Karolina Suchecka Nathalie Gasiglia
Resp. ULR ALITHILA et UMR STL (Université de Lille)
27 juin 2022

TAL et Littérature comparée.
Détection automatique des correspondances textuelles entre les réécritures d'un mythe

Atelier TAL et Humanités numériques

Le corpus des réécritures : 53 textes

- 2 principales sources antiques :
  1. Virgile, *Les Géorgiques*, chant IV, 37-30 av. J.-C.
  2. Ovide, *Les Métamorphoses*, livres X et XI, 1er s. après. J.-C.
- 53 réécritures en français, en vers et en prose, de différents genres et états de la langue :
  - **Poésie :** L'Hermite, « La Lyre d'Orphée » (1662)
  - **Opéra :** Gluck, Calzabigi & Moline, *Orphée et Eurydice* (1774)
  - **Prose :** Ballanche, *Essais de palingénésie sociale. Orphée* (1827-1829)
  - **Théâtre :** Cocteau, *Orphée. Tragédie en un acte et un intervalle* (1927)
  - **Littérature de jeunesse :** Jimenes, *Orphée l'enchanteur* (2004)

Le corpus des traductions : 41 textes

« La **flexibilité** permet de suggérer la souplesse d'adaptation et en même temps la résistance de l'élément mythique dans le texte littéraire, les modulations surtout dont ce texte lui-même est fait. » (Brunel, *Mythocritique : théorie et parcours*, 1992)

1. Virgile, *Les Géorgiques*, chant IV, 37-30 av. J.-C.

   Publiées entre 1540 et 1932 :
   - 6 traductions en vers
   - 8 traductions en prose
   - 1 transcription d'un manuscrit ancien
2. Ovide, *Les Métamorphoses*, livres X et XI, I^{er} s. après. J.-C.

   Publiées entre 1493 et 2008 :
   - 7 traductions en vers
   - 13 traductions en prose
   - 4 transcriptions de manuscrits anciens

Logiciels de détection des réutilisations textuelles

Tracer, Marco Büchler (dir.), eTRAP, Georg-August-Universität Göttingen. URL : https ://www.etrap.eu/research/tracer/.

TextPAIR, Clovis Gladstone (dir.), ARTFL, Université de Chicago. URL : https ://artfl-project.uchicago.edu/text-pair/.

Techniques informatiques :

- Plongement lexical (Mikolov *et al.*, 2013) ;
- Alignement séquentiel (Wise, 1993 ; Bergroth *et al.*, 2000)
- *N-grammes* (Jurafsky & Martin, 2009 ; Forstall *et al.*, 2015)

