Sommaire

1 - Résumé

Le lexique *Morphalou* est un lexique ouvert des formes fléchies du français. Il s'agit d'un lexique extensionnel, c'est-à-dire d'un lexique qui liste explicitement toutes les formes fléchies d'un lemme. Les données initiales (celles de la version 1.0) proviennent du *TLFNome*, la nomenclature du Trésor de la Langue Française, produit par le laboratoire ATILF (Nancy Université - CNRS), qui a fourni 540.000 formes fléchies, réparties sur à peu près 68.000 lemmes. Dans la version 2.0, la couverture a été augmentée par environ 30.000 lemmes, extraits automatiquement du Trésor de la Langue Française Informatisé. Il s'agit essentiellement de formes composés, dérivées ou locutions n'ayant pas un statut d'entrée principale dans le TLFi. Par conséquent, le calcul automatique des informations lexicales (forme de l'entrée, catégorie grammaticale) a pu introduire certaines erreurs (cf. liste des problèmes connus), dues aux structures idiosyncrasiques du TLF. Elles seront corrigées au fur et à mesure par des lexicographes. Actuellement, le lexique comporte 95.810 entrées lexicales. Le format de représentation suit les recommendations de normalisation pour les ressources lexicales du TAL à l'ISO (TC 37/SC 4) : il s'agit d'une instanciation actualisée du *Lexical Markup Framework*, version 9. *Morphalou 2.0* est en accès libre sous acceptation de la licence. L'hébergement, la maintenance et la mise à jour du lexique sont assurés par ORTOLANG.

2 - Contributions et remerciements

Morphalou 2.0 a été dévéloppé par

- Christiane Jadelot (ATILF) : traitement lexicographique
- Mathieu Mangeot (ATILF, puis Université de Chambéry) : ajouts du TLFi
- Etienne Petitjean (ATILF) : développement informatique
- Susanne Salmon-Alt (ATILF) : coordination et conception

Morphalou 2.0 a bénéficié de l'aide et des contributions de Laurent Romary (INRIA), Ingrid Falk (Loria), Pascale Bernard (ATILF), Lou Burnard et Sebastian Rahtz (Oxford) et Jean-Marie Pierrel (ATILF), ainsi que de nombreux encouragements venus des utilisateurs de la version 1.0. Le projet a été soutenu de 2004 à 2006 dans le cadre du CPER Lorraine.

3 - Couverture linguistique et décisions lexicographiques

Actuellement, *Morphalou 2.0* couvre :

- 60.940 noms communs
- 8790 verbes
- 22790 adjectifs
- 1579 adverbes
- 1.450 mots fonctionnels (classes fermées)
- 193 interjections
- 68 onomatopées

3.1 Lemmes et formes fléchies

Morphalou 2.0 contient des entrées lexicales <lexicalEntry>. En théorie, une entrée lexicale est composée d'informations sur la forme et le sens de la lexie décrite. Pour l'instant, seuls les aspects relatifs à la forme sont traités dans *Morphalou 2.0*. Ils sont regroupés dans un élément <formSet>, contenant un lemme <lemmatizedForm> et des formes fléchies <inflectedForm>. Une forme fléchie correspond à un *mot-forme (word form)* au sens de Polguère (2003). La notion de *lemme* a été définie comme une forme arbitraire, mais conventionnelle, abstraite sur l'ensemble des formes d'un paradigme flexionnel (cf. *lexème* chez Polguère 2003).Un lemme <lemmatizedForm> est décrit par son orthographe, une categorie grammaticale et, pour les noms, un genre grammatical. Il est également prévu d'y ajouter sa prononciation ainsi que sa fréquence dans des corpus de référence. Une forme fléchie <inflectedForm> comporte, en plus, un ensemble de traits flexionnels (mode, temps, genre, personne, nombre).

```
<lexicalEntry>
 <formSet>
  <lemmatizedForm>
   <orthography>actrice</orthography>
   <grammaticalCategory>commonNoun</grammaticalCategory>
   <grammaticalGender>feminine</grammaticalGender>
  </lemmatizedForm>
  <inflectedForm>
   <orthography>actrice</orthography>
   <grammaticalNumber>singular</grammaticalNumber>
  </inflectedForm>
  <inflectedForm>
   <orthography>actrices</orthography>
   <grammaticalNumber>plural</grammaticalNumber>
  </inflectedForm>
 </formSet>
</lexicalEntry>
```

Concernant la variation flexionnelles, deux cas peuvent se présenter: la variation systématique (deux paradigmes flexionnels pour un même lemme) et la variation accidentelle (variation particulière à une forme fléchie).

