Stratégie de développement des technologies langagières pour des langues avec peu de ressources : le cas du basque

Xabier Artola

Groupe de recherche Ixa (*Ixa taldea*)
Faculté de Informatique
Université du Pays Basque

http://ixa.si.ehu.es

Billère/Vilhèra, 2014-06-24

Contenu de la présentation

- Les langues dans le contexte des technologies de l'information et des communications (TIC), et de la technologie du langage (TL).
- Le groupe de recherche Ixa.
  - Stratégie de développement : le traitement de la langue basque dans le groupe Ixa.
- Conclusions

Les langues dans le contexte des TIC et de la TL

- Réseau d'Excellence META-NET.
  - Les langues dans le Web, les technologies de la langue, des opportunités que ces technologies peuvent nous offrir.
  - Résultats pour le basque (comparés au français et à l'anglais)
- Comment les langues font-elles face aux défis des TIC et de la TL ?
  - Certains paramètres pour classer les langues.
  - Quelles sont les langues avec « moins de ressources » ? Six niveaux différents.

META-NET

- META-NET, l'Alliance Technologique pour une Europe multilingue : réseau d'excellence soutenu par la Commission Européenne.
- 50+ laboratoires de recherche du domaine des sciences et technologies de la langue, dans une trentaine de pays.
- Collection de livres blancs sur les technologies de la langue : analyse de l'état des ressources et des technologies de la langue pour 31 langues européennes.

Technologies de la langue : traitement automatique d'une langue

- Correcteurs de texte, moteurs de recherche sur la toile, systèmes de réponse aux questions, reconnaissance et synthèse automatique de la parole, dialogue oral, traduction automatique (écrite ou vocale)…
- Mais aussi : la reconnaissance du locuteur ou de la langue parlée, l'extraction d'information ou le résumé automatique des textes…

Les langues sur le Web

- Il y a quelques années : la vaste majorité des contenus sur le Web étaient en anglais.
- Aujourd'hui : la quantité de contenus en ligne dans d'autres langues (européennes, mais aussi asiatiques et l'arabe en particulier) a explosé.
- *Fossé numérique* causé par les frontières linguistiques : Quelles langues vont prospérer et persister dans la société de l'information et du savoir en réseau, et quelles sont celles qui sont susceptibles de disparaître ?

Nos langues en danger

- Arrivée de l'imprimerie : inestimable échange d'information en Europe, mais aussi l'extinction de certaines langues européennes.
- Langues minoritaires : rarement imprimées, limitées par leur forme orale, ce qui a restreint leur adoption, diffusion et utilisation par rapport aux langues imprimées.

- L'Internet aura-t-il le même impact sur nos langues actuelles ?

Nos langues en danger

Alors que les langues largement répandues comme l'anglais ou l'espagnol vont certainement maintenir leur présence dans la société numérique émergente et sur le marché international, beaucoup de langues européennes pourraient être coupées de la communication numérique et devenir sans importance dans une société en réseau.

Technologies de la langue : des technologies-clés habilitantes

- Les technologies de la langue peuvent aider les individus :
  - à collaborer ;
  - à entretenir des échanges commerciaux ;
  - à partager des connaissances ;
  - à participer à des débats sociaux ou politiques (indépendamment des barrières linguistiques ou des compétences informatiques) ;
  - …

Technologies de la langue : des technologies-clés habilitantes

- Elles opèrent souvent de façon cachée dans des logiciels complexes lorsque nous :
  - trouvons des informations avec un moteur de recherche sur Internet ;
  - vérifions l'orthographe et la grammaire dans un traitement de texte ;
  - obtenons des recommandations de produits dans un magasin en ligne ;
  - entendons les instructions verbales d'un système de navigation routière ;
  - traduisons des pages web, des courriels, des blogs, etc. avec un service de traduction en ligne ;
  - …

Technologies de la langue : des technologies-clés habilitantes

- Sans les technologies de la langue, nous ne serons pas en mesure de donner aux utilisateurs des moyens de communiquer réellement interactifs, multimédias et multilingues dans un futur proche.

Des opportunités pour les technologies de la langue

- Les technologies de la langue peuvent :
  - simplifier et automatiser les processus de traduction, de production de contenus, de traitement de l'information et de gestion des connaissances ;
  - favoriser le développement d'interfaces vocales pour les appareils électroniques domestiques, les machines, les véhicules, les ordinateurs, les téléphones ou les robots ;
  - …

Des opportunités pour les technologies de la langue

- Les applications commerciales et industrielles sont encore dans les premiers stades de développement, même si les récentes réalisations en R&D ont créé un éventail d'opportunités.
- Par exemple, la traduction automatique atteint déjà une qualité raisonnable dans des domaines spécifiques :
  - des applications expérimentales permettent d'effectuer la fourniture d'information multilingue, la gestion des connaissances ainsi que la production de contenus dans de nombreuses langues.