Observation d'un échantillon de 1 000 résultats

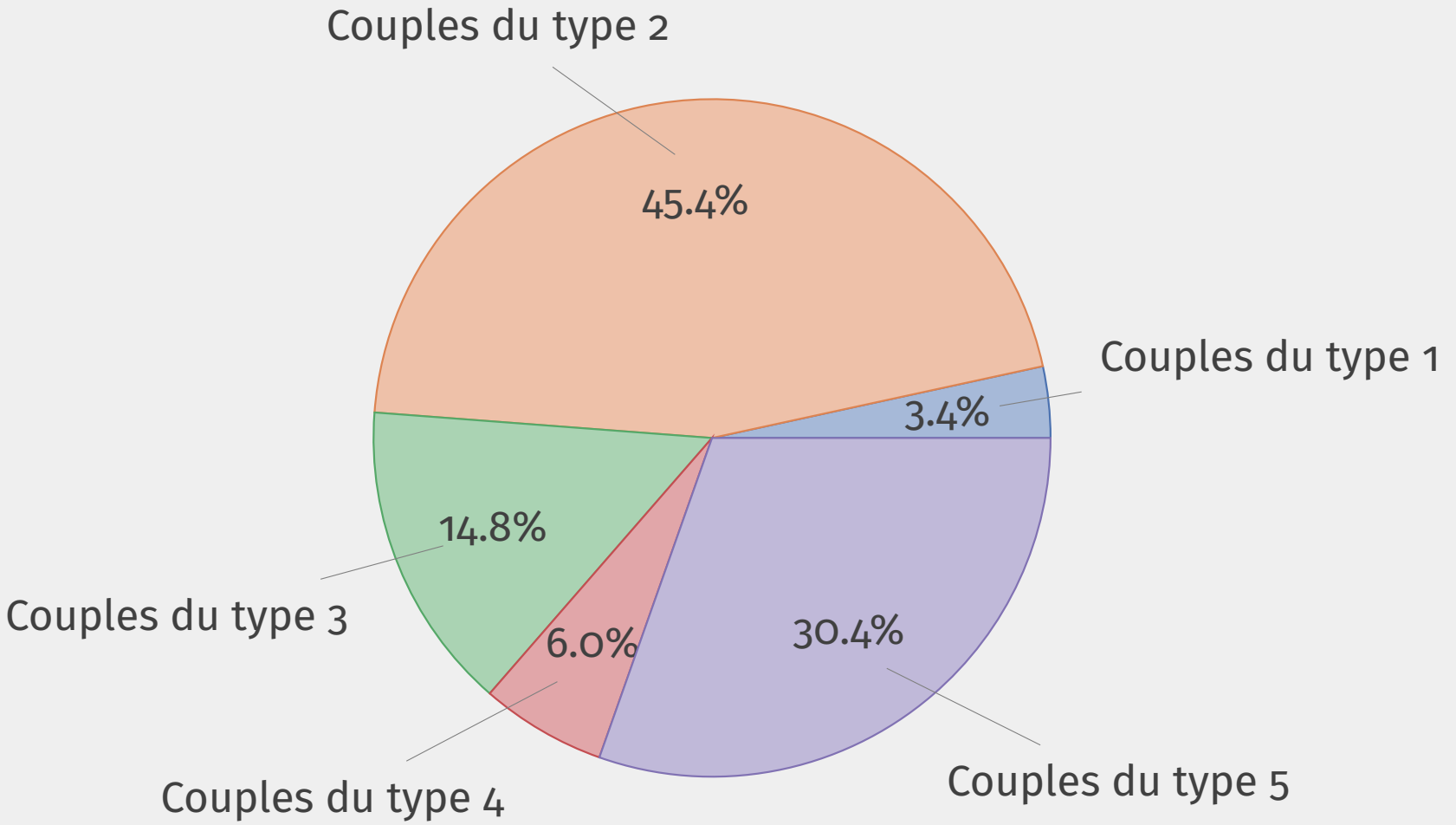

Figure – Observation des 1 000 couples choisis aléatoirement parmi les résultats de Tracer (500/10 414) et TextPAIR (500/8 8851)

Observation d'un échantillon de 1 000 résultats

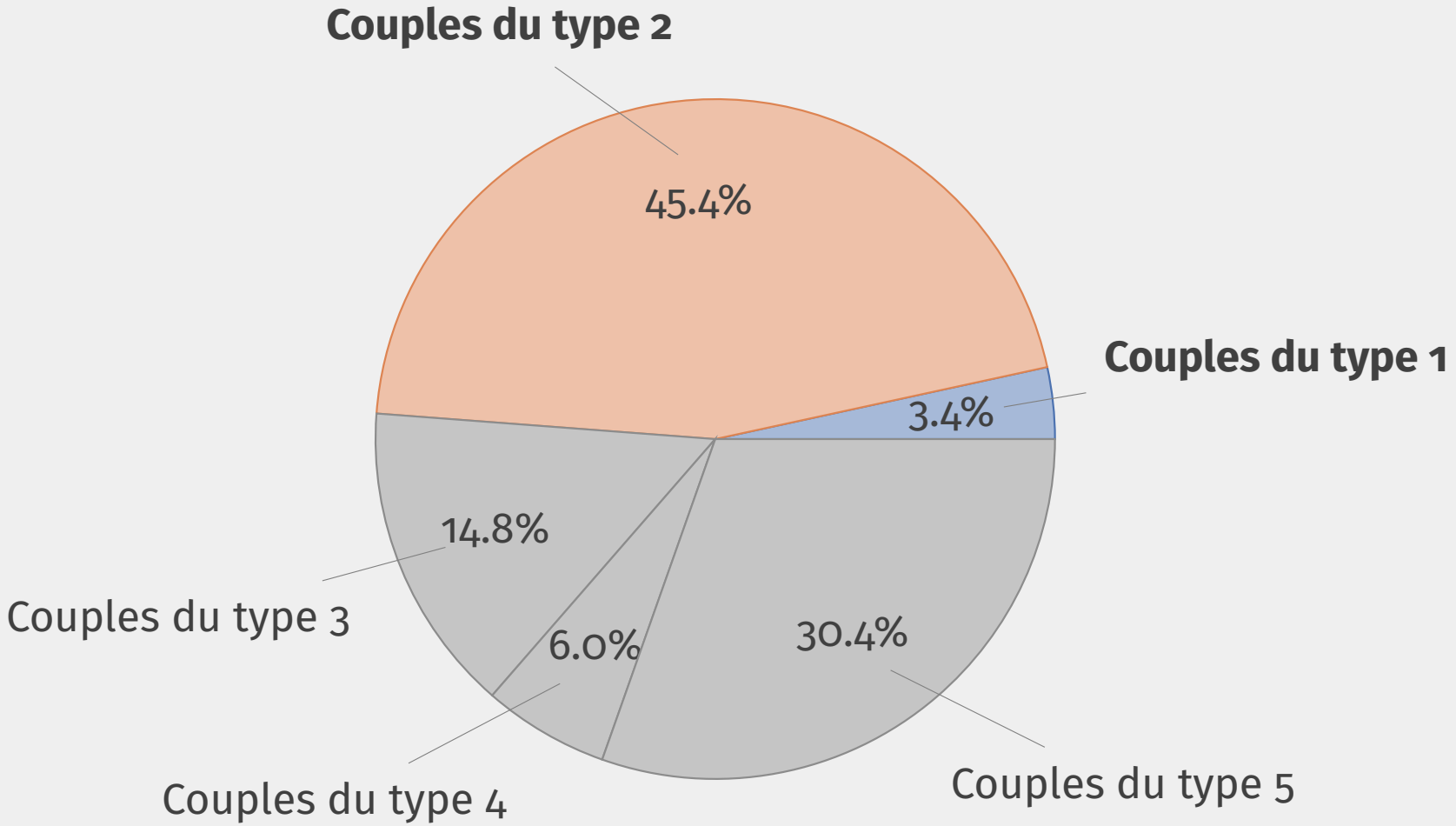

Figure – Type 1 (3,4%) : relations fausses issues du mauvais paramétrage des outils ou préparation inadaptée du corpus soumis. **Type 2** (45,4%) : relations fausses ou peu pertinentes appuyées sur des mots ou expressions faibles