- Variation systématique : pour un même lemme appartenant simultanément à deux paradigmes flexionnels ("une alto - des altos / des alti"), le lexique pourrait contenir, en principe, autant de paradigmes flexionnels <formSet> que nécessaires pour une même entrée lexicale. Toutefois, dans une perspective purement extensionnelle, *Morphalou* les traite comme variations accidentelles.
- Variation accidentelle : Lorsqu'une des formes fléchies d'un lemme présente une variation non systématique en langue ("courbaturer - courbattu / courbaturé"), nous parlons d'une variante accidentelle. Dans la perspective extensionnelle du lexique, cette variante est simplement listée parmi les formes fléchies du lemme. Dans un lexique intensionnel - qui indiquerait l'appartenance à des paradigmes plutôt que toutes les formes fléchies - il y aurait deux possibilités de traitement : soit, la création d'un nouveau paradigme ad hoc (cas de "assoir" dans le Bécherelle), soit le renvoi à un paradigme existant, en ajoutant explicitement les formes supplémentaires (cas de "coubaturer" dans le Bécherelle).

```
<lexicalEntry>
  <formSet>
   <lemmatizedForm>
    <orthography>lied</orthography>
    <grammaticalCategory>commonNoun</grammaticalCategory>
    <grammaticalGender>masculine</grammaticalGender>
   </lemmatizedForm>
   <inflectedForm>
    <orthography>lied</orthography>
    <grammaticalNumber>singular</grammaticalNumber>
   </inflectedForm>
```

```
    <inflectedForm>
     <orthography>lieds</orthography>
     <grammaticalNumber>plural</grammaticalNumber>
    </inflectedForm>
    <inflectedForm>
     <orthography>lieder</orthography>
     <grammaticalNumber>plural</grammaticalNumber>
    </inflectedForm>
   </formSet>
  </lexicalEntry>
```

Morphalou 2.0 contient un certain nombre de lexies complexes :

- Mots composés : "casse-pied", "pomme de terre", "cervico-spinal", "parce que", etc. Ces lexèmes comportent des blancs ou des tirets, selon l'orthographe caclulée d'après le TLFi.
- Locutions : "à la je-m'en-fiche", "up to date", etc. Ces lexies comportent des blancs ou des tirets, selon l'orthographe calculée d'après le TLFi. L'ordre des composants n'est pas toujours canonique et dépend de l'encodage dans le TLF.
- Verbe pronominaux : "se pommeler", "se abader", etc. La forme lemmatisée des verbes pronominaux commence toujours par "se" (même pour les verbes commençant par une voyelle), suivi d'un espace, puis du verbe. Les formes fléchies ne comportent pas de pronom réflexif.

NB : Un certain nombre de lemmes dans *Morphalou 2.0* ne sont pas encore assortis de leurs formes fléchies. Dans ce cas, seule la <lemmatizedForm> y figure.

3.2 Lemmes et variantes

Dans *Morphalou 2.0*, toute forme de variation sur un lemme (contrairement à une forme fléchie, cf. supra) donne lieu à une nouvelle entrée lexicale. Ceci concerne en particulier les variantes orthographiques ("cheik - cheikh"), les abréviations et troncations ("sanatorium - sana") et les variantes morphologiques ("soprano - soprane"). Cette décision se justifie au regard de notre définition de lemme, et plus précisément par le fait que ces variantes sont en général associées à un paradigme flexionnel qui leur est propre. Lorsque la sémantique est commune à un ensemble d'entrées lexicales, il est possible de les relier par un pointeur typé. Pour l'instant, nous utilisons <spellingVariantOf> pour relier des variantes orthographiques entre elles. Ce pointeur part d'une variante et renvoie à une forme arbitrairement choisie comme étant la forme principale (en général, la première forme dans l'ordre alphabétique.)

```
<lexicalEntry>
 <spellingVariantOf>clef</spellingVariantOf>
 <formSet>
  <lemmatizedForm>
   <orthography>clé</orthography>
   <grammaticalCategory>commonNoun</grammaticalCategory>
   <grammaticalGender>feminine</grammaticalGender>
  </lemmatizedForm>
  <inflectedForm>
   <orthography>clé</orthography>
   <grammaticalNumber>singular</grammaticalNumber>
  </inflectedForm>
  <inflectedForm>
   <orthography>clés</orthography>
   <grammaticalNumber>plural</grammaticalNumber>
  </inflectedForm>
 </formSet>
</lexicalEntry>
```