Des opportunités pour les technologies de la langue

- *Les technologies de la langue représentent une formidable opportunité pour l'Union européenne. Elles peuvent aider à traiter la délicate question du multilinguisme en Europe – le fait que plusieurs langues coexistent naturellement dans les entreprises européennes, les organisations et les écoles.*

- *Les technologies de la langue peuvent être vues comme des sortes de technologies « d'assistance » qui aident à résoudre le « handicap » de la diversité des langues et rendent les communautés linguistiques plus accessibles les unes aux autres.*

Acquisition de la langue par les humains et les machines

- Les systèmes TL « acquièrent » leurs capacités linguistiques d'une manière semblable à celles des humains.
  - Approches statistiques (fondées sur les données) : ils acquièrent les connaissances linguistiques à partir de vastes collections d'exemples concrets de texte dans une seule langue ou à partir de ce qu'on appelle des *textes parallèles.*

    ⇒ *il faut avoir des grandes collections de textes ; c'est difficile a réussir pour des langues minoritaires!*
  - Systèmes de règles : des experts de la linguistique, de la linguistique computationnelle et de l'informatique doivent d'abord coder la grammaire de la langue (ou les règles de traduction) et compiler des lexiques (vocabulaire).

    ⇒ *c'est très cher : les TL à base de règles n'ont été développées que pour les principales langues!*
- Les avantages et inconvénients des systèmes statistiques ou à base de règles tendent à être complémentaires : les recherches actuelles s'orientent vers des approches hybrides.

META-NET: Conclusions et résultats pour le basque

- Dans le domaine des technologies de la langue, la langue basque montre un certain nombre de produits, de technologies et de ressources.
- Le basque est l'une des langues de l'UE qui ont besoin encore des recherches plus poussées pour que les solutions technologiques soient prêtes pour une utilisation quotidienne.
- Le développement de technologie de haute qualité pour le basque est urgent et d'une importance capitale pour la préservation de la langue.
- On va pas entrer dans des détails sur les produits et technologies recensés dans le rapport…

META-NET: résultats pour le basque

	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability
Language Technology: Tools, Technologies and Applications							
Speech Recognition	2	1	1	1	4	3	2
Speech Synthesis	2	3	4	4	4	3	3
Grammatical analysis	4	2.5	4	4	4	2.5	2.5
Semantic analysis	1	1.5	2	1	1	1	1
Text generation	1	0	0	0	0	0	0
Machine translation	3	5	2	3	3	2	2
Language Resources (Resources, Data and Knowledge Bases)							
Text corpora	2	4	3	2	3	4	2.5
Speech corpora	3	2	3	2	3	3	2
Parallel corpora	2	4	2	2	2	2	1
Lexical resources	4	4	4	5	5	4	3
Grammars	2	2	2	2	2	2	2

7: State of language technology support for Basque

META-NET: résultats pour le français

	Quantité	Disponibilité	Qualité	Couverture	Maturité	Pérennité	Adaptabilité
Technologies de la langue							
Reconnaissance de la parole	4	3	4	4	4	3	3
Synthèse vocale	4	3	4	4	4	3	3
Analyse grammaticale	4	4	4	4	4	3	3
Analyse sémantique	3	3	3	3	3	2	2
Génération de texte	3	2	3	3	3	2	2
Traduction Automatique	5	4	4	4	4	3	3
Ressources linguistiques							
Corpus de textes	4	3	4	4	4	4	3
Corpus de parole	4	3	4	4	4	4	3
Corpus parallèles, Mémoires de traduction	4	3	4	4	4	4	3
Ressources lexicales	4	3	4	4	4	4	3
Grammaires, Modèles de langage	3	3	4	4	3	3	3

14 : Tableau réduit de la situation estimée des technologies de la langue et des ressources linguistiques pour le français.

META-NET: résultats pour l'anglais

	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability
Language Technology: Tools, Technologies and Applications							
Speech Recognition	5	3	5	5	4	2	3
Speech Synthesis	5	3	4.5	5.5	4	2	3
Grammatical analysis	5	5	5.5	4.5	4.5	3	4
Semantic analysis	3	2	3	3	2.5	2	2
Text generation	3	3	3.5	2.5	2.5	2	2.5
Machine translation	4	4	3.5	4	4	2	2
Language Resources: Resources, Data and Knowledge Bases							
Text corpora	5	4	5.5	4	5	2.5	4
Speech corpora	5	2	6	5.5	5	3	3
Parallel corpora	4.5	4.5	5	5	3.5	3	3
Lexical resources	4	6	5	5	4.5	4.5	4.5
Grammars	3.5	2.5	4	4	2.5	4	1.5

8: State of language technology support for English

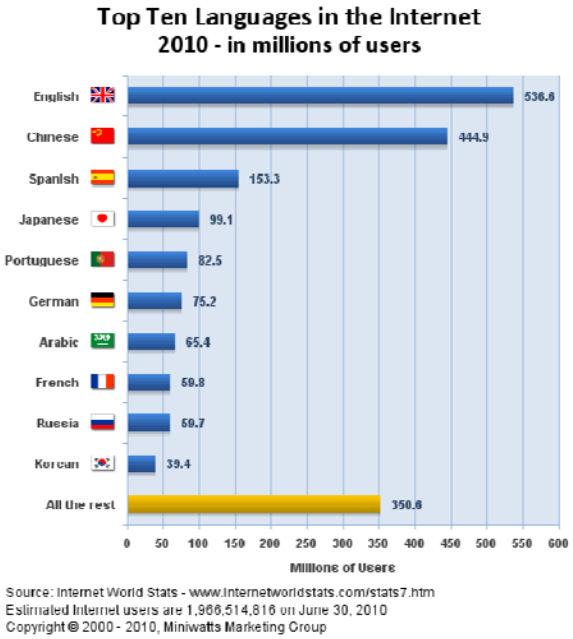
Comment les langues font-elles face aux défis des TIC et des TL ?