Observation d'un échantillon de 1 000 résultats

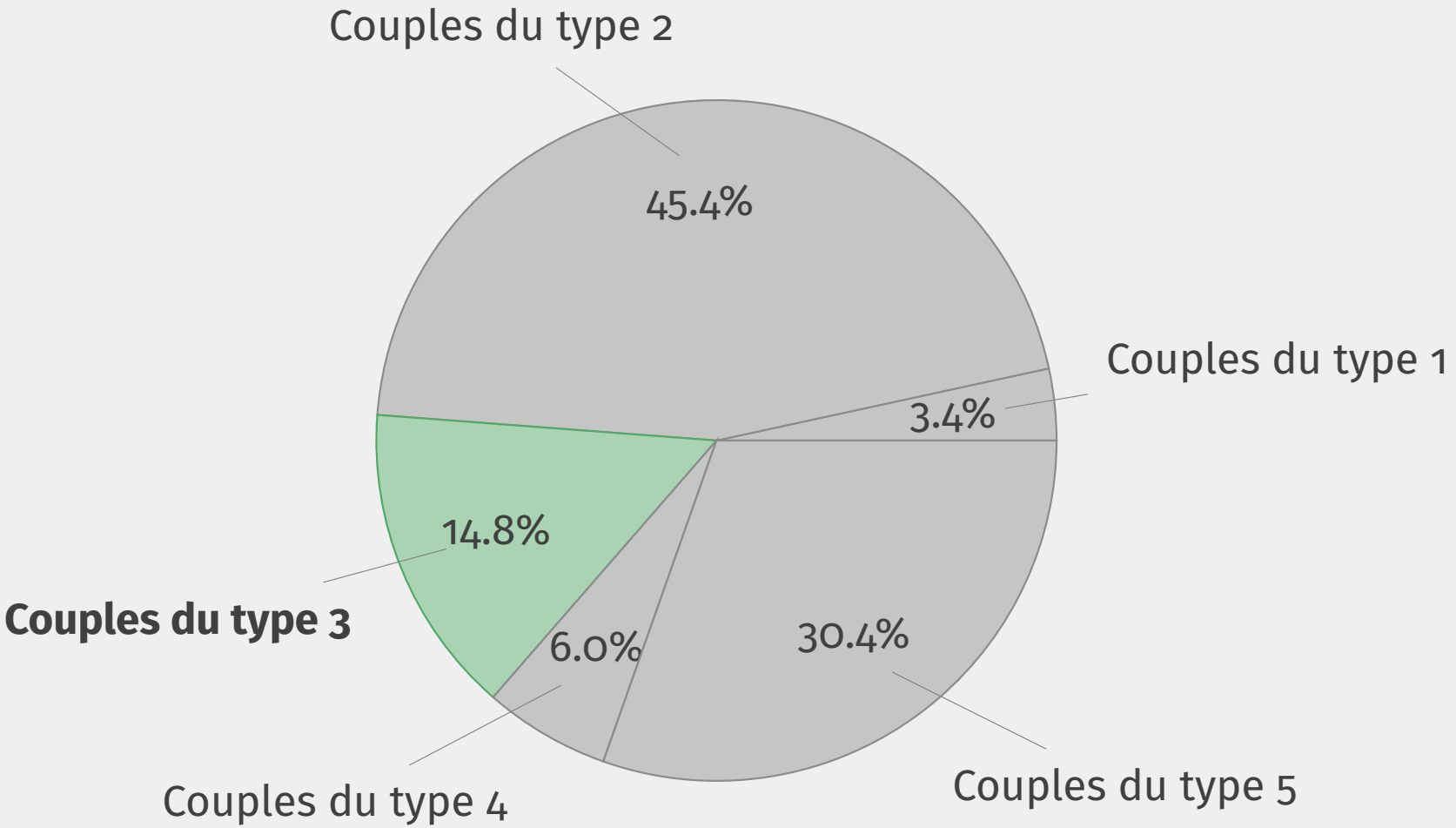

Figure – Type 3 (14,8%) : relations appuyées sur des mots forts, mais dont le contexte fourni est insuffisant pour évaluer leur pertinence