Par ailleurs, les lexiques à l'origine de *Morphalou 2.0* disposaient d'une information indiquant la dérivation nominale pour les formes féminines ("avocat - avocate"). Bien que la portée d'une telle information dépasse le cadre stricte d'un lexique morphologique flexionnel (cf. ci-dessous), ces liens ont été maintenus dans Morphalou. Toutefois, plusieurs raisons linguistiques nous ont amenés à ne pas considérer le genre comme un trait flexionnel des noms communs en français (notamment le caractère non paradigmatique de la variation, la différence du sens dénotationnel et des collocations potentiellement différentes). Par conséquence, les formes féminines des noms communs concernés ne sont pas codées en tant que formes fléchies du lemme masculin, mais en tant que lemmes à part entière. Le lien de dérivation est maintenu par un trait <feminineVariantOf> reliant la forme féminine au lemme de son correspondant masculin. Idéalement, ce lien devrait par ailleurs pointer sur un sens particulier de cette forme masculine ("avocat" : profession), mais c'est ici que l'intégration d'informations dérivationnelles va au-delà d'un lexique flexionnel au sens strict. La génération des nouvelles entrées pour les variantes féminines à été opérée semi-automatiquement : dans une première passe, toutes les formes nominales féminines renvoyant à un lemme masculin non homographe ont été considérées comme féminisations : avocate - avocat, illettrée - illettré etc. Ensuite, les formes ambiguës en genre à lemme homographe (un/une absentéiste) ont été considérées comme candidats potentiels à la féminisation par dérivation. Elles ont été triées manuellement pour séparer les féminisations par conversion (un/une absentéiste) des homographes sans lien dérivationnel (un/une voile).

```
<lexicalEntry>
 <feminineVariantOf>acteur</feminineVariantOf>
 <formSet>
  <lemmatizedForm>
   <orthography>actrice</orthography>
   <grammaticalCategory>commonNoun</grammaticalCategory>
   <grammaticalGender>feminine</grammaticalGender>
  </lemmatizedForm>
  <inflectedForm>
   <orthography>actrice</orthography>
   <grammaticalNumber>singular</grammaticalNumber>
  </inflectedForm>
  <inflectedForm>
   <orthography>actrices</orthography>
   <grammaticalNumber>plural</grammaticalNumber>
  </inflectedForm>
 </formSet>
</lexicalEntry>
```

NB : Un certain nombre de liens de variation orthographique ou de féminisation peuvent manquer ou être faux. Ces liens ont été déduits (semi-)automatiquement des sources du TLFi, et toutes les configurations n'ont pas pu faire l'objet d'un contrôle manuel.

Le calcul de l'orthographe des formes composées avec un tiret a fait l'objet d'un traitement particulier. Compte tenu de leur encodage sous-spécifié dans le TLFi (encodage du tiret entre parenthèses, signifiant une forme possible avec tiret et/ou une forme séparée par un espace et/ou une forme sans séparation graphique, cf. "co(-)souveraineté", "papier(-)ministre", etc.), l'explicitation des variantes n'a pas pu se faire de façon uniforme et systématique. Pour (1) ne pas proposer une surgénération systématique sous forme de trois variantes orthographiques, et (2) ne pas laisser des formes orthographiques de surface contenant des signes méta-linguistiques (i.e. des parenthèses), nous avons décidé de proposer, dans *Morphalou 2.0*, seulement la forme avec tiret. Pour signaler que ces formes peuvent avoir d'autres orthographes (cf. <originatingEntry>, l'orthographe est assortie d'un marqueur spécial *provisionallyProcessed*.

```
<lexicalEntry>
 <formSet>
```

```
        <lemmatizedForm>
          <orthography processStatus="provisionallyProcessed">prêt-à-porter</orthography>
          <grammaticalCategory>commonNoun</grammaticalCategory>
          <grammaticalGender>masculine</grammaticalGender>
        </lemmatizedForm>
      </formSet>
      <originatingEntry target="68340">
        PRÊT(-)À(-)PORTER,(PRÊT À PORTER, PRÊT-À-PORTER)subst. masc.
      </originatingEntry>
    </lexicalEntry>
```

3.3 Les mots fonctionnels

Morphalou 2.0 contient des entrées lexicales catégorisées en tant que mots grammaticaux (/functionWord/). Cette classe comprend les pronoms (personnels, indéfinis, démonstratifs, relatifs, interrogatifs, numériques...), les conjonctions (coordinatives et subordinatives), certains adverbes (de quantité, interrogatifs, présentatifs), les déterminants (indéfinis, définis, possessifs, démonstratifs, numéraux), les prépositions, les formes contractées ("aux", "desquels", "du") et des inclassables ("ergo"). Certaines de ces formes possèdent des variantes en genre ou en nombre (les pronoms et les déterminants), des variantes allomorphiques ("je" : "j' "; "lorsque" : "lorsqu'") ou des variantes morpho-phonétiques ("le": "au" "du" "des" "aux"). Actuellement, ces variantes sont majoritairement traitées comme des formes fléchies d'un même lemme. Certaines peuvent exister comme entrée à part. Ne disposant pas d'un référentiel stable pour la classification de ces catégories fermées, et sachant que différents analyseurs les classent différemment, nous n'avons pas jugé prioritaire de proposer des sous-classes à ces mots fonctionnels. A terme, il serait souhaitable que cette classe fasse l'objet d'une sous-catégorisation consensuelle et compatible avec les catégories de données pour les partie du discours en cours d'élaboration à l'ISO TC 37/SC 4. Sur la base de cette sous-classification, il sera également possible d'envisager un traitement plus approprié de la variation (paradigmatique) des pronoms et déterminants (en tant que variation flexionnelle par un élément /inflectedForm/) et de la variation (non paradigmatique) allomorphique et morpho-phonétique (en tant que variante par un renvoi /spellingVariantOf/).