- Il n'est pas facile à obtenir des chiffres concernant les quantités de ressources pour les différentes langues sur l'Internet.
- On devra utiliser des données publiques plus spécifiques :
  - nombre d'usagers
  - nombre de documents sur l'Internet
  - nombre d'articles dans la Wikipedia
  - etc.

Comment les langues font-elles face aux défis des TIC ?

Nombre d'usagers
[*Internet World Stats 2010*]

- Anglais :
  - 536 millions d'usagers
  - 27%
- *Top 10 langues* :
  - 1.616 millions d'usagers
  - 82%
- Le reste des langues :
  - 351 millions d'usagers
  - 17,8% des usagers



Comment les langues font-elles face aux défis des TIC ?

- Nombre de documents sur le Web
  - Il y peu de statistiques fiables pour les différentes langues
  - Une étude sur la présence des langues romanes (Latin Union, 2007) indiquait :
    - 45% des pages Web sont écrites en anglais
    - 7,80% en espagnol
    - 5,9% en allemand
    - 4,41% en français
    - 2,66% en italien
    - 1,39% en portugais
    - 0,28% en roumain
    - 0,14% en catalan
    - …

Comment les langues font-elles face aux défis des TIC ?

Nombre d'entrées dans Wikipedia

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

- Articles en 286 langues (Juin 2014).
- *Top 10* :
  - Anglais: 4,54 millions d'articles
  - Hollandais : 1,78 M
  - Allemand : 1,73 M
  - Suédois : 1,63 M
  - Français : 1,52 M
  - Italien, russe, espagnol, polonais et waray-waray.
- Après:
  - 14ème: Portugais (830 K)
  - 17ème: Catalan (429 K)
  - 35ème: Basque (181 K)
  - 54ème: Occitan (87 K)
  - 71ème: Bréton (50 K)
  - …

Comment les langues font-elles face aux défis des TL ?

- Plusieurs référentiels publics (ressources et outils) :
  - ELRA : *European Language Resources Association*
  - LDC : *Linguistic Data Consortium*
  - ACLWiki : *Association for Computational Linguistics*
  - NLSR : *Natural Language Software Registry* (DFKI)
  - *yourdictionary.com* : site web de dictionnaires
  - …

Comment les langues font-elles face aux défis des TL ?

- Ces sources d'information ne sont pas toujours complètes
  - les référentiels mentionnent toujours les produits qu'ils offrent
  - ils gèrent les ressources et vendent certains d'entre eux
  - les sites du type wiki sont gérés par des volontaires (valables juste pour consultation)
- On peut trouver des ressources et d'outils pour le basque dans ces référentiels : 6 dans ELRA, 15 dans ACLWiki, 3 dans NLSR, 9 dicos dans *yourdictionary.com*…

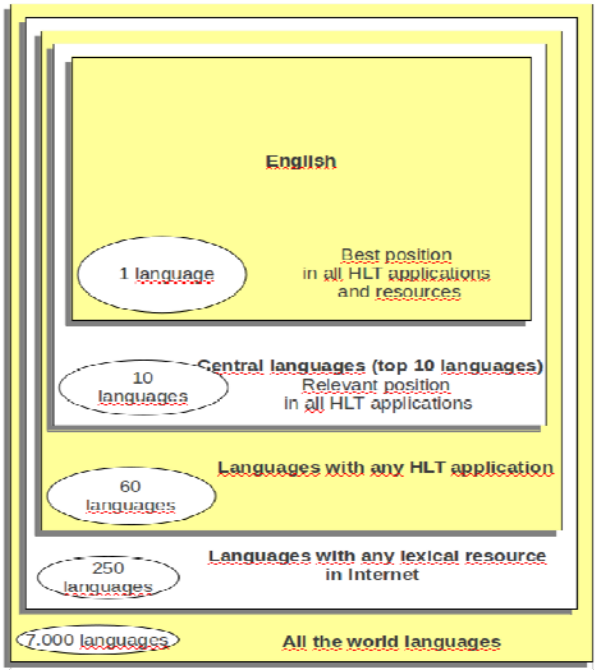
Comment les langues font-elles face aux défis des TL ?

- Présence/absence des TL dans des services les plus populaires :
  - traitement de texte
  - moteurs de recherche
  - traduction automatique
  - …
- Traitement de texte : le basque est présent dans les deux programmes les plus utilisés (vérification et correction d'orthographe, notamment).
  - *MS Word* : 91 langues
  - *Libreoffice* : 104 langues
- Moteurs de recherche :
  - *Google* : ~50 langues sont identifiées
- Systèmes de traduction automatique :
  - *Babelfish* : 14 langues
  - *Google Translate* : ~80 langues (y compris le basque)

Les langues du monde et leurs ressources langagières

Quelles sont les langues avec « moins de ressources » ?

