42), supériorité (40), droit (39), intégration (38)
ADJECTIFS	civil (112), puissant (110), français (90), étranger (89), américain (78), politique (69), utilisateur (57), aérien (56), médiatique (54), national (40), majeur (40), réel (39), mauvais (39), économique (33), ancien (32), mondial (27), social (26), iranien (26), francophone (26), armé (26)

Exemple (anglais)

MILITARY PLANE	
NOMS	crash (166), aircraft (165), air (141), world (77), fighter (69), time (64), transport (60), area (60), airport (60), security (52), missile (51), fire (49), airplane (49), aviation (46), war (45), pilot (43), jet (43), governement (41), airspace (40), cargo (38)
ADJECTIFS	russian (111), civilian (77), american (73), iranian (66), commercial (48), chinese (48), german (47), strategic (40), iraqi (38), international (37), french (35), least (34), foreign (34), vintage (32), venezuelan (32), free (32), added (31), political (30), turkish (29), regular (29)

III- Comparaison des mondes lexicaux

- Intersection des mots communs via un dictionnaire bilingue

APPAREIL COMPACT / COMPACT CAMERA	
NOMS	boîtier/case, dimension/size, équipement/equipment, flash/flash, gamme/range, marché/market, mémoire/memory, mesure/time, monde/world, objectif/lens, photographie/photography, produit/product, qualité/quality, sac/bag, série/series, taille/size, technologie/technology, zoom/zoom
ADJECTIFS	automatique/automatic, digital/digital, étanche/waterproof, faible/low, idéal/ideal, léger/light, manuel/manual, optique/optical, parfait/perfect, portable/portable, professionnel/professional, puissant/powerful, rapide/fast, rare/rare