Observation d'un échantillon de 1 000 résultats

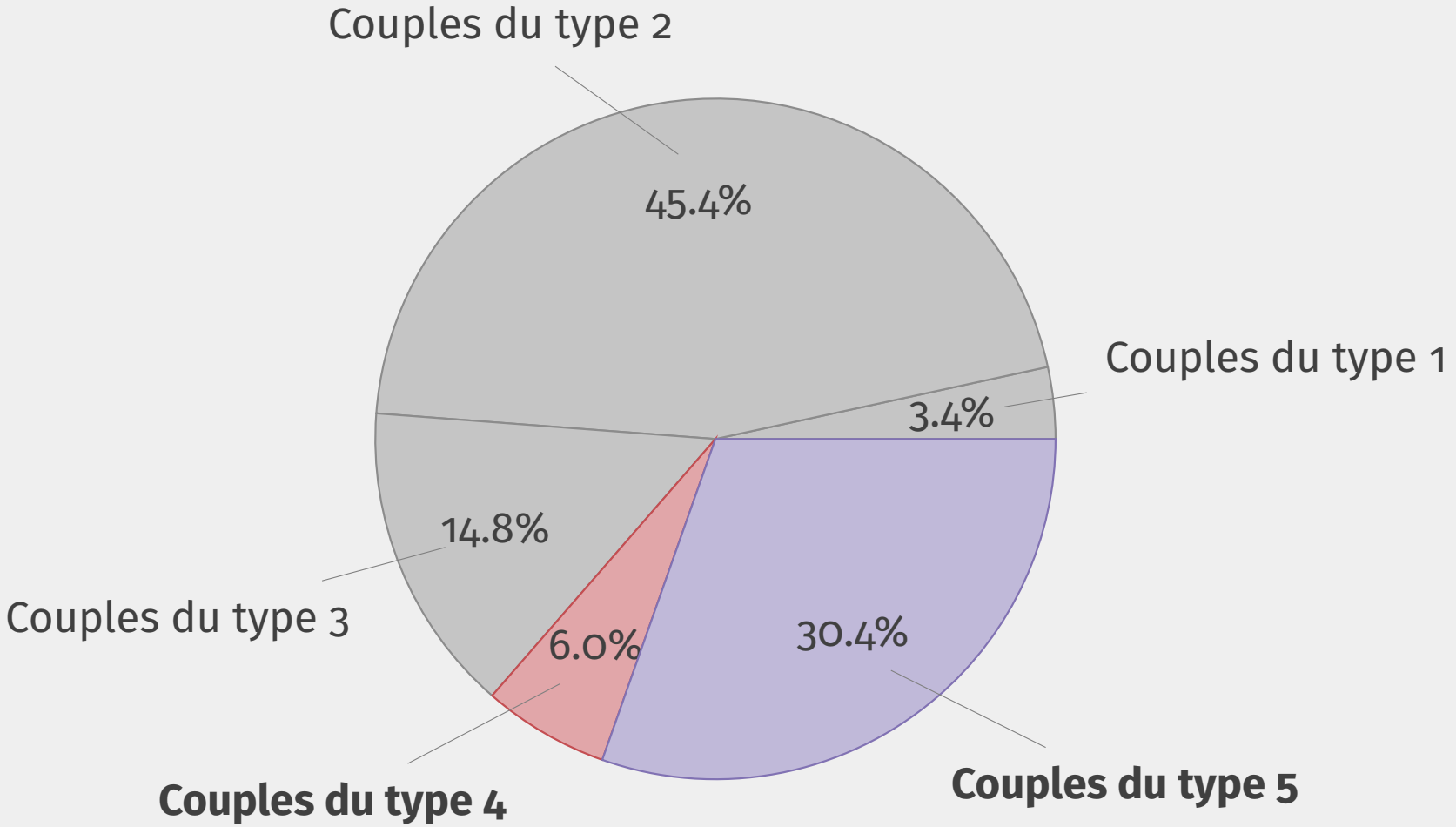

Figure – Type 4 (6%, dont 5,4% incluant des réécritures) : relations pertinentes, mais très allusives. **Type 5** (30,4%, dont 1,57% incluant des réécritures) : relations pertinentes et explicites

Quelques difficultés à surmonter

- Taux d'erreurs élevé
- Nombre d'extraits impliqués trop important pour examiner chaque couple

↪ pour dégager les liens les plus pertinents, nous exploitons ce qui les motive : identités de mots, de lemmes ou de racines, ou synonymes communs

Traitement des couples issus des 41 traductions

- Précision élevée : 848/860 couples de Tracer pertinents
- Post-traitement au moyen du langage XML

```
<phr type="GN" select="mari">
        <w xml:id="2400028_14" n="14" lemma="son" pos="DET:POS">son</w>
        <w xml:id="2400028_15" n="15" lemma="mari" pos="NOM"
        sameAs="époux conjoint homme">mari
                <xr corresp="6900027_17" type="forme" cert="5"/>
                <xr corresp="7400048_48" type="lemme" cert="5"/>
                <xr corresp="1300018_19" type="lemme_synonyme" cert="2"/>
                <xr corresp="1600020_13" type="synonyme_synonyme" cert="1"/>
        </w>
</phr>
```

Figure – Exemple de structuration d'un groupe nominal (`<phr>`). Chaque mot (`<w>`) est enrichi des informations linguistiques : lemme (`@lemma`), catégorie grammaticale (`@pos`) et synonymes (`@sameAs`). Chaque mot plein est lié à ceux d'autres couples (`<xr>`)

Traitement des couples issus des 41 traductions

- Précision élevée : 848/860 couples de Tracer pertinents
- Post-traitement au moyen du langage XML

```
<phr type="GN" select="mari">
        <w xml:id="2400028_14" n="14" lemma="son" pos="DET:POS">son</w>
        <w xml:id="2400028_15" n="15" lemma="mari" pos="NOM"
        sameAs="époux conjoint homme">mari
                <xr corresp="6900027_17" type="forme" cert="5"/>
                <xr corresp="7400048_48" type="lemme" cert="5"/>
                <xr corresp="1300018_19" type="lemme_synonyme" cert="2"/>
                <xr corresp="1600020_13" type="synonyme_synonyme" cert="1"/>
        </w>
</phr>
```

Figure – Exemple de structuration d'un groupe nominal (`<phr>`). Chaque mot (`<w>`) est enrichi des informations linguistiques : lemme (`@lemma`), catégorie grammaticale (`@pos`) et synonymes (`@sameAs`). Chaque mot plein est lié à ceux d'autres couples (`<xr>`)

Traitement des couples issus des 41 traductions

- Précision élevée : 848/860 couples de Tracer pertinents
- Post-traitement au moyen du langage XML

```
<phr type="GN" select="mari">
        <w xml:id="2400028_14" n="14" lemma="son" pos="DET:POS">son</w>
        <w xml:id="2400028_15" n="15" lemma="mari" pos="NOM"
        sameAs="époux conjoint homme">mari
                <xr corresp="6900027_17" type="forme" cert="5"/>
                <xr corresp="7400048_48" type="lemme" cert="5"/>
                <xr corresp="1300018_19" type="lemme_synonyme" cert="2"/>
                <xr corresp="1600020_13" type="synonyme_synonyme" cert="1"/>
        </w>
</phr>
```

Figure – Exemple de structuration d'un groupe nominal (`<phr>`). Chaque mot (`<w>`) est enrichi des informations linguistiques : lemme (`@lemma`), catégorie grammaticale (`@pos`) et synonymes (`@sameAs`). Chaque mot plein est lié à ceux d'autres couples (`<xr>`)