- La réponse est relative
- On peut distinguer six niveaux différents



Les langues du monde et leurs ressources langagières

- Premier niveau : l'anglais.
  - 27% des usagers de l'Internet.
  - 45% des pages web.
  - 62% des ressources langagières dans le LDC.
  - 51% des ressources langagières dans ELRA.
  - Pratiquement tous les types d'applications du langage existent pour l'anglais.

Les langues du monde et leurs ressources langagières

- Deuxième niveau : *top 10* langues dans le Web
  - 82% des usagers de l'Internet (y compris l'anglais).
  - Le développement actif de ressources langagières continue.
  - La plupart des applications de la TL y sont représentées.
  - La plupart des ressources décrites dans LDC ou ELRA sont disponibles pour ces langues :
    - 45,79% pour l'allemand, 41,27% pour le français, 40,76% pour l'espagnol, 36,24% pour l'italien, 31,31% pour le portugais…
  - Streiter *et al.* (2006) utilisent le terme *central languages* pour se référer a cet ensemble de langues.

Les langues du monde et leurs ressources langagières

- Troisième niveau : les langues qui possèdent une ou plusieurs applications de technologie langagière
  - 60 langues dans ELRA
  - 82 dans LDC
  - 73 dans ACLWiki
  - 30 dans NLSR

Les langues du monde et leurs ressources langagières

- Quatrième niveau : des langues qui possèdent des ressources lexicales, voire des dictionnaires, en ligne
  - 307 langues in *yourdictionary.com*
  - Pratiquement le même ensemble de langues présentes dans la Wikipedia (286 langues).

Les langues du monde et leurs ressources langagières

- Cinquième niveau : des langues qui possèdent de système d'écriture (Borin, 2009)
  - 2.000+ langues
- Sixième niveau : des langues non-écrites
  - 4.500+ langues

Les langues du monde et leurs ressources langagières

- Cette typologie de 6 niveaux nous donne une définition relative de « langue avec peu de ressources » :
  - En comparant avec l'anglais, on peut considérer toutes les autres langues comme ayant peu de ressources.
    - Ou... sauf les *top 10* langues, le reste peut être considéré comme ayant peu de ressources.
  - Les langues des niveaux 3ème et 4ème sont des langues considérées comme ayant peu de ressources dans le domaine des TL.
  - On peut considérer que les langues des niveaux 5ème et 6ème sont vraiment en danger (du point de vue de leur utilisation dans les TIC).

Les langues du monde et leurs ressources langagières

- Cette classification n'est pas stricte…
- …mais elle peut être utile pour reconnaître des domaines d'application et pour dessiner d'éventuelles stratégies pour le développement des ressources langagières.

Les langues du monde et leurs ressources langagières

- Et il y a des risques en ce qui concerne l'application de ces indicateurs :
  - Langues avec des promoteurs très actifs peuvent avoir une grande visibilité sur Wikipedia, n'étant pas cependant significative de la présence de la langue sur l'Internet en général, ou du nombre et de la qualité des ressources langagières.
    - Par exemple, le catalan apparaît dans une bonne position dans le classement du nombre d'articles de la Wikipedia, mais il s'agit d'une langue généralement considérée comme ayant peu de ressources.
  - Néanmoins, l'indicateur Wikipedia est très accessible, car il est mis à jour automatiquement pour toutes les langues, et utile lorsqu'il est utilisé en conjonction avec d'autres indicateurs.

Ixa : groupe de recherche en TALN

- Groupe Ixa : groupe de recherche créé en 1988, à la Faculté d'Informatique de Saint-Sébastien (UPV / EHU) http://ixa.si.ehu.es
- Objectif principal : faire face au défi de l'adaptation du basque aux technologies de la langue, établir une infrastructure (ressources et outils) pour le traitement automatique du basque
  - 1988 : 5 enseignants d'université (informatique)
  - 2014 : équipe interdisciplinaire
    - 45+ informaticiens, 15+ linguistes et 3 assistants de recherche
    - ~30 enseignants de l'université

Stratégie de développement des TL pour le basque (groupe Ixa)

- Nous avons présenté, déjà en 1998 (Aduriz *et al.*, 1998) une proposition ouverte pour faire des progrès dans les TL.
- Idée principale: *ne pas mettre la charrue avant les bœufs!*
  - **d'abord, les fondations => après, les applications**
- Les mesures proposées ne se correspondent pas exactement à celles observées dans l'histoire du traitement automatique de l'anglais. Les ressources langagières dans le cas de l'anglais…
  - n'ont pas évolué à la suite d'un plan unique et coordonné
  - beaucoup d'efforts indépendants ont produit ces ressources, à fin de répondre aux besoins spécifiques de projets concrets
- Les ressources pour le traitement du basque ont été développées d'une façon différente, plus planifiée (au sein du groupe).