Test de validité

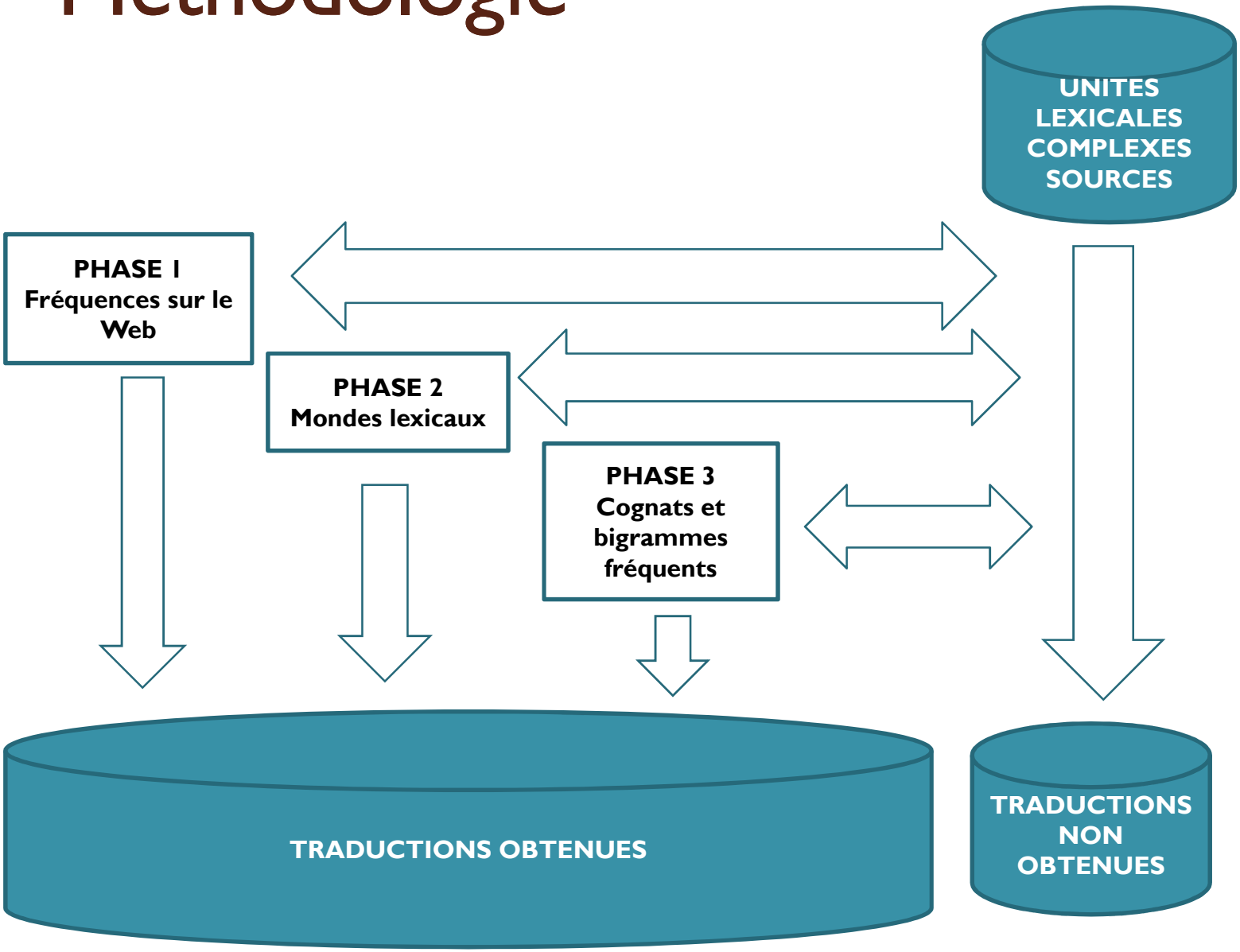
- Si valide : traduction stockée

course de karting > karting race
éclat naturel > natural shine

- Si invalide : module suivant

*caisse claire > * clear case*

Méthodologie



Phase 3

Traductions non compositionnelles et/ou inconnues

Exemples prototypiques

caisse claire
appareil circulatoire

- Pas de cumul des traductions (snare drum)
- Traduction inconnue d'un des éléments (*circulatoire* non présent dans le dictionnaire)

Méthodologie

- Repérer les traductions directement sur le Web

- Collecte de résumés « mixtes »

Jacques Delécluse: Douze Etudes Pour **Caisse Claire** at Musicroom ... - [Traduire cette page]
Sheet Music - £16.95 - For Snare drum (French, English, German and Spanish text).
www.musicroom.com/se/ID_No/0252876/details.html - 42k - En cache - Pages similaires

Translation **caisse claire** in the French-English Collins dictionary - [Traduire cette page]
caisse claire translation French - English : **caisse claire** nf (MUSIQUE) snare drum ... .
French - English, Collins dictionary, synonyms, translation.
dictionary.reverso.net/french-english/caisse%20claire - Pages similaires

Méthodologie

- Requête en français limitée à la langue anglaise

- Exemple

« souris d'agneau »

Langue sélectionnée : anglais

Méthodologie

Souris d'agneau confite, légumes oubliés en mash on Flickr - Photo Sharing! - Traduire
Braised lamb shank, root vegetable mash. Urbane, 12 rue Arthur Groussier in the 10th (01 42 40 74 75) ... Explore Page Video on Flickr Last 7 Days Interesting ...
www.**flickr.com**/photos/clotilde/446968461 - En cache

Souris d'agneau, condiment figue-gambas on Flickr - Photo Sharing! - Traduire
Lamb shank, fig-jumbo shrimp condiment. Itinéraires, 5 rue de Pontoise, Paris 5ème (01 46 33 60 11) ... Explore Page Video on Flickr Last 7 Days Interesting ...
www.**flickr.com**/photos/clotilde/2589024731 - En cache

http://www.bloodyfrench.com - Traduire
Brick de chèvre chaud au miel /£7.9. Hot goat cheese pastry topped with honey ... **Souris d'agneau** aux légumes et aux abricots / £13.5 ...
www.**bloodyfrench.com** - En cache

Restaurant Vin des Pyrénées - Traduire
A La Carte Menu: €30 - €40 per person. Photos n° ... de Norvège rôti à l'anis étoilé, and **Souris d'agneau** braisée aux choux verts. ...
parismarais.com/.../vins-des-pyrenees/vins-des-pyrenees.htm - En cache

Méthodologie

- Méthodes d'identification des traductions

- Collecte de cognats, puis de bigrammes fréquents

I- Repérage de cognats

- Occurrences identiques ou qui se ressemblent graphiquement

  - Formes identiques

    code > *code*

  - Bases communes

    circulatoire > *circulatory*

II- Repérage des bigrammes fréquents

- Bigrammes les plus fréquents au sein des résumés mixtes

souris d'agneau "lamb shank"
souris d'agneau "geneve pays"
souris d'agneau "detail produit"
souris d'agneau "lamb shanks"
souris d'agneau "weekly letter"
souris d'agneau "anglais discussion"
souris d'agneau "zucchini recipe"
souris d'agneau "weather forecast"
souris d'agneau "username password"

Exemples de traductions

acide nucléique > nucleic acid

casque de protection > protective helmet

étui de protection > protective cover

applique murale > wall lamp

aurore boréale > nothern light

IV- Résultats et évaluation

Evaluation humaine

- Traductrice professionnelle

	Français	English	Recherche	Evaluation
1	absence temporaire	temporary absence	Recherche	A
2	accès libre	free access	Recherche	A
3	accident grave	serious accident	Recherche	A
4	accident vasculaire	vascular disease	Recherche	C
5	accord de contribution	contribution agreement	Recherche	A
6	accord global	overall understanding	Recherche	C
7	accord mutuel	mutual understanding	Recherche	C
8	achat immédiat	immediate purchase	Recherche	A
9	acide aminé	amino acid	Recherche	A
10	acide nucléique	nucleic acid	Recherche	A
11	acier inox	stainless steel	Recherche	A
12	acte législatif	legislative proceedings	Recherche	B
13	acte de résistance	act of resistance	Recherche	A
14	acteur économique	economic actor	Recherche	A
15	acte de vente	bill of sale	Recherche	A
16	action commune	joint action	Recherche	A
17	action directe	direct action	Recherche	A
18	action internationale	international action	Recherche	A
19	action nouvelle	new share	Recherche	C
20	action stratégique	strategic action	Recherche	A