Graphes et délimitation des segments du mythe

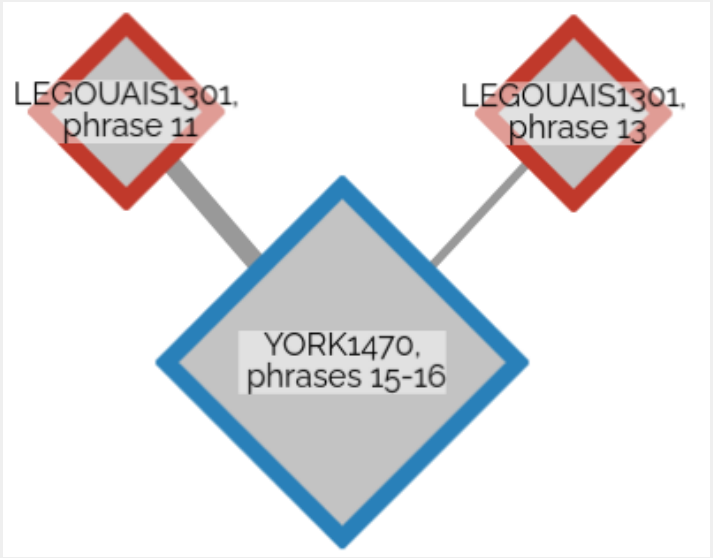

Figure – Principe de visualisation à l'aide de graphes

- Diamants = traductions d'Ovide
- Nœud = extrait textuel (phrase(s))
- Bordures rouges = textes en vers
- Bordures blues = textes en prose
- Épaisseur de liens = similarité de deux extraits

Relation entre les traductions de Virgile

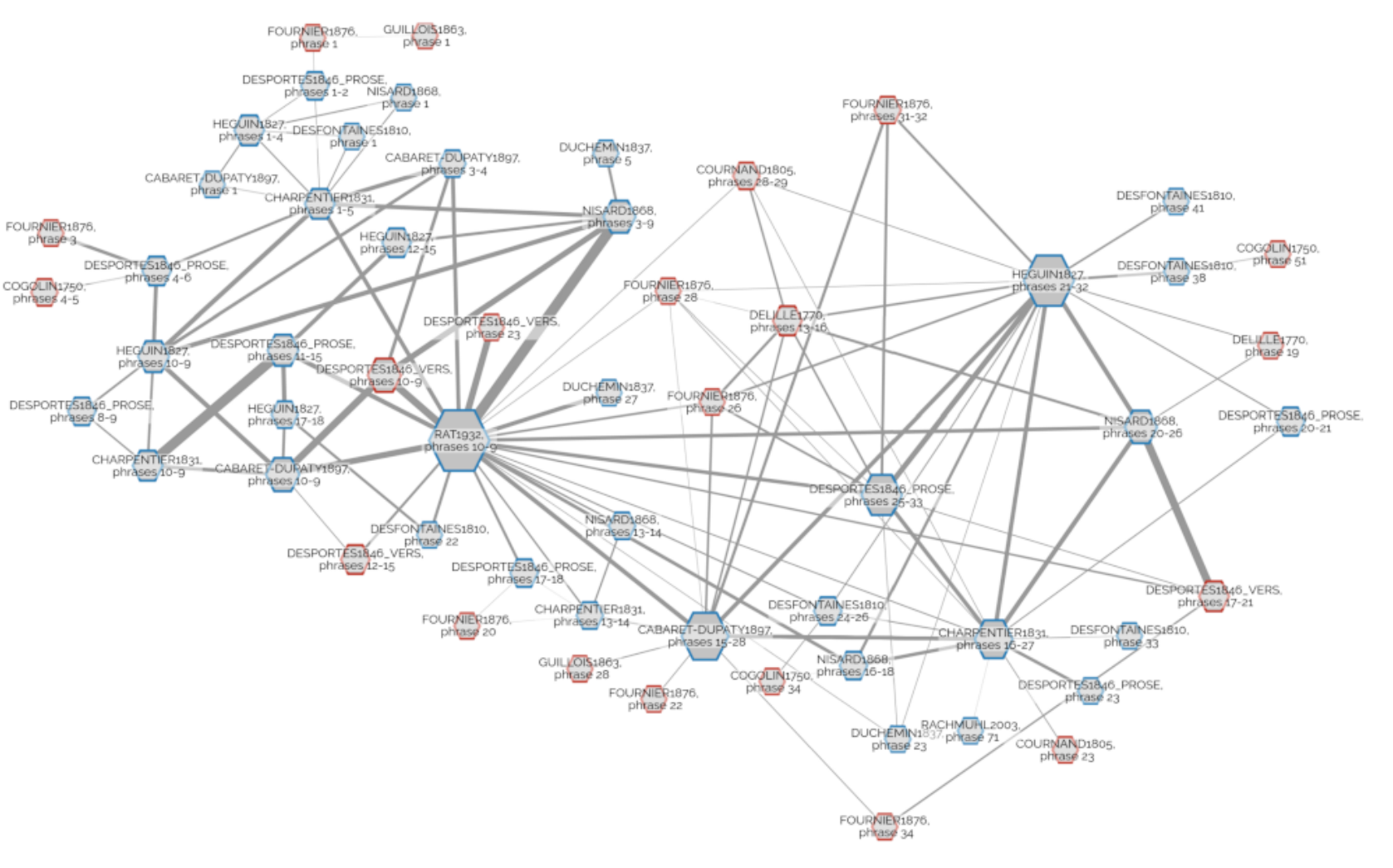

Figure – Graphe de la totalité des relations entre les traductions de Virgile (zoomable dans l'interface)