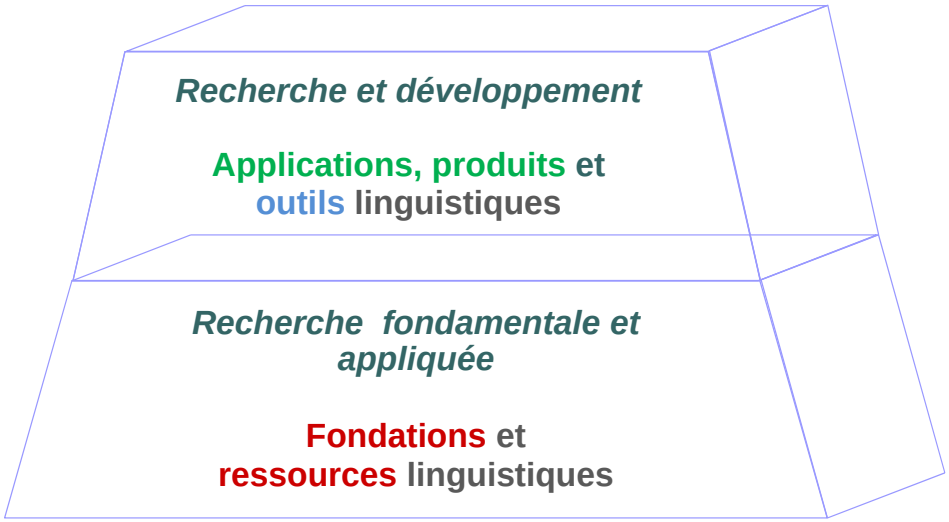
Stratégie de développement des TL pour le basque (groupe Ixa)

- Conception et développement des bases de la technologie langagière, des outils et des applications
  - d'une manière progressive et planifiée
  - afin d'en tirer le meilleur bénéfice
- Normalisation des ressources afin de les utiliser:
  - dans des recherches variées
  - pour développer des outils divers
  - dans des applications et produits différents
  - adoption de TEI et de standards comme XML comme base pour l'étiquetage linguistique aux différents niveaux de traitement (méthodologie générale pour l'annotation des corpus)

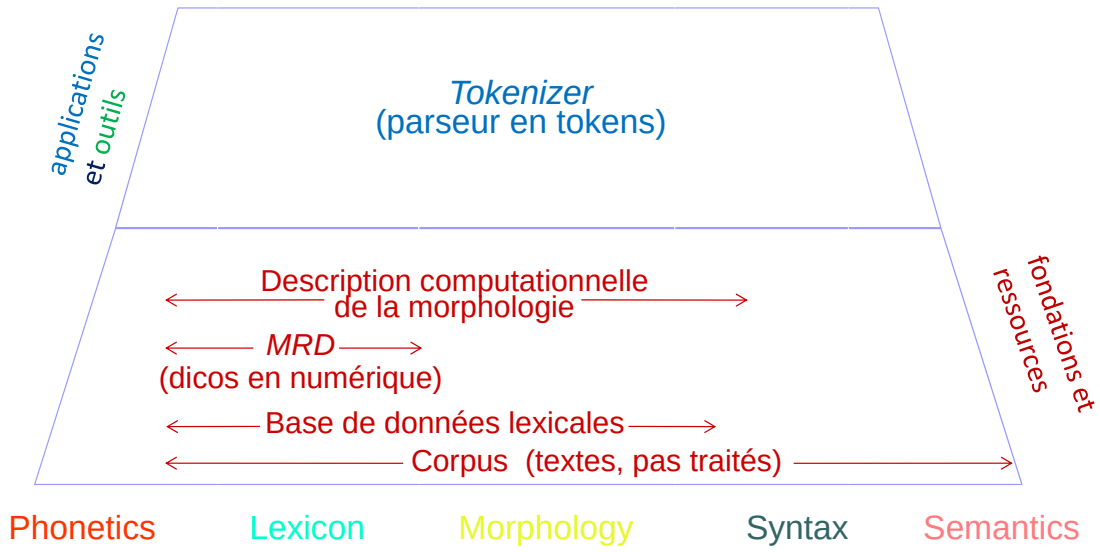
Stratégie de développement des TL pour le basque (groupe Ixa)

- En prenant comme référence notre expérience dans la conception et le développement de ressources et d'outils :
  - nous proposons un stratégie général à quatre phases pour le développement d'une infrastructure du traitement automatique d'une langue (Alegria *et al.*, 2011)

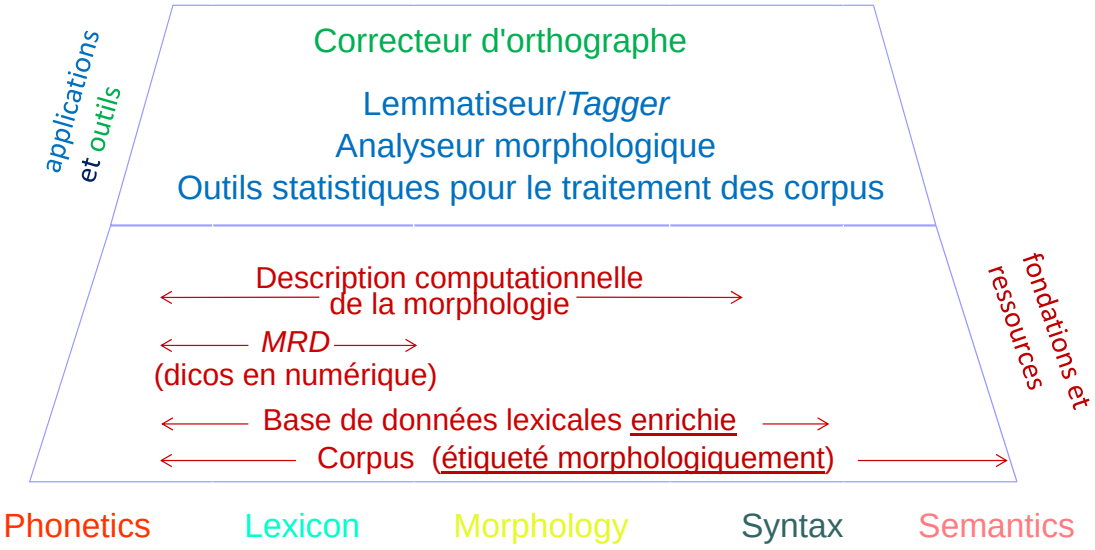
La priorité stratégique: de la recherche fondamentale vers le développement d'applications