Résultats

Catégories d'évaluation	Nombre de traductions	Pourcentage
A	792	89,29%
B	45	5,07%
C	50	5,64%

Catégories d'évaluation	Nombre de traductions	Pourcentage
Acceptable	837	94,36%
Non acceptable	50	5,64%

Erreurs lexicales

- Choix thématiquement proche

  expression orale > oral communication

- Choix lexical erroné

  voie commerciale > commercial road

Erreurs morpho-syntaxiques

- Choix erroné de structure syntaxique

analyse de marché > analysis of market

- Structure syntaxique non traitée

lait de femme > woman milk

Possessif : woman's milk

Erreurs idiomatiques

- Sens lexical adéquat mais association non idiomatique

  fête d'anniversaire > *anniversary party*

V- Discussion et évolutions

Discussion : Apports de la méthode

- Modules de traduction adaptés aux caractéristiques linguistiques
- Réflexion théorique et pratique sur l'utilisation du Web :
  - Acquisition d'ULC sources
  - Acquisition et filtres des traductions
- Comparaison de mondes lexicaux à partir du Web

Discussion : Apports de la méthode

- 82,51% de traductions obtenues
- 94,36% de traductions acceptables

Evolutions : Bruit

- Améliorer la comparaison des mondes lexicaux

  campagne publique > state country

- Pondération des fréquences (accorder plus de « poids » aux mots les plus fréquents)

Evolutions : Bruit

- Elargissement des patrons morpho-syntaxiques cibles

villa þprovençale > þprovencal style villa

Evolutions : Silence

- Traductions absentes des résumés mixtes

futur antérieur (future perfect)

- Collecte de résumés « comparables »

verbe, anthologie… > *verb +anthology*

Evolutions : Silence

- Mondes lexicaux non homogènes

	MOIS D'ABSENCE
NOMS	membre (28), série (25), match (21), sport (19), musique (18), football (18), championnat (17), santé (16), monde (16), saison (15), foot (14), équipe (14), film (13), voyage (12), succès (12), journée (12), discussion (12), connexion (12), cinéma (12), accueil (12)
ADJECTIFS	français (19), beau (17), francophone (12), ivoirien (10), professionnel (9), jeune (9), bienvenu (9), politique (9), ancien (8), rapide (7), national (7), live (7), informatique (7), social (6), présent (6), longue (6), virtuel (5), sportif (5), public (5), prochain (5)

Evolutions : Méthodologie

- Nouvelle utilisation du Web : cibler des genres ou des thématiques
  - Genres : news, Wikipédia, etc.
  - Thématiques : forums thématiques, etc.
- Améliorer les cas d'ambiguïté lexicale restants

Evolutions : Exemple d'application

- Ressources de type ontologique (hiérarchisée)

- Classification sémantiques des ULC

PHOTOGRAPHIE
appareil numérique
appareil compact
…

MENAGER
appareil électrique
appareil électroménager
…

Repères bibliographiques

Grefenstette, G. (1999). *The World Wide Web as a Resource for Example-Based Machine Translation Tasks*. ASLIB "Translating and the Computer" conference, Londres, Angleterre.

Léon, S., Millon, C. (2005). *Acquisition semi-automatique de relations lexicales bilingues (français-anglais) à partir du Web*. Rencontre des Etudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL), Dourdan, France.

Nie, J.-Y., Simard, M., Isabelle, P., Durand, R. (1999). *Cross-Language Information Retrieval Based on Parallel Texts and Automatic Mining of Parallel Texts from the Web*. ACM SIGIR'99.

Ma, X., Liberman, M. (1999). *Bits: A method for bilingual text search over the web*. Machine Translation Summit VII, Singapour, Singapour.

Morin E., Dufour-Kowalski S., Daille B. (2004), Extraction de terminologies bilingues à partir de corpus , *Actes de TALN'2004*, Fès (Maroc).

Nagata, M. (2001). *Using the Web as a bilingual dictionary*. 39th ACL Workshop on Data-Driven Methods in Machine Translation.

Polguère, A. (2003). *Lexicologie et sémantique lexicale. Notions fondamentales*, Les Presses de l'Université de Montréal.

Resnik, P. (1999). *Mining the web for bilingual text*. 37th Annual Meeting of the Association for Computational Linguistics.

Resnik, P., Smith, N. A. (2003). "The Web as a parallel corpus." *Computational Linguistics, Special issue on web as corpus* 29(3): 349 - 380.

Tutin, A., Grossmann, Francis (2002). "Collocations régulières et irrégulières : esquisse de typologie du phénomène collocatif." *Revue française de linguistique appliquée, Lexique : recherches actuelles* VII: 7-25.

Véronis J. (Ed.) (2000). Parallel Text Processing: Alignment and use of translation corpora, Kluwer Academic Publishers.

Véronis, J. (2003). *Cartographie lexicale pour la recherche d'information*. Actes de la Conférence Traitement Automatique des Langues (TALN'2003), Batz-sur-Mer, France, ATALA.

Véronis, J. (2004). "HyperLex: lexical cartography for information retrieval." *Computer Speech & Language* 18(3): 223–252.