Relations entre les traductions d'Ovide

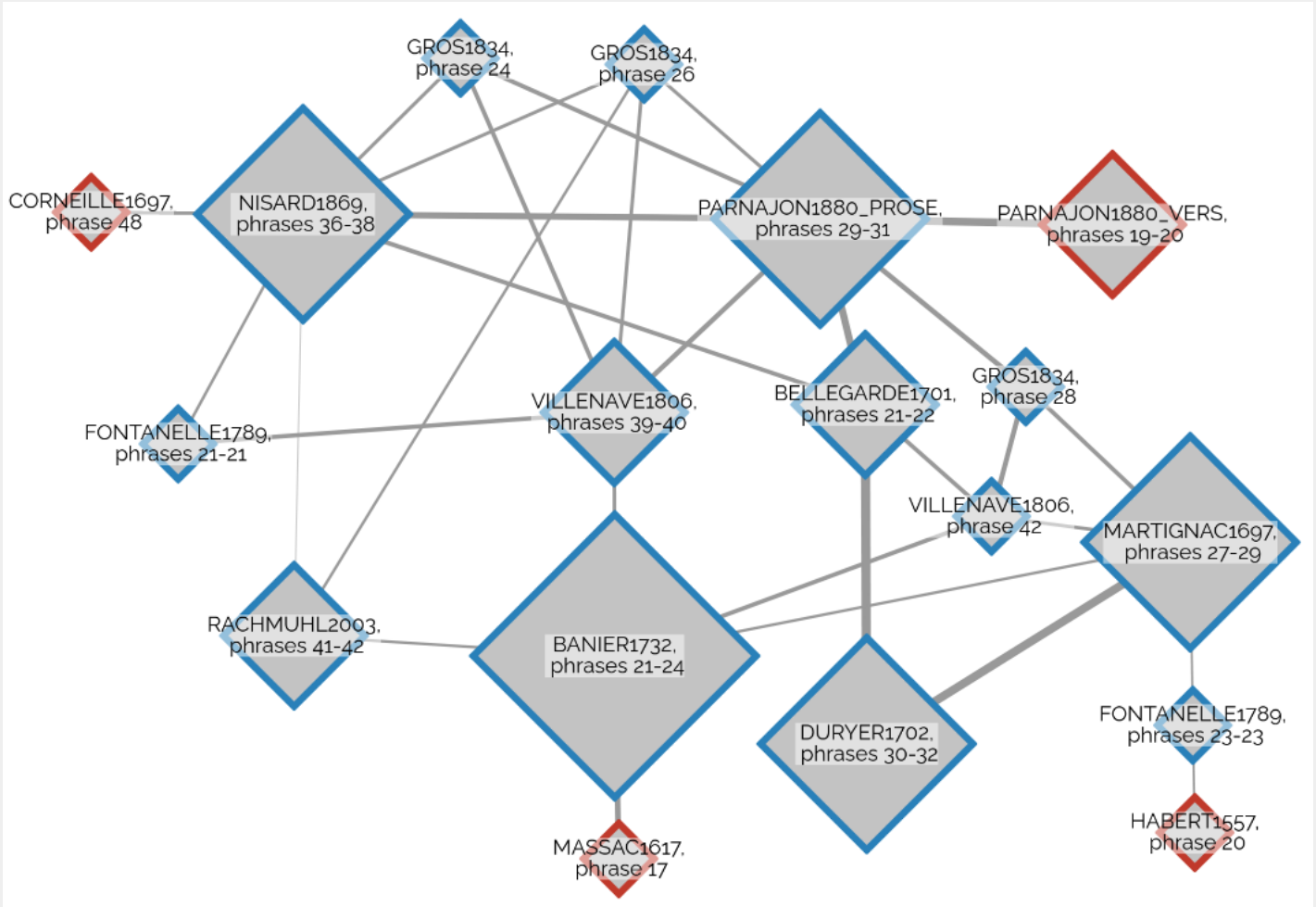

Figure – Graphe de 18 nœuds évoquant la seconde mort d'Eurydice

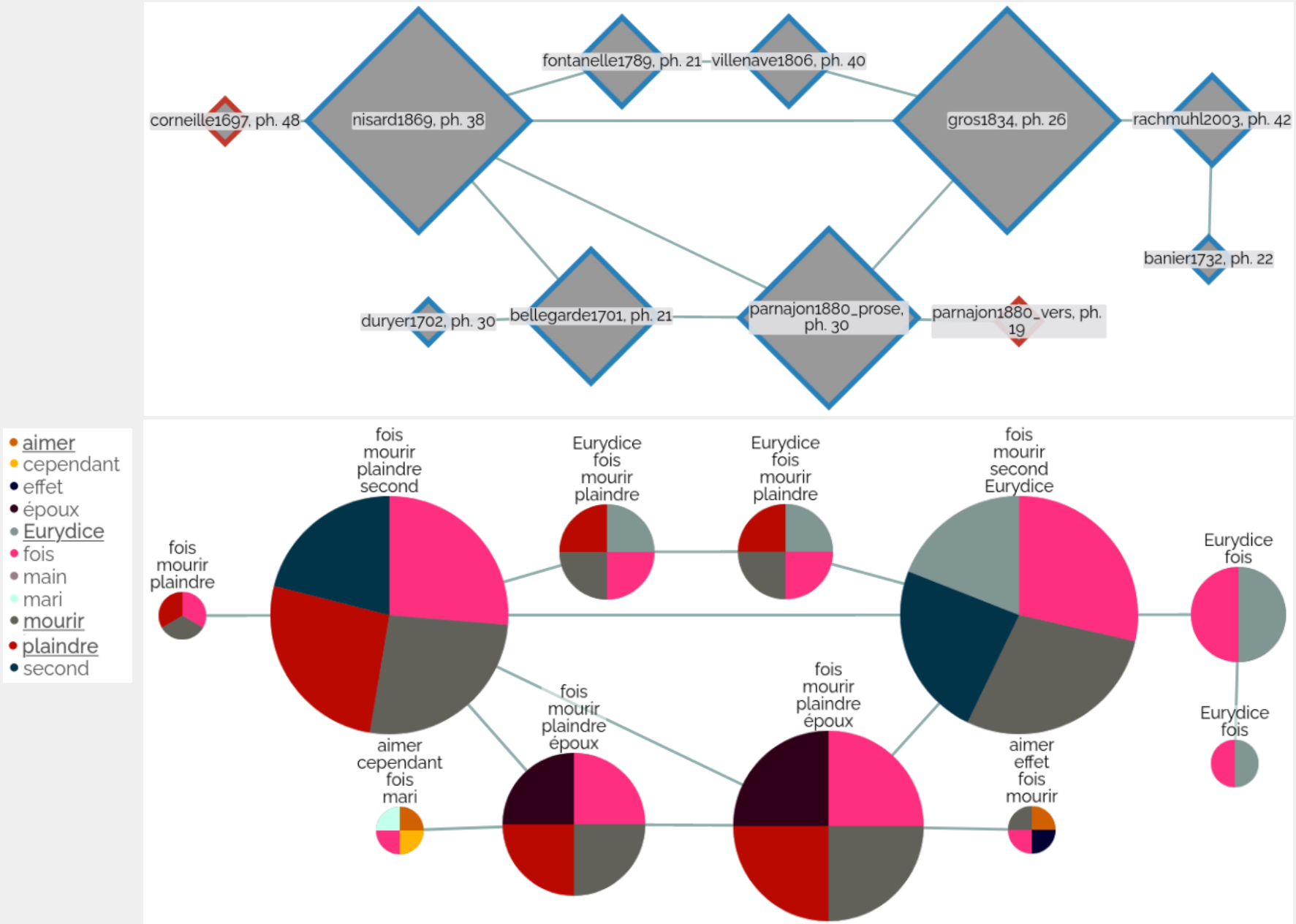

Figure – Visualisations complémentaires du sous-graphe de 11 nœuds (seconde mort d'Eurydice dans les traductions d'Ovide)

