IBM SolutionsConnect 2013

L'IBM TechSoftware nouvelle génération

28, 29 et 30 août

IBM Client Center Paris

#solconnect13

Transformez vos opportunités en succès

IBM

ISO03 - Découvrir les connaissances dans les sources de données textuelles

Conception d'annotateurs sémantiques avec IBM Content Analytics Studio

Jean-Marc Langé

ICA Studio

- **Un environnement de conception permettant de paramétrer les annotateurs chargés d'identifier des concepts métier à partir des sources textuelles, sans nécessité d'utiliser un quelconque langage de programmation.**

- **Permet de gérer les projets, de concevoir et tester les annotateurs dans un environnement graphique, et de les déployer vers le moteur ICA.**

- **Utilise deux ressources essentielles pour générer des annotations:**
  - **dictionnaires** qui contiennent la terminologie pertinente pour l'objectif métier visé;
  - **règles** qui combinent des mots et des annotations déjà identifiées pour générer une annotation de niveau supérieur;

Comment ça marche?

Content Analytics Studio

Patient : Blessure membre inférieur

Extraction de concept

Personne | Pathologie | Partie du corps

NP	NP	Verbe	Dét	Nom	Prép	Dét	Nom
Jean	Dupont	présente	une	entorse	à	la	cheville

Annotateurs (UIMA)

API REST Temps Réel

Données non structurées

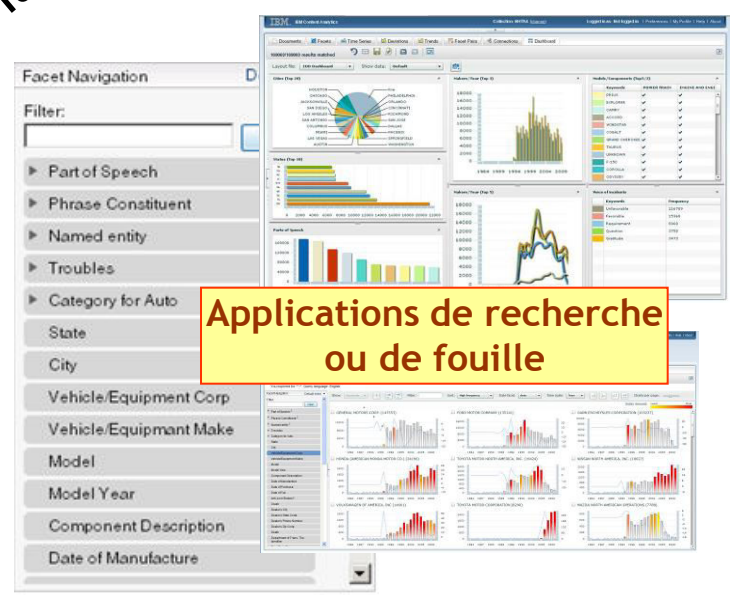
Entreprise (Centre de Contact, logs de test, notes des concessionnaires / conseillers, messagerie,ECM, etc.) et externes (Web, forums utilisateurs, réseaux sociaux, etc.)