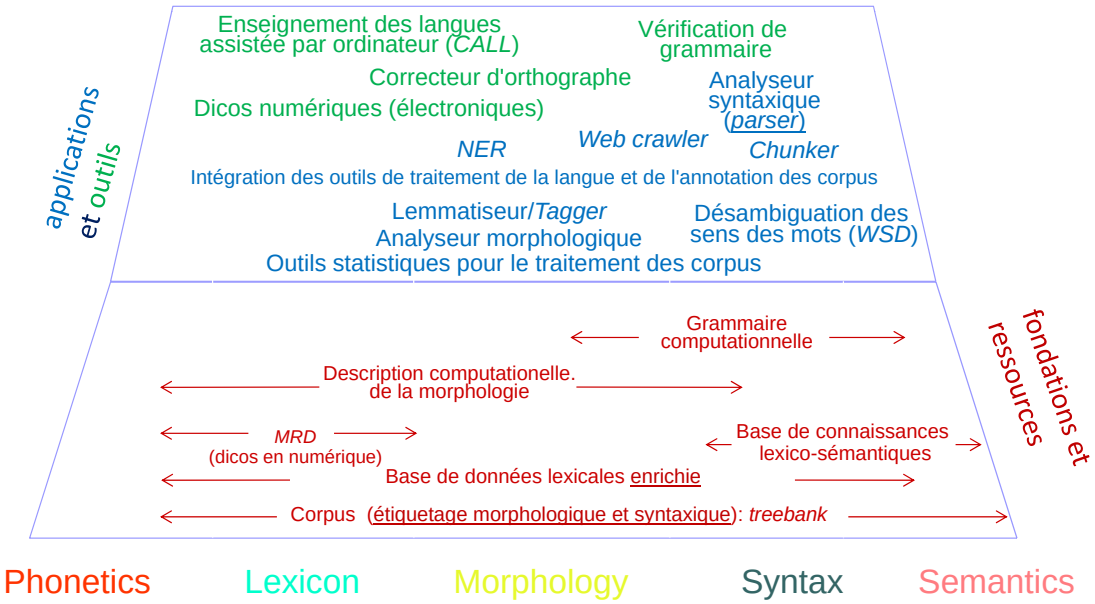
Phase I: pose des fondations



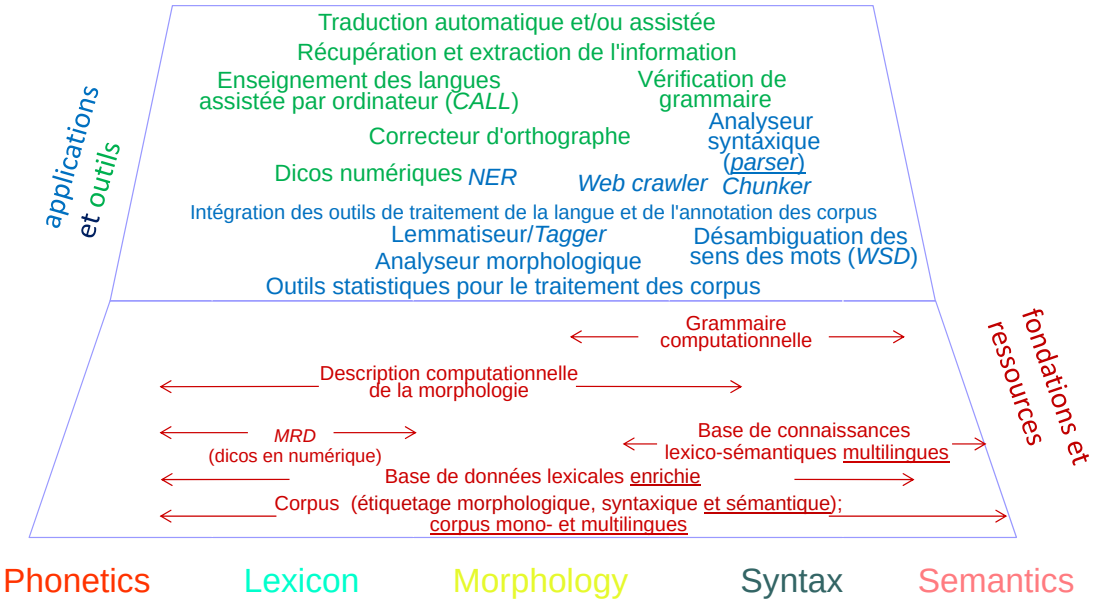
Phase II: premiers outils basiques et applications