3.4 Limites de couverture et problèmes connus

Bien que *Morphalou 2.0* ait bénéficié d'un enrichissement considérable par rapport à *Morphalou 1.0*, certains phénomènes ne sont pas couverts de façon systématique, et/ou leur traitement n'est pas satisfaisant d'un point de vue linguistique. Ces problèmes sont dus à l'incohérence lexicographique inhérente à tout dictionnaire éditorial ainsi qu'aux difficultés d'extraction automatique qui s'ensuivent. Les erreurs connus feront l'objet de corrections manuelles, et les versions corrigées seront distribuées librement.

3.4.1 Noms propres

Le lexique ne comprend, en principe, pas de noms propres ("Paris", "Matignon"). Toutefois, certains noms propres, s'ils figuraient dans le TLFi, s'y trouvent, mais catégorisés en tant que nom commun ("Saint-Marcellin") et sans majuscule. C'est en particulier le cas des noms propres utilisés pour désigner des espèces (fromages, vins etc.). De la même façon, les gentilés et noms de peuples y figurent, le plus souvent comme noms et adjectifs, mais sans exhaustivité ("brésilien", "parisien" etc.).

3.4.2 Abréviations et troncations

Le lexique contient certaines abréviations ("C.R.S."), mais sans exhaustivité aucune, ni traitement satisfaisant : elles ont été transformées en minuscules, ne contiennent pas de lien vers leurs formes pleines et sont catégorisées en tant que noms communs (ou une autre catégorie), mais elles ne sont pas marquées en tant qu'abréviations. Par ailleurs, la présence des points dans le lemme n'est pas un indicateur sûr, il peut y avoir d'autres entrées (erronées) contenant des points, et certaines abréviations peuvent ne pas comporter des points ("etc"). Enfin, certaines formes tronquées figurent dans le lexique, mais là aussi, elles ne sont pas marquées en tant que telles ("sana", "parano").

3.4.3 Eléments lexicaux non autonomes

Le TLF contient des entrées pour des éléments formants ("gyn(o)-"), des préfixes ("a-") et des suffixes ("-able). En raison de l'inconsistence du traitement de ces éléments et des questions linguistiques liées à leur classification, nous ne les avons pas encore intégrés dans *Morphalou 2.0*.

3.4.4 Lemmes contenant des tirets

Le TLF contient des entrées, dont la forme graphique inclut un tiret facultatif. Dans la version papier, ces entrées se présentent sous une forme parenthésée, comme "auto(-)féconder" ou "saute(-)en(-)barque". L'ambiguïté d'expansion des parenthèses - orthographe sans ou avec espace possible - nous a amenés à retenir seule la version avec tiret (mais sans parenthèses). Pour indiquer la possibilité d'autres orthographes, la forme orthographique calculée a été assortie d'un marqueur spécial ("provisionallyProcessed")

3.4.5 Lemmes contenant des signes de ponctuation

En raison de l'extraction automatique de lemmes du TLF, certaines formes risquent d'être dégradées ou fausses. Cela concerne entre autres les lemmes contenant des signes de ponctuation (des points, en particulier) pour lesquels le tri et la correction doivent s'opérer majoritairement à la main ("polyodontie. 1", "subst. fém.", "mi-ciel ." etc.).

3.4.6 Lexies complexes : composés et locutions

L'ordre des composants d'une locution est, pour l'instant, l'ordre lexicographique tel qu'adopté dans le TLF. Cet ordre ne correspond pas toujours à l'ordre linéaire des occurrences dans un texte ("je-m'en-fiche à la", "tourne-main en un", "affilée d'"). Par ailleurs, il peut y avoir des problèmes de ségmentation et /ou de perte de composants encodés dans TLF de façon ambiguë ("s'il te/vous plaît", "tantet (un)", "tapinois (en)", "tire-d'aile (à)", "vade retro satanas ou satana"). Les composés discontinus ne sont pas traités correctement ("ne ... miette"). La résolution de ce problème demanderait une validation manuelle de toutes les entrées dont le code source du TLF contient la mention *locution/loc.* et la supression de doublons éventuels ("visé" et "visé au").