Crawlers

Crawlers

Documents et données analysés

Avec identification de concepts



Applications de recherche ou de fouille

RDB

Cognos. software

SPSS AN IBM COMPANY

IBM MDM

RDB

i2

Cognos. software

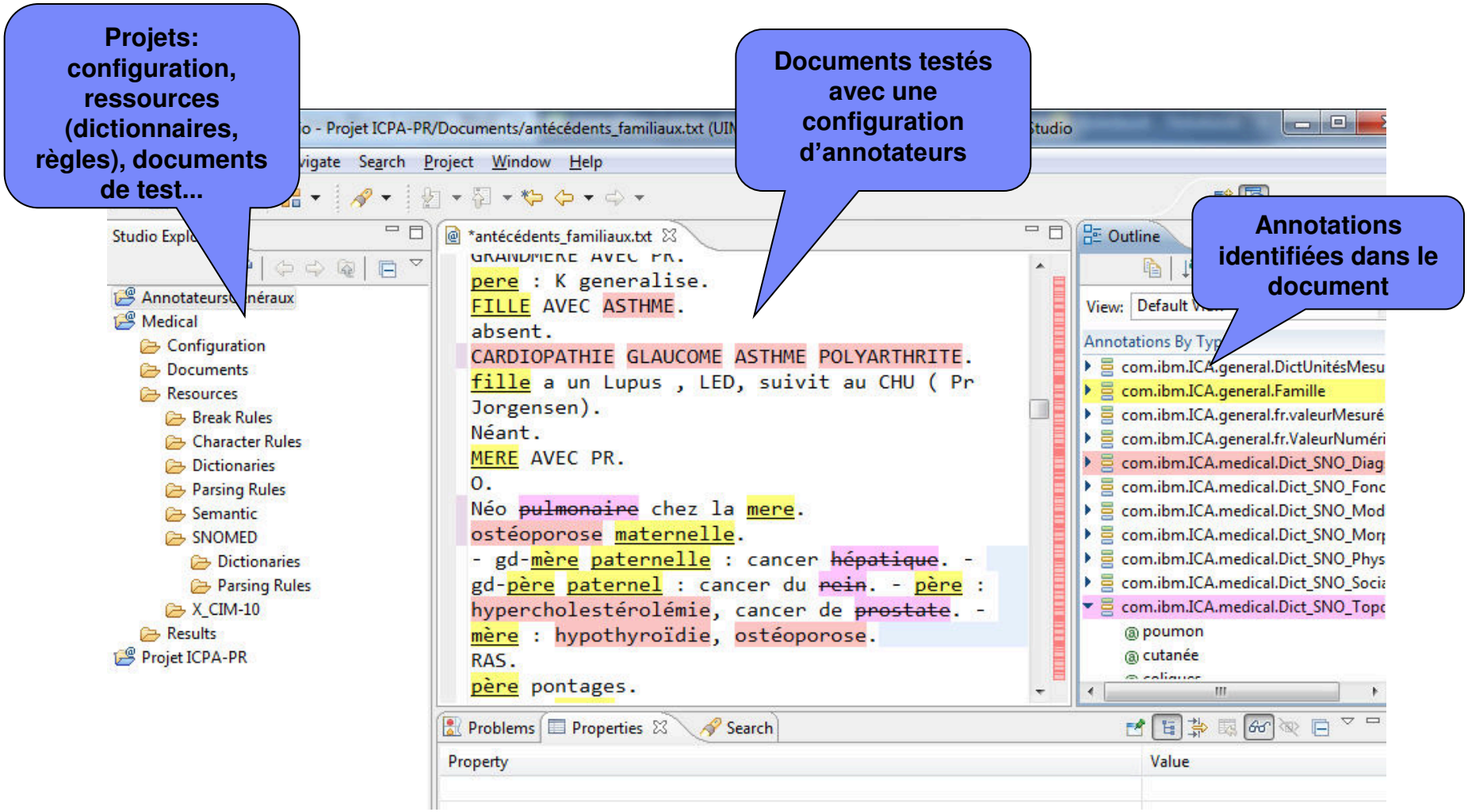
Applications tierces d'analyse / exploration