Phase III: des outils et des applications plus avancées



Phase IV: applications multilingues et générales



Groupe Ixa : création de ressources, outils et applications

- 1988-1996
  - *EDBL* : base de données lexicales d'usage général
  - Description de la morphologie du basque (*two-level morphology*)
  - *Morfeus* : analyseur morphosyntaxique
  - *Xuxen* : vérificateur et correcteur d'orthographe

Groupe Ixa : création de ressources, outils et applications

- 1997-2005
  - *Euskal WordNet* : wordnet du basque
  - *EPEC* : (petit) corpus général de référence
  - *Erreus* : corpus d'erreurs (apprentissage de la langue)
  - *PATR-IXA* : grammaire computationnelle, syntaxe (constituants)
  - *Eustagger* (*EusLem*) : lemmatiseur/*tagger*
  - *Ixati* (*Zatiak*) : *chunker*, identificateur de syntagmes et chaînes verbales
  - *Eihera*: identificateur/classeur d'entités nommées (noms, prénoms, dates…)
  - Intégration de *Xuxen* dans de divers environnements (traitement de texte, navigateurs, etc.) et version en ligne
  - Dictionnaires électroniques intégrés dans des traitement de textes (Elhuyar-Word, *eu-es*, *eu-fr*; UZEI-Word, synonymes)
  - *Multimeteo* : génération automatique des prévisions météorologiques

Groupe Ixa : création de ressources, outils et applications

- 2006-
  - *ZTc* : Corpus de Science et Technologie (usager final)
  - *LB* : Observatoire du Lexique (corpus, usager final)
  - *MCR* : Multilingual Central Repository (EuroWordnet)
  - *EPEC-EuSemCor* : *EPEC* étiqueté avec des sens des mots (wordnet du basque)
  - *EPEC-AnCora* : *EPEC* intégré dans AnCora, avec des corpus *es* et *ca,* étiqueté syntaxiquement (dépendances)
  - *EDGK* : grammaire de dépendances (règles)
  - *Basyque* : Base de Données Syntaxique Basque (usager final)
  - *e-ROLda* : outil de consultation de verbes (arguments, rôles)
  - *Euskal RST Treebank* : petit corpus annoté au niveau du discours

Groupe Ixa : création de ressources, outils et applications

- 2006-
  - *Maltixa* : analyseur syntaxique de dépendances (statistique)
  - *libiXaml* : librairie basique d'annotation linguistique
  - *UKB* : collection de programmes pour la désambiguïsation des sens des mots (indépendante de la langue)
  - *WSD-IXA* : système de désambiguïsation des sens des mots pour le basque (en ligne)
  - *Eulia / Armiarma* : outils de consultation et traitement du corpus (étiquetage, désambiguïsation)
  - *lexKit* : environnement d'édition de dictionnaires

Groupe Ixa : création de ressources, outils et applications

- 2006-
  - *Anhitz* : expert virtuel (3D) en science et technologie (*Question Answering, MT, IE/IR*)
  - *Matxin (KBMT) / EusMT (SMT)* : traduction automatique *es-eu* (basée sur la connaissance / statistique)
  - *Ihardetsi* : système de réponse aux questions en langage naturel (*Question Answering*)
  - *BertsolariXa* : système de recherche de mots rimés
  - *Berbatek Tutor* : tuteur personnel pour l'enseignement de la langue (exercices de grammaire et de compréhension à la lecture)
  - *Berbatek Dubbing* : doublage automatiques de documentaires

Sans quoi on ne peut pas se passer…

...si on veut traiter la langue écrite :

- Base de données lexicales
- Lemmatiseur/*tagger*
- Corpus

et (après) ils viendront :

- l'analyse syntaxique, sémantique…
- l'identification des entités nommées
- …
- et les applications et produits, bien sûr!

Une question importante : existe-t-il de langue standard?

Groupe Ixa : lignes de recherche actuelles

- Recherche fondamentale en lexicographie, morphologie, syntaxe et sémantique computationnelles
- Recherche sur le discours et les aspects pragmatiques de la langue (coréférence, structure rhétorique du discours)

- Recherche fondamentale sur les aspects opérationnels du traitement du langage : traitement de grands collections de textes, traitement parallèle…
- Annotation linguistique des corpus

- Récupération et extraction d'informations: réponse aux questions, résumé automatique de textes… sur domaines divers (médecine, tourisme, financier…)
- Traduction automatique
- Apprentissage des langues

Groupe Ixa : recherche, résultats et projets

- ~50 publications annuelles (congrès et revues).
- Impliqué dans la création de la société *spin-off Eleka*, de la Fondation Elhuyar (2002).
- Collabore actuellement avec plusieurs entreprises du Pays Basque et de l'étranger.
- On travaille sur le basque, mais aussi sur d'autres langues (notamment sur l'anglais).
- Projets actifs :
  - Communauté Européenne : 6
  - *Ministerio de Economia y Competitividad* (Espagne) : 3
  - *Eusko Jaurlaritza* (Gouvernement Basque):
    - 1 projet ETORTEK (2012-2014) : recherche stratégique dans la CAPV
    - Groupe de recherche consolidé (2010-2015)
  - Avec d'autres entités : 2