3.4.7 Entrées fantômes

Malgré un certain nombre de filtrages, il est possible qu'il reste des entrées fantômes, créées au cours de l'analyse automatique des structures idiosyncrasiques du TLF. Les problèmes connus sont

- les suffixes flexionnels adjectivaux ("uque")
- des doublons pour des locutions, sans ou avec une préposition ("tue-tête", "tue-tête à")
- des polycatégorisations abusives, souvent en /functionWord/ et une autre catégorie ("vis-à-vis")
- des formes féminines d'un adjectif, non reconnues comme forme fléchié ("vaine")
- des formes mal segmentées ("avenantà l'", "s'il te", "refile aller au", cf. aussi ci-dessus pour les locutions et les entrées contenant des ponctuations)
- des verbes simples issus d'une forme pronominale ("méfier", "absenter")

Nous "nettoyons" régulièrement le lexique pour diminuer ce genre de problèmes. Des contributions extérieures sont également bienvenues.

3.4.8 Mauvais renvois à des variantes orthographiques

Le calcul des variantes orthographiques a également souffert des encodages hétérogènes dans la source. En particulier, pour des lexies complexes présentant des éléments optionnels et/ou des variantes d'écriture dans le TLF, il est possible que ce calcul soit faux : "vanvale à la" renvoie à "venvole", "visé à" renvoie à "visé" etc. Par ailleurs, il y a des variantes orthographiques masculines pour des substantifs féminins dont les masculins possèdent des variantes ("abdominienne" est considéré comme variante de "abdominal", est "surfeuse" est variante de "surfer"). Enfin, les liens peuvent être incomplets, puisque certaines sous-vedettes du TLF indiquent des variantes

orthographiques plûtot que de nouveaux lemmes ("pappermane", "papé") sans que ce lien soit explicite dans le TLF. Il reste aussi des cas assez particuliers (pour rire un bon coup, allez voir "vous plaît"...).

3.4.9 Le choix entre participe et adjectif

Le TLF possède un certain nombre d'entrées catégorisés comme participes ("accouchant", "sourdant", "symphonisant"). N'ayant pas considéré les participes comme une catégorie grammaticale, nous avons dû trancher entre deux solutions : supprimer ces entrées ou les sauvegarder sous une autre catégorie, e.g. comme adjectif. Nous avons opté pour la deuxième solution, en tenant compte des paramètres suivants : si l'entrée était aussi marquée comme adjectif, nous l'avons gardée comme adjectif ("sourdant"). Si l'entrée était marquée uniquement comme participe, nous l'avons supprimée lorsque la forme existait déjà en tant que forme fléchie du verbe correspondant ("accouchant"). Sinon, elle a été gardée comme entrée adjectivale ("symphonisant"). Une conséquence potentiellement dérangeante de ce choix consiste en la perte des propriétés flexionnelles des participes présents considérés comme non adjectivaux (on prévoit "accouchant", mais pas "accouchante"...).

3.4.10 Lemme contenant des parenthèses

Certaines entrées du TLF contiennent des parenthèses pour indiquer des variantes orthographiques. Quand il y en a plusieurs dans un même lemme, comme par exemple dans "se a(c)couf(f)ler" ou "sans(-)desse(i)n(e)", la probabilité de génération d'erreurs augmente évidemment. Ces formes, ainsi que les liens entre variantes, doivent être revues et validées à la main.

3.4.11 Absence de formes fléchies

La majorité des nouvelles entrées extraites du TLF pour *Morphalou 2.0* ne contient pas les formes fléchies. D'ailleurs, nous sommes à la recherche d'un flechisseur "open source" pour le français: des contributions sont les bienvenues !

3.4.12 Absence d'entrées

Certaines entrées extraites du TLF sont encore en cours de vérification. Cela concerne en particulier toutes celles pour lesquelles nous n'avons pas réussi à calculer automatiquement une catégorie grammaticale. Ces quelques 7000 entrées seront intégrées dans la version suivante. Par ailleurs, nous avons mis en place un outil de veille lexicographique qui proposera régulièrement des candidats à la néologie formelle, catégorielle ou sémantique.

4 - Format de représentation : Lexical Markup Framework (LMF)

The Lexical Markup Framework is a currently elaborated ISO standard for encoding lexical resources (ISO Committee Draft 24613:2006). As a basic principle, it proposes to assembly components (such as /lexicon/, /lexicalEntry/, /form/, or /sense/) with data categories (such as /grammaticalCategory/, /grammaicalGender/, /definition/ etc.). Components might be seen as containers, whereas data categories might be thought of as terminal nodes of a lexicographical description. Data categories are either user defined or (preferably) imported from the Data Category Registry (DCR), which is also an ISO initiative and aims at providing a uniform and widely approved description of basic linguistic concepts.

