Representation and Processing of Composition, Variation and Approximation in Language Resources and Tools

Agata Savary

Centre Val de Loire UNIVERSITÉ
Spirit of the Loire Valley

Université François Rabelais Tours

UNIVERSITÉ FRANÇOIS - RABELAIS TOURS

Année universitaire (*Academic Year*): 2013-2014

Discipline (*Domain*): Informatique (*Computer Science*)

Dissertation en vue d'obtention d'une Habilitation à diriger des recherches (*Habilitation dissertation in view of an accreditation to supervise research*)

présentée et soutenue publiquement par (*orally presented by*)

Agata SAVARY

27 Mars 2014

devant le jury suivant (*in front of the following jury*):

Anne ABEILLÉ	Professeur des universités (*Full professor*)	Université Paris 7, France
Jean-Yves ANTOINE	Professeur des universités (*Full professor*)	Université François Rabelais Tours, France
Béatrice DAILLE	Professeur des universités (*Full professor*)	Université de Nantes, France
Jan HAJIČ	Professeur (*Full professor*)	Charles University in Prague, République Tchèque
Denis MAUREL	Professeur des universités (*Full professor*)	Université François Rabelais Tours, France
Agnieszka MYKOWIECKA	Chargée de recherche, HDR (*Accredited research fellow*)	Polish Academy of Sciences, Varsovie, Pologne
Joachim NIEHREN	Directeur de recherche (*Research director*)	Institut national de recherche en informatique et en automatique (INRIA), Lille, France

Representation et traitement automatique de la composition, de la variation et de l'approximation dans des ressources et outils linguistiques

(Representation and Processing of Composition, Variation and Approximation in Language Resources and Tools)

Agata Savary

Université François Rabelais Tours
campus de Blois
Laboratoire d'informatique
3 place Jean-Jaurès
41000 Blois, France
http://www.info.univ-tours.fr/~savary/English/indexgb.html
agata.savary@univ-tours.fr

14 Novembre 2013

Foreword

This volume contains my dissertation in view of the French HDR diploma (*Habilitation à Diriger les Recherche*) in computer science. Candidates to this diploma are supposed to demonstrate their substantial personal contribution to research, their capacity to supervise research activities, and their experience and maturity in research-related tasks such as project management, event organization, research evaluation, etc.

Since my PhD diploma in 2000, I have been active in natural language processing (NLP), computer science and linguistics. My research interests focus on two central challenges in language modeling and processing: the composition of linguistic units and the related compositionality property, as well as the variation in complex structures, notably in multi-word expressions (WMEs) and named entities (NEs). I address these challenges by defining linguistically-motivated description paradigms, as well as by automating the creation of the corresponding language resources such as electronic lexicons and annotated corpora. Additionally, I am concerned with the problems of data incorrectness, imprecision and evolution, which call for approximation and correction methods, such as approximate string matching, spelling correction or XML document correction.

I am particularly motivated by multilingual considerations about language processing. I have dedicated my efforts to different languages from different language families, notably English, French, Polish and Serbian. I deeply believe that a multilingual point of view acts in favor of a better understanding of language phenomena, and of the appropriateness and universalism of formalisms and methods.

This dissertation is organized as follows. In chapter 1 I present an extended summary of my contributions in French. I then give a more detailed description of these contributions in English. In chapter 2 I provide a general introduction of the research context, which includes the two major phenomena mentioned above: composition and variation. In chapter 3 I discuss multi-word expressions by addressing, notably, their morphosyntactic (non-)compositionality and their lexical description. In chapter 4 I focus on named entities as particular subtypes of MWEs, and I discuss their annotation, their automatic recognition and their representation in ontologies and knowledge bases. In the same chapter I extend NEs to more generally understood mentions of discourse-world entities, and I refer to the problem of coreference annotation and resolution. Chapter 5 is dedicated to formal methods based on finite-state tools for the representation and processing of linguistic data and of XML documents. In chapter 6 I describe the general framework of my work, as well as my main contributions and experiences in organizing and supervising research activities. Finally, in chapter 7 I draw conclusions from my previous work and I sketch the major perspectives for the future.

Research is not a solitary activity. The contributions presented here would not have been achieved without long-lasting or occasional support from many people and institutions. My acknowledgements go to the members of my BdTln (*Bases de Donnée et Traitement des Langues Naturelles*) research team in Blois/Tours for the inspiring and friendly atmosphere, collaboration, encouragements and advice. In their company I learned to better organize my work, to

develop curiosity about seemingly distant domains, and to draw cross-domain parallels. They also taught me the wisdom of sharing daily coffee breaks and the impact it has on collaboration and productivity.

I am grateful to my colleagues from the Linguistic Engineering Group in Warsaw, with whom I have been carrying on intensive collaboration, particularly since my sabbatical stay in 2009-2010. I highly esteem their competence and expertise in natural language processing and in computer science, and I frequently draw my inspiration from their analyses and decisions. I consider the Institute of Computer Science of the Polish Academy of Sciences as my second informal affiliation, and I owe it a large part of my scientific results.

I give thanks to my other external collaborators from the Universities of Belgrade, Gdańsk, Marne-la-Vallée, Olsztyn, Orléans, Poznań and Tomsk, as well as from the PARSEME COST action. Contacts with these excellent experts and friendly colleagues increased my open-mindedness and provided motivation to my work.

I am greatly honored by the presence of prominent researchers in my habilitation jury. I highly appreciate their interest in my work and I am looking forward to their expert and demanding evaluation of my contributions.

I am also indebted to dozens other researchers from different countries most of whom I do not personally know but who inspire and lead me via their revisions of my publications and projects, their efficient research event organization, and especially via their high quality publications. Many great papers which I read made me open my eyes on new problems and gain a better understanding of my subjects of study.

Last but not least, my professional achievements would not be possible without continuous support from my family and friends. It is to them that I dedicate this volume.

Agata Savary
Blois,
November 2013

Contents

Chapter 1

Résumé

Cet volume contient une dissertation en vue de l'obtention du diplôme de l'Habilitation à Diriger des Recherches (HDR) dans le domaine de l'informatique. Je présente ici mes travaux de recherche effectués depuis ma thèse de doctorat en 2000. Il s'agit d'un travail pluridisciplinaire concernant des thèmes liés au traitement automatique des langues (TAL), à la linguistique et à l'informatique.

1.1 Composition et variation

Depuis plus de dix ans, je m'occupe des phénomènes de *composition* et de *variabilité* des unités linguistiques. Dans le chapitre 2 je me penche sur la définition des ces deux propriétés essentielles. D'après des travaux en philosophie et mathématiques, tels que (Pagin & Westerståhl, 2001a; Kracht, 2007), la composition, évoquée déjà par Frege (Janssen, 2001), n'a pas été rigoureusement décrite jusqu'aux années 2000, même si elle a été depuis longtemps considérée comme propriété essentielle en linguistique, philosophie du langage, logique et informatique. La définition largement acquise, citée par Kracht (2007), est la suivante : *une expression composée est compositionnelle si sa signification est une fonction des significations de ses constituants