SPSS AN IBM COMPANY

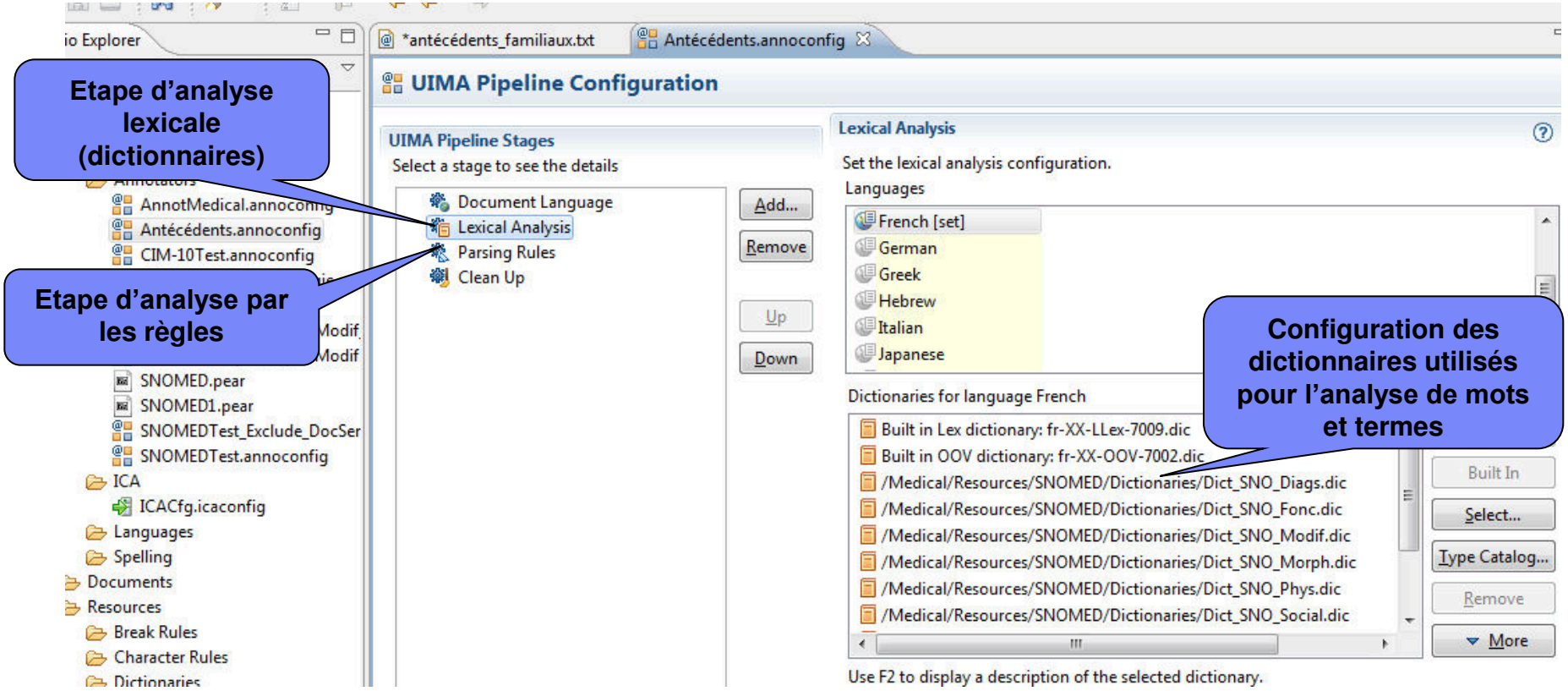
Qu'est-ce que TAL (NLP) et UIMA

- Traitement automatique du langage naturel (TAL) ou *Natural Language Processing (NLP)*, **le lien qui rend possible l'interaction entre les ordinateurs et les langages humains**
  - Watson utilise **IBM Content Analytics** pour les fonctions NLP
- Unstructured Information Management Architecture (UIMA), **une architecture ouverte pour le traitement des données textuelles et la définition de solutions d'analyse de texte**
  - Standard Open Source OASIS
  - Projet supporté par Apache Software Foundation
  - Mis en oeuvre dans plusieurs solutions IBM