Morphalou 2.0 implements the LMF proposal in the following way (see Figure 1):

- A /lexicon/ component contains /lexiconInformation/ and one or many /lexicalEntries/;
- /lexiconInformation/ contains metadata about the lexicon, in particular information about /originatingData/, that is source databases and responsible institutions;
- A /lexicalEntry/ comes with relations or data categories (pointers to spelling variants, to related masculine forms for feminine entries, and to an entry in a source database), and subordinate components;
- Subordinate components for a /lexicalEntry/ are /formSet/, as well as /sense/. The latter is not (yet) needed in Morphalou;
- /formSet/ (not in LMF !) is a container we use for grouping together /inflectedForm/ components and /lemmatizedForm/ components;
- a /lemmatizedForm/ has specific data categories (/grammaticalGender/, /inflectionalParadigm/, /grammaticalCategory/, /frequency/ and so on) and components which implements the LMF /representationFrame/;
- /pronunciation/ implements the LMF /representationFrame/ for phonetic and phonological information. It's refined by /script/, /transcription/, /syllabification/ and /liaison/;
- /orthography/ implements the LMF /representationFrame/ for information about written forms. It's refined by /script/ and /syllabification/.
- /inflectedForm/ also uses /pronunciation/ and /orthography/, additionally to specific data categories (/grammaticalGender/, /grammaticalMood/, /grammaticalTense/, /grammaticalPerson/, /grammaticalNumber/, /frequency/).

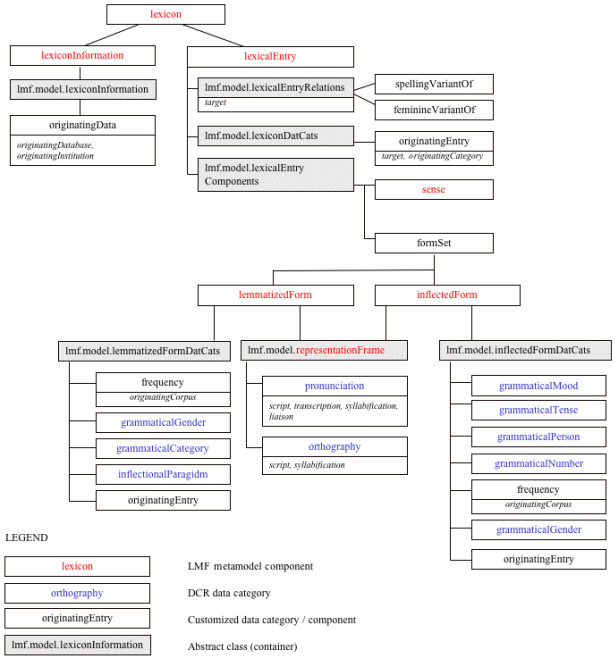

Figure 1. LMF implementation for Morphalou 2.0

5 - Technical Documentation : LMF implementation for Morphalou 2.0

Classes defined

lmf.model.inflectedFormDatcats

lmf.model.inflectedFormDatcats	Set of inflection features, applicable to inflected forms
Members	frequency grammaticalGender grammaticalMood grammaticalNumber grammaticalPerson grammaticalTense originatingEntry
Module	module-from-LMF-Morphalou

lmf.model.lemmatizedFormDatcats

lmf.model.lemmatizedFormDatcats	Set of data categories used for description of lemmatized forms (POS, Gender etc.).
Members	frequency grammaticalCategory grammaticalGender inflectionalParadigm originatingEntry
Module	module-from-LMF-Morphalou

lmf.model.lexicalEntryComponents

lmf.model.lexicalEntryComponents	This is a simple place holder, allowing for customization of metamodel components anchored on lexical entries.
Members	formSet sense
Module	module-from-LMF-Morphalou

lmf.model.lexicalEntryDatcats

lmf.model.lexicalEntryDatcats	This is a simple place holder, allowing for customization of data categories related to lexical entries.
Members	originatingEntry
Module	module-from-LMF-Morphalou

lmf.model.lexicalEntryRelations

lmf.model.lexicalEntryRelations	This is a simple place holder, allowing for customization of relations between lexical entries.
Members	feminineVariantOf spellingVariantOf
Module	module-from-LMF-Morphalou

lmf.model.lexiconInformation

lmf.model.lexiconInformation	Placeholder class for customization of metadata related information.
Members	originatingData
Module	module-from-LMF-Morphalou

lmf.model.representationFrame

lmf.model.representationFrame	Set of elements used for description of writing and pronunciation properties of a form
Note	INcludes information about script, orthography etc.
Members	orthography pronunciation
Module	module-from-LMF-Morphalou