Enseignement

- Dégrée en Informatique : *Traitement du langage naturel* (matière facultative, depuis 1994)
- Masters
  - Hiztek (Diplôme spécialisé en Technologie de la Langue; UPV/EHU + UEU): années 2001/2005
  - HAP, master sur le TALN : années 2005/2014
  - Erasmus Mundus master on *Language and Communication Technologies* + HAP/LAP master on *Language Analysis and Processing* (basque et anglais) : à partir de 2014
- Programmes de doctorat sur le TALN : 11 thèses soutenues dans les cinq dernières années

En promouvant la coopération entre les divers acteurs liés aux Industries de la Langue

LANGUNE
ASSOCIATION DES INDUSTRIES DE LA LANGUE DU PAYS BASQUE

- **Langune,** association d'entreprises crée en 2010 (35+ entreprises): www.langune.com
  - Encourager, renforcer et fournir un cadre cohérent à l'Industrie de la Langue en Euskal Herria (Pays Basque), afin principalement d'améliorer la compétitivité et la visibilité de ses associés.
- L'industrie linguistique est le secteur d'activité chargé de concevoir, produire et commercialiser des produits et des services en rapport avec le traitement des langues.

En promouvant la coopération entre les divers acteurs liés aux Industries de la Langue

- On parle ici de:
  - Entreprises de traduction, localisation de logiciel, doublage et sous-titrage…
  - Apprentissage de la langue: enseignement en ligne, certifications sur la connaissance et l'usage des langues, etc.
  - Multilinguisme et gestion de contenus, ressources langagières…
- Défis stratégiques
  - Création et développement de l'association
  - Coopération et compétitivité des entreprises associées
  - Internationalisation
  - Développement technologique et innovation
- En 2012, le Département de l'Industrie, l'Innovation, le Commerce et le Tourisme du Gouvernement Basque a concédé *Langune* le titre de *Cluster* des Industries de la Langue.

Conclusions

- De notre expérience, nous défendons que la recherche et le développement pour les langues avec « moins de ressources » devraient suivre ces points:
  - conception et développement progressifs : *bottom up*
  - réutilisation des fondations, des ressources et des outils
  - normalisation
  - *open-source* (code source ouvert): si on est peu, il faut partager!
- Nous pensons que notre stratégie visant à développer des technologies de la langue pourrait être utile pour d'autres langues : celles qui possèdent une norme écrite? celles qui ont déjà quelques ressources lexicales initiales?

Conclusions

- Nous pensons que si le basque est maintenant dans une assez bonne position dans le domaine des technologies de la langue est parce que pendant les 25 dernières années ces lignes directrices ont été appliquées…
  - même quand il était plus facile de construire des ressources et des outils « jouet », utiles pour obtenir de « bons résultats académiques » à court terme, mais pas toujours réutilisables dans des développements futurs
- Il y a des expériences similaires avec d'autres langues : le cas du tchèque, par exemple, est une autre exception : il y a un bon nombre de ressources langagières pour le tchèque, grâce aux efforts coordonnés de certains chercheurs ambitieux et productifs.

Conclusions

- La recherche ciblée sur chaque langue et sur l'application des techniques générales au traitement de chaque langue sont nécessaires; par ailleurs, elle contribue à la recherche générale sur le TALN.
- Un langage qui cherche à survivre dans la société de l'information exige des produits de technologie de la langue.

Quelques références

- Aduriz, I., Agirre, E., Aldezabal, I., Alegria, I., Ansa, O., Arregi, X., Arriola, J.M., Artola, X., Díaz de Ilarraza, A., Ezeiza, N., Gojenola, K., Maritxalar, M., Oronoz, M., Sarasola, K., Soroa, A., Urizar, R.. A framework for the automatic processing of Basque. In: *Proceedings of Workshop on Lexical Resources for Minority Languages* (1998).
- Alegria I., Aranzabe M., Arregi X., Artola X., Díaz de Ilarraza A., Mayor A., Sarasola K.. Valuable Language Resources and Applications Supporting the Use of Basque. In: Z. Vetulani (Ed.) : LTC 2009, LNAI 6562, pp. 327–338, 2011. Springer-Verlag, Berlin Heidelberg : 2011.
- Borin, L.. Linguistic diversity in the information society. In: *SALTMIL 2009 Workshop: IR-IE-LRL Information Retrieval and Information Extraction for Less Resourced Languages*. Université du Pays Basque (2009).
- Krauwer, S.. The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap. In: *International Workshop Speech and Computer*, Moscou, Russia (2003).
- META-NET, La collection des livres blancs : http://www.meta-net.eu/whitepapers/overview
- Streiter, O., Scannell, K., Stuflesser, M.. Implementing NLP projects for non-central languages: instructions for funding bodies, strategies for developers. *Machine Translation* 20 (4), 267–289 (2006).

merci de votre attention

mercés hèra hòrt

eskerrik asko

xabier.artola@ehu.es

ixa.si.ehu.es