Aperçu général de ICA Studio

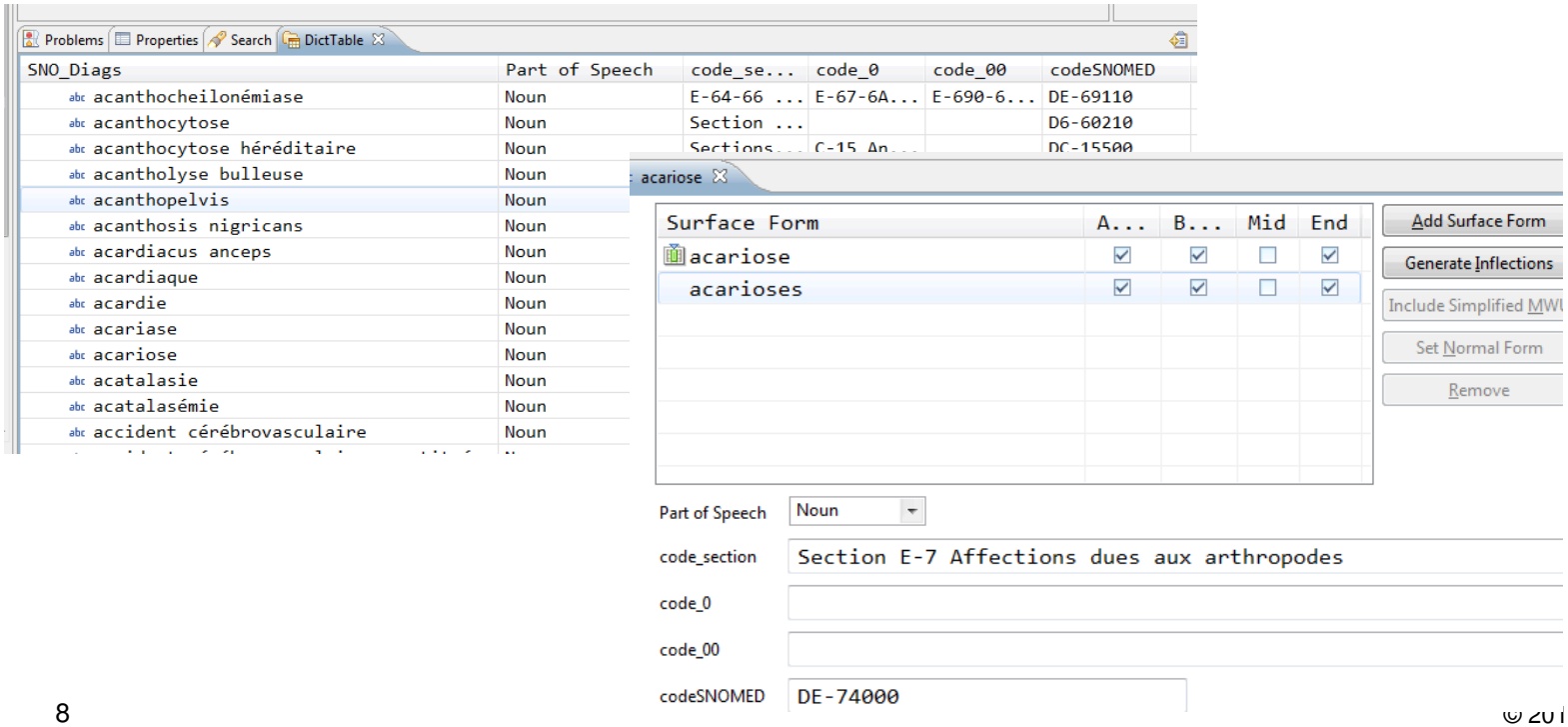


Configuration de la chaîne d'annotations



Paramétrage des dictionnaires

- **Les dictionnaires peuvent être remplis « à la main » ou par importation de fichiers en format délimité**
- **Ils contiennent les entrées, leur catégorie grammaticale, et les variantes «de surface»: formes fléchies (pluriel, féminin, formes conjuguées), abréviations...**
- **On peut associer à une entrée des attributs définis par l'utilisateur (type, classification, codification, ...)**