Elements defined

<feminineVariantOf>

feminineVariantOf	A pointer from a separate entry for a feminine noun to the entry for the corresponding masculine noun.
Class	lmf.model.lexicalEntryRelations
Declaration	`element feminineVariantOf { attribute target { xsd:IDREF }?, text }`
Attributes	(In addition to global attributes) target Status: Datatype: `xsd:IDREF`
Module	module-from-LMF-Morphalou

<formSet>

formSet	A /formSet/ -- not in LMF -- rassembles a unique lemma and all its inflected forms, that is one inflection paradigm.
Class	lmf.model.lexicalEntryComponents
Declaration	`element formSet { lemmatizedForm, inflectedForm* }`
Attributes	(Global attributes only)
Note	This is a customized container which differs slightly from LMF where the grouping of lemmata and inflected forms is suggested to be done by embedding inflected forms within the lemmatized form. We consider that embedding inflected forms into a lemmatized form does not well reflect the linguistic relationship between lemmas and inflected forms, that is abstraction instead of composition. Furthermore, embedding seems also odd with regard to the conceptual inheritance of both inflected and lemmatized forms from an underlying /form/ class, as suggested in LMF.
Module	module-from-LMF-Morphalou

<frequency>

frequency	normalized frequency of a word form or a lemma, in a given reference corpus
Class	lmf.model.lemmatizedFormDatcats lmf.model.inflectedFormDatcats
Declaration	`element frequency { attribute originatingCorpus { text }?, text }`
Attributes	(In addition to global attributes) originatingCorpus Status: Datatype: `text`
Module	module-from-LMF-Morphalou

<grammaticalCategory>

grammaticalCategory	
Class	lmf.model.lemmatizedFormDatcats
frequency	normalized frequency of a word form or a lemma, in a given reference corpus

Class	lmf.model.lemmatizedFormDatcats lmf.model.inflectedFormDatcats
Declaration	element frequency { attribute originatingCorpus { text }?, text }
Attributes	(In addition to global attributes) originatingCorpus Status: Datatype: text
Module	module-from-LMF-Morphalou

<grammaticalNumber>

grammaticalNumber	
Class	lmf.model.inflectedFormDatcats
Declaration	element grammaticalNumber { text }
Attributes	(Global attributes only)
Module	module-from-LMF-Morphalou

<grammaticalPerson>

grammaticalPerson	
Class	lmf.model.inflectedFormDatcats
Declaration	element grammaticalPerson { text }
Attributes	(Global attributes only)
Module	module-from-LMF-Morphalou

<grammaticalTense>

grammaticalTense	
Class	lmf.model.inflectedFormDatcats
Declaration	element grammaticalTense { text }
Attributes	(Global attributes only)
Module	module-from-LMF-Morphalou

<inflectedForm>

inflectedForm	A wordform as to be used in context, and to be observed in corpora. An /inflectedForm/ is characterized by a set of data categories for inflectional features, and components related to its graphical and phonetical representation.

Declaration	element inflectedForm
	{
	attribute id { xsd:ID }?,
	lmf.model.representationFrame*,
	lmf.model.inflectedFormDatcats*
	}

Attributes	(In addition to global attributes) id Status: Datatype: xsd:ID
Note	In our application, any lemmatized form has a counterpart as an inflected form.
Module	module-from-LMF-Morphalou

<inflectionalParadigm>

inflectionalParadigm	reference to an inflection class for nouns, adjectives and verbs
Class	lmf.model.lemmatizedFormDatcats
Declaration	element inflectionalParadigm { text }
Attributes	(Global attributes only)
Module	module-from-LMF-Morphalou

<lemmatizedForm>

lemmatizedForm	This element implements the /lemmatizedForm/ component of the LMF metamodel. It is refined by customzed data categories as well as members of the LMF /representationFrame/ class.

Declaration	element lemmatizedForm
	{
	attribute id { xsd:ID }?,
	lmf.model.representationFrame*,
	lmf.model.lemmatizedFormDatcats*
	}

Attributes	(In addition to global attributes) id Status: Datatype: xsd:ID
Note	In our application, a lemma is a conventionnally chosen written word form - in the default script and orthography of the lexicon - that represents an abstraction over a homogeneous set of inflected forms. As a consequence, formal (e.g. spelling or abbreviation) variants should lead to different lemmata, and hence to different lexical entries.
Module	module-from-LMF-Morphalou

<lexicalEntry>

lexicalEntry	This element implements the /lexicalEntry/ component of the LMF core metamodel.