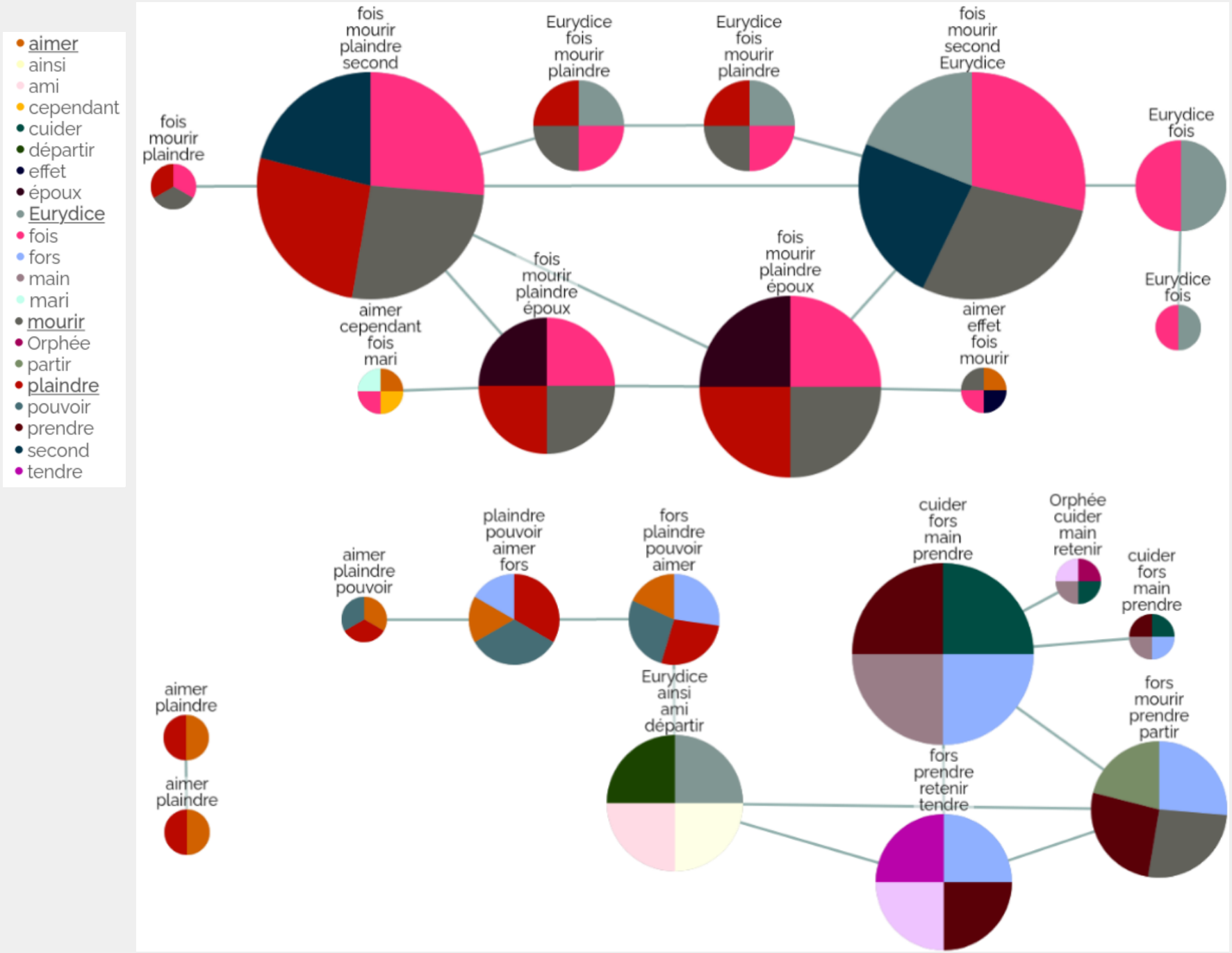

Figure – Rapprochement de trois graphes indépendants

Détail des phrases regroupées

- 445 relations lexicales pour 57 phrases de 18 textes
- 130 lemmes communs
- 30 lemmes appariés au moins 10 fois
  - 3 lemmes les plus fréquents (34-32 occ.) : *second*, *mourir*, *plaindre*
  - Les noms et périphrases de deux amants : Orphée (14 occ.), Eurydice (14 occ.), *ami* (14 occ.), *époux* (13 occ.)
  - Beaucoup de verbes : *entendre* (21 occ.), *tendre* (21 occ.), *retourner* (14 occ.), *aimer* (12 occ.), *embrasser* (12 occ.) et *regarder* (12 occ.)

Seconde mort d'Eurydice chez Gluck (1774)

Gluck (1774), acte III, scène I :
38 répliques, 110 phrases

vs.

- **Traductions de Virgile :**
  15 phrases maximum
- **Traductions d'Ovide :**
  8 phrases maximum

Seconde mort d'Eurydice chez Gluck (1774)

- 385 mots forts dont 175 mots-clés de l'épisode (80 de Virgile, 71 d'Ovide et 24 occ. des noms propres de deux amants)
- Les mots restants ont un sens interprétatif
  - Incompréhension et refus d'Eurydice : *abandonner*, *abhorrer*, *affliger*, *barbare*, *barbarie*, *ennemi*, *fuir*, *indifférence*, *ingrat*, *injustice*, *jaloux*, *outrager*, *soupçon*
  - Désarroi d'Orphée : *contrainte*, *désespérer*, *désoler*, *effroi*, *implorer*, *martyre*, *refuser*, *secret*, *trouble*

Synonymie

« sur son visage il détournât ses yeux » :

- *œil* : parmi les mots-clés de l'épisode
- *détourner* : équivalence synonymique avec *se retourner*

Mieux : *détourner les yeux sur* ∼{*regarder en arrière* ; *se retourner [et] poser les yeux sur* ; *tourner le regard vers*}

Résultats d'une expérimentation avec *word2vec*

`$enquête` : rang | lemme apparié | similarité

`$orphée` :

1.	aristée	0,95
3.	hermès	0,93
5.	eurydice	0,89

`$mourir` :

1.	consoler	0,93
2.	rassurer	0,92
4.	suivre	0,31

`$eurydice` :

1.	aristée	0,92
2.	érigone	0,91
5.	orphée	0,89

`$aimer` :

1.	obéir	0,89
4.	abandonner	0,89
5.	souffrir	0,89

Relations flexionnelles et dérivationnelles

- Stemmatisation avec l'algorithme Snowball (Porter, 2001)
  - $aimer_{\text{aim}}$ = $aimait_{\text{aim}}$ = $aimant_{\text{aim}}$