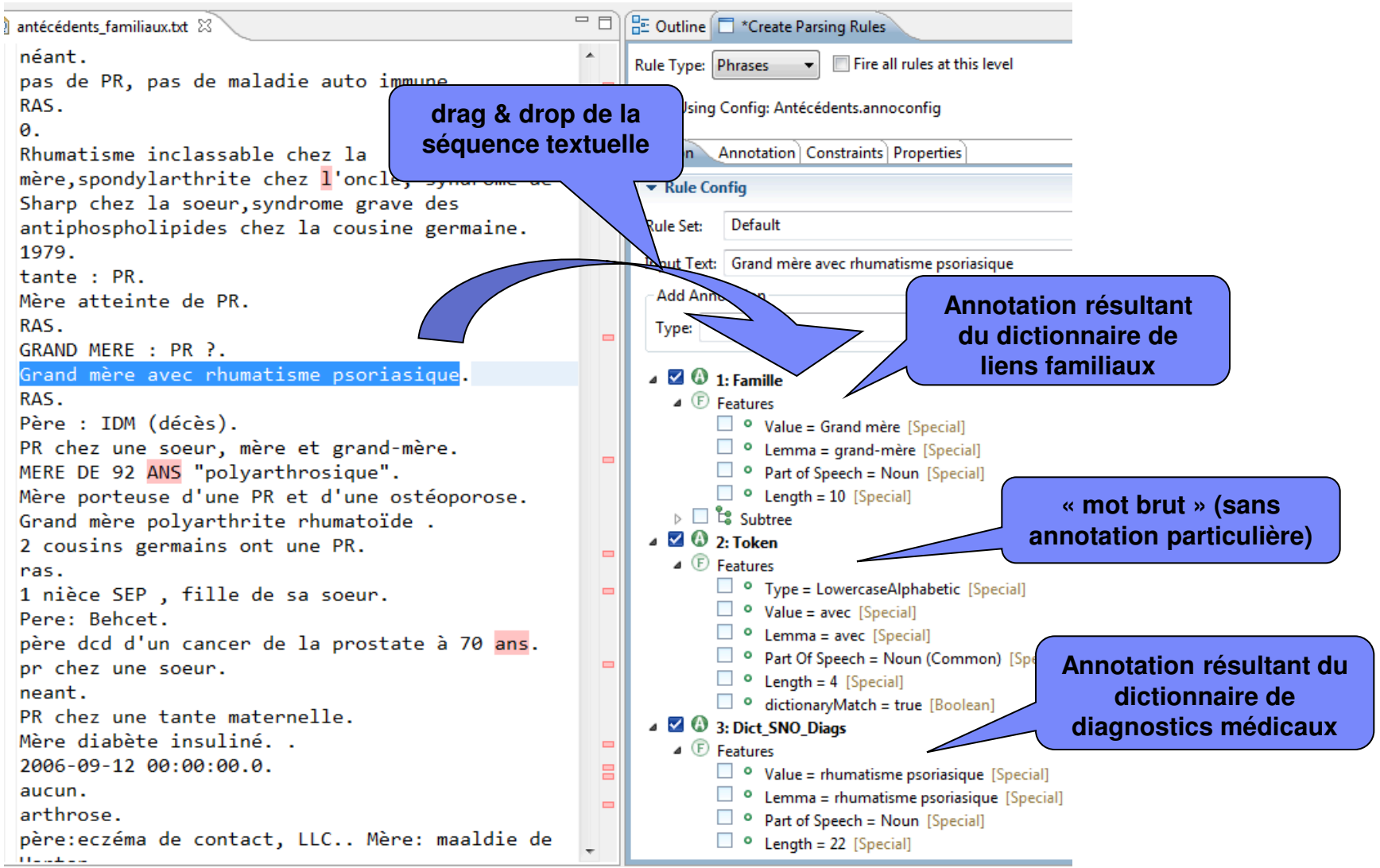


Création de règles

- Les règles sont créées en sélectionnant dans un document une séquence de mots correspondant au concept que l'on souhaite annoter, et en la faisant glisser dans l'espace de création des règles
- Dans l'espace de création, on retrouve les différentes annotations déjà identifiées (dictionnaires ou règles) et des mots « bruts » (non annotés)
- Pour chacun de ces éléments (annotations ou mots bruts), on peut sélectionner des critères (maj/min, catég. grammaticale, valeur, nbre de caractères, etc), ainsi qu'une possible répétition.
- Une fois définie la séquence à annoter et les critères, on lui associe une nouvelle annotation, et optionnellement des attributs, qu'on peut piocher dans les mots ou annotations composant cette séquence

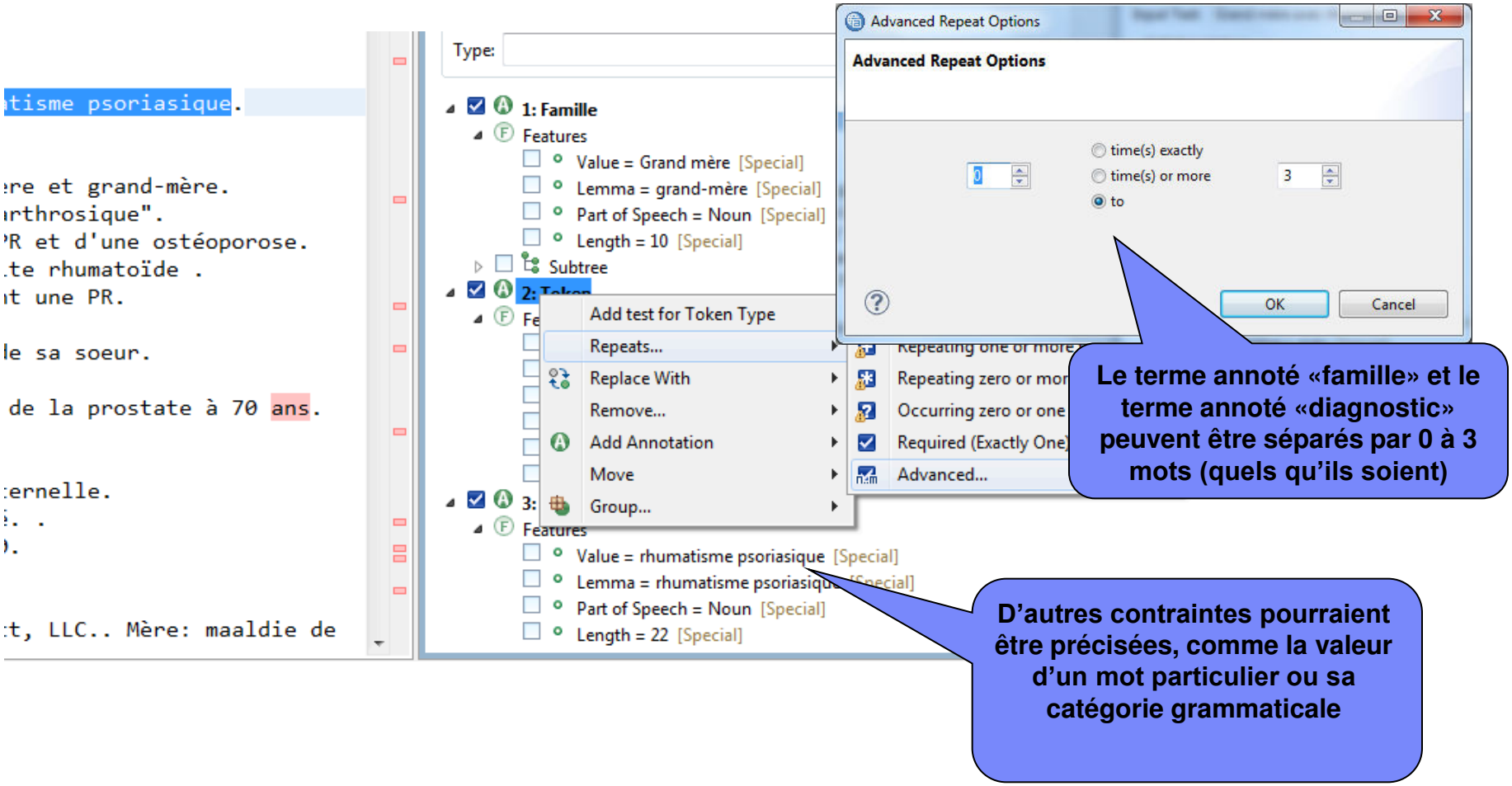
Création de règle: antécédents médicaux familiaux

Etape 1: drag&drop dans la zone de création de règles



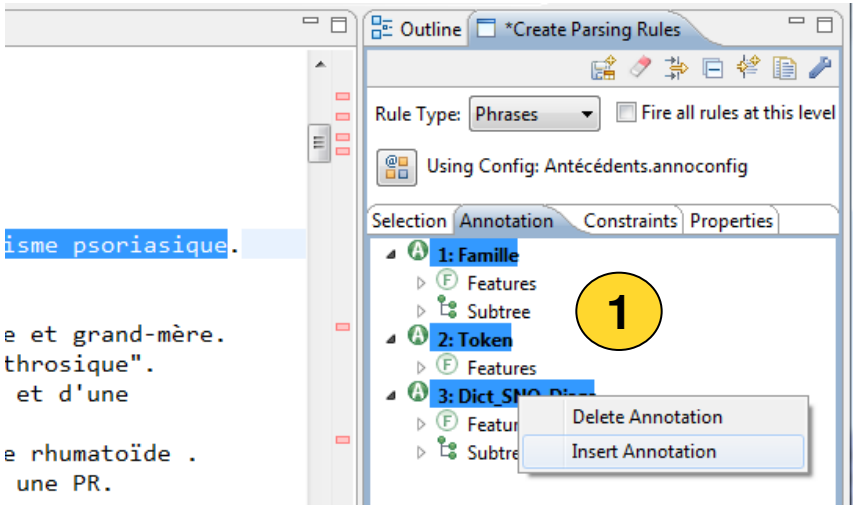
Création de règle: antécédents médicaux familiaux

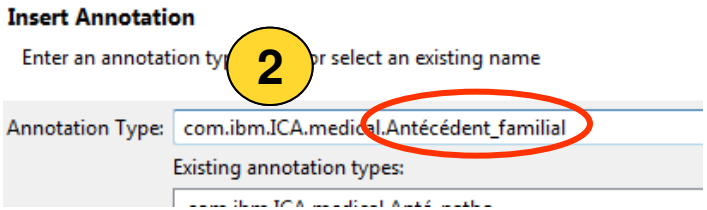
Etape 2: spécification de contraintes sur les éléments de la règle

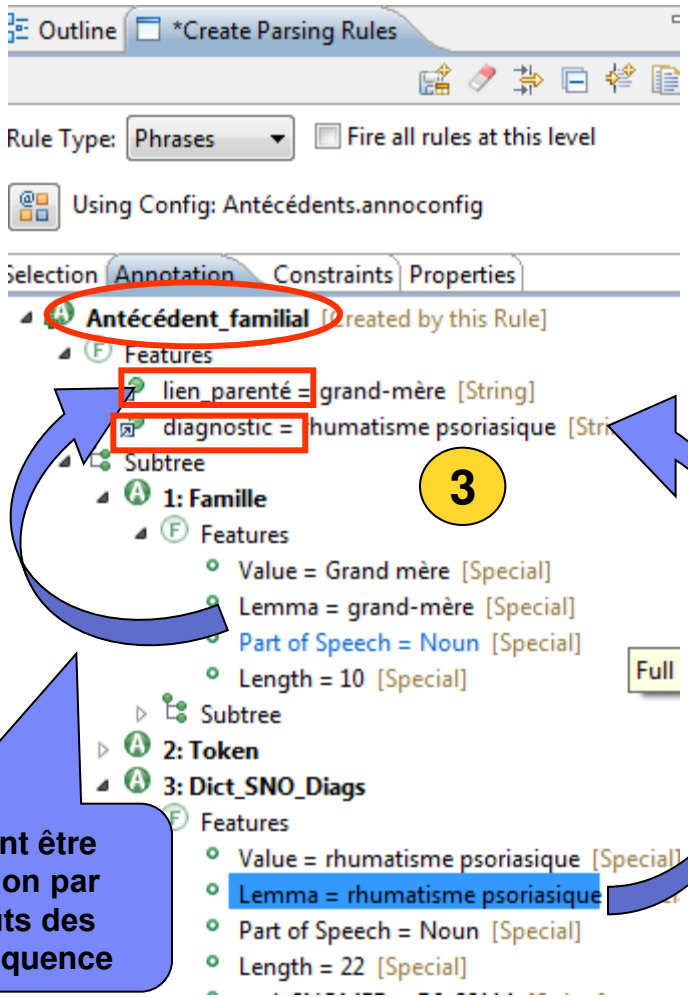


Création de règle: antécédents médicaux familiaux

Etape 3: création de l'annotation pour la séquence paramétrée



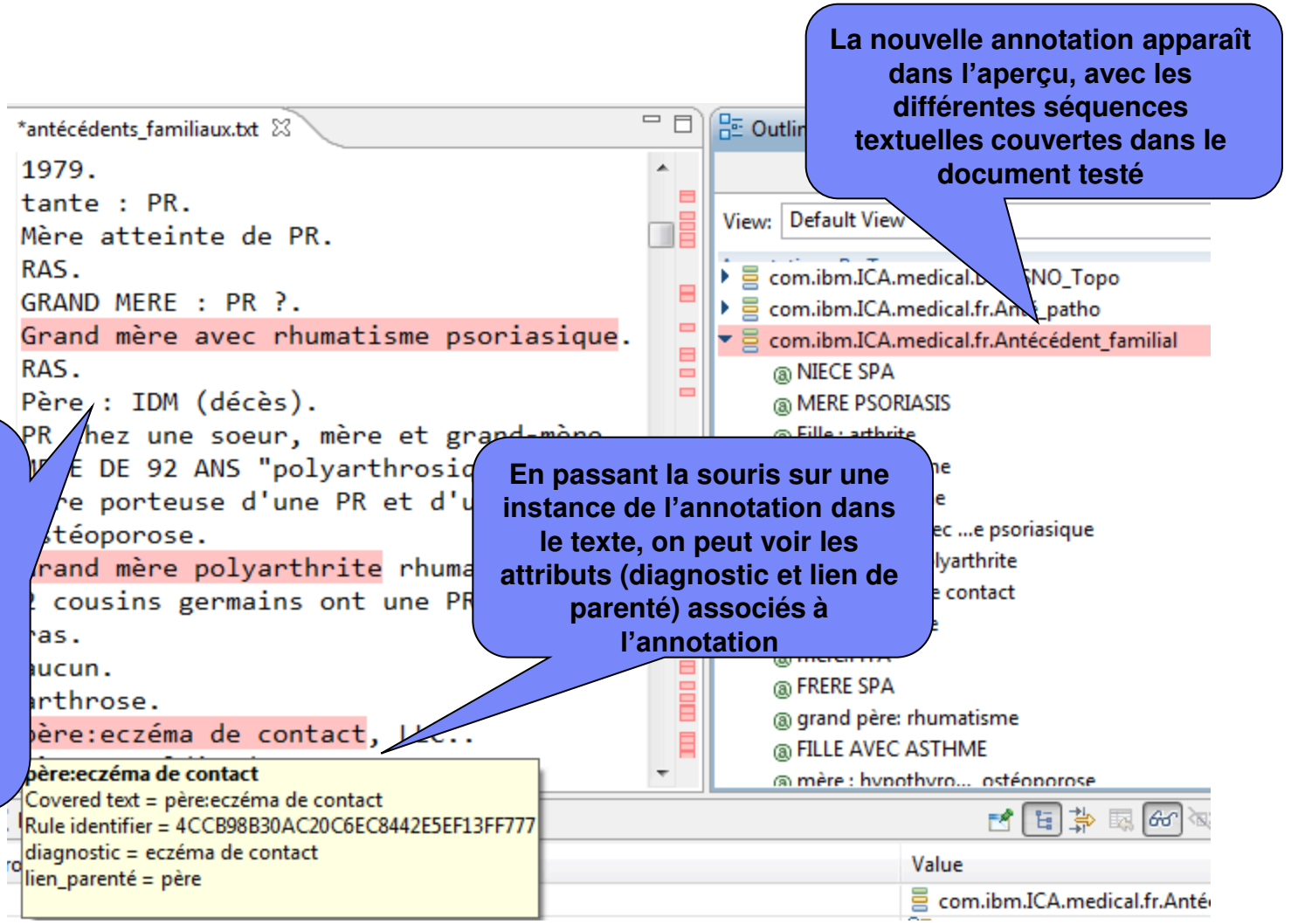




Des attributs peuvent être rajoutés à l'annotation par drag&drop d'attributs des composants de la séquence

Création de règle: antécédents médicaux familiaux

Etape 4: vérification de l'application de la règle à l'échantillon de texte



Déploiement des annotateurs vers ICA

- **Les annotateurs sont déployés vers ICA sous forme d'un fichier .pear qui inclut toutes les ressources nécessaires (dictionnaires, règles, configuration...)**
- **Un assistant permet de sélectionner les annotations à exporter, si nécessaire les attributs de ces annotations, et les facettes que ces annotations vont générer à la suite de l'indexation dans ICA.**
- **Ce même assistant permet de déclencher jusqu'à la réindexation par ICA, ce qui fait qu'aucune intervention d'administration n'est nécessaire dans ICA pour que les utilisateurs bénéficient de ces nouvelles facettes.**

La plate-forme d'échange Vivastream™

- Développez votre réseau
- Découvrez les experts sur les sujets qui vous intéressent
- Echangez avec les speakers et les experts
- Regardez qui participe aux sessions auxquelles vous êtes inscrit
- Evaluez les sessions auxquelles vous avez participé

Inscrivez-vous sur le web ou avec votre smartphone

http://www.vivastream.com/events/ibmimt-solutionsconnect-frfr

Viva | Available on the App Store | ANDROID APP ON Google play | BlackBerry App World