Declaration	element lexicalEntry
	{
	attribute id { xsd:ID }?,
	(
	lmf.model.lexicalEntryRelations*,
	lmf.model.lexicalEntryComponents*,
	lmf.model.lexicalEntryDatcats*
	)
	}

Attributes	(In addition to global

	attributes) id Status: Datatype: `xsd:ID`
Note	A /lexicalEntry/ is associated with a set of (user-customized) data categories, relations to other entries and a set of core or extending components from the LMF metamodel (form, sense etc). In our application, a /lexicalEntry/ is extended with one /formSet/ component, grouping together one lemmatized and zero to more inflected forms. Furthermore, there might be pointers to entries with spelling or feminisation variants. The model also integrates a sense slot, not yet used in Morphalou.
Module	module-from-LMF-Morphalou

<lexicon>

lexicon	Implements the LMF /lexicon/ component. Corresponds to the root of the lexical database.
Declaration	`element lexicon { lexiconInformation, lexicalEntry* }`
Attributes	(Global attributes only)
Module	module-from-LMF-Morphalou

<lexiconInformation>

lexiconInformation	Implements the /lexiconInformation/ of LMF. Includes metadata on contributing source databases and institutions.
Declaration	`element lexiconInformation { lmf.model.lexiconInformation* }`
Attributes	(Global attributes only)
Module	module-from-LMF-Morphalou

<originatingData>

originatingData	An element for identification of a source database and the responsible institution.
Class	lmf.model.lexiconInformation
Declaration	`element originatingData { attribute originatingDatabase { text }?, attribute originatingInstitution { text }?, empty }`
Attributes	(In addition to global attributes) originatingDatabase Datatype: `text`
Module	module-from-LMF-Morphalou

<orthography>

orthography	A shortened LMF representation frame for standard orthography (e.g. Academic French) in a default script (e.g. Latin). Possibility of specifying syllabification.
Class	lmf.model.representationFrame
Declaration	`element orthography { attribute processStatus { text }?, attribute script { text }?, attribute syllabification { text }?, text }`
Attributes	(In addition to global attributes) processStatus Status: Datatype: `text` script Status: Datatype: `text` syllabification Status: Datatype: `text`
Module	module-from-LMF-Morphalou

<pronunciation>

pronunciation	A particular LMF representation frame for phonetic or phonological transcription in a specified script and transcription system. Possibility of specifying syllabification and liaison.
Class	lmf.model.representationFrame
Declaration	`element pronunciation { attribute script { text }?, attribute transcription { text }?, attribute syllabification { text }?, attribute liaison { text }?, text }`
Attributes	(In addition to global attributes) script Status: Datatype: `text` transcription Status: Datatype: `text` syllabification Status: Datatype: `text` liaison Status: Datatype: `text`

Module	module-from-LMF-Morphalou

<sense>

sense	Implements the lmf /sense/ component
Class	lmf.model.lexicalEntryComponents
Attributes	(Global attributes only)
Note	Has to be elobarated. Not used so far in Morphalou.
Module	module-from-LMF-Morphalou

<spellingVariantOf>

spellingVariantOf	A pointer from a separate entry for a spelling variant to a main entry
Class	lmf.model.lexicalEntryRelations
Declaration	`element spellingVariantOf { attribute target { xsd:IDREF }?, text }`
Attributes	(In addition to global attributes) target Status: Datatype: `xsd:IDREF`
Note	The choice of the "main entry" for spelling variants has been done automatically: we chose the alphabetically first lemmatized form.
Module	module-from-LMF-Morphalou

6 - References

The Lexical Markup Framework is a currently elaborated ISO standard for encoding lexical resources (ISO Committee Draft 24613:2006). As a basic principle, it proposes to assembly components (such as /lexicon/, /lexicalEntry/, /form/, or /sense/) with data categories (such as /grammaticalCategory/, /grammaicalGender/, /definition/ etc.). Components might be seen as containers, whereas data categories might be thought of as terminal nodes of a lexicographical description. Data categories are either user defined or (preferably) imported from the Data Category Registry (DCR), which is also an ISO initiative and aims at providing a uniform and widely approved description of basic linguistic concepts.

- Romary L., Salmon-Alt S., Francopoulo G. (2004). Standards going concrete : from LMF to Morphalou. Workshop on Electronic Dictionaries, Coling 2004, Geneva, Switzerland.
- Salmon-Alt S., Akrout A., Romary L. (2005). Proposals for a normalized representation of Standard Arabic full form lexica. Second International Conference on Machine Intelligence (ACIDCA-ICMI 2005), Tozeur, Tunisia.
- Polguère A. (2003) Lexicologie et sémantique lexicale. Notions fondamentales, coll. "Paramètres", Montréal: Presses de l'Université de Montréal
- Francopoulo G., Monte G. (2006). Lexical Markup Framework (LMF aka ISO-24613), CD revision 9 : 15 mars 2006
- Ide N., Romary L. (2004). A Registry of Standard Data Categories for Linguistic Annotation, 4th International Conference on Language Resources and Evaluation - LREC'04, May 2004.
- Burnard L., Rahtz S. (2004). RelaxNG with Son of ODD. Proceedings of Extreme Markup Languages 2004®.