  - *fer* = $fermer_{\text{fer}}$
  - $savons_{\text{savon}}$ ≠ *savez* $_{\text{sav}}$ ≠ $savent_{\text{savent}}$
- Exploration des ressources morphologiques, comme DériF ou Morphonette
  - 2 007/11 653 de lemmes liés à leur base morphologique
  - *noceur* = *nocif*, *voyant* = *voyage*, *mauve* = *mauvais*, *vin* = *divin*
  - 13 appariements de dérivés des classes grammaticales différentes :
    admirable = *admiration*, *inconsolable* = *consolation*, *embrassant* = *embrassement*, *étonnant* = *étonnement*, *perdurable* = *perdurablement*, *successif* = *successivement*, *berceuse* = *bercer*, *charmant* = *charmer*, *déplorable* = *déplorer*, *entendement* = *entendre*, *flottant* = *flotter*, *touchant* = *toucher*, *sonnant* = *sonnette* = *sonneur*.

Correction de l'annotation

- Annotation morpho-syntaxique avec TreeTagger
- Correction manuelle des résultats
  - Modernisation de l'ancien et du moyen français (*aimoit*)
  - Regroupement des mots composés (*pomme de terre*)
  - Correction des erreurs de la tokenisation (*dit-on*, *serait-ce*)
- Soumission de la liste des formes et leurs lemmes à TextPAIR

	Résultats de l'annotation automatique	Entrées corrigées manuellement
Total de mots-occurrence	424 729	33 527 (7,89 %)
Formes (graphies différentes)	24 692	7 108 (28,79 %)
Lemmes	11 653	3 971 (34,08 %)

Figure – Résultats de l'annotation automatique et sa correction manuelle

Conclusion

- Observation et analyse d'un concept théorique issu de la recherche littéraire à l'aide des outils informatiques.
- Analyse des résultats grâce à l'enrichissement des résultats initiaux, l'affinement des détections, l'annotation des unités communes et la visualisation à l'aide de graphes.
- Motivation par des objectifs littéraires, mais manipulations qui sollicitent des compétences linguistiques-informatiques.

- Besoins :
  - chaînes de traitement cohérents et dont les produits sont compréhensibles et manipulables
  - dialogue et échange interdisciplinaire (des connaissances, des pratiques, des données)

↪ Pour produire des données mieux exploitables et mieux réutilisables par des recherches linguistiques-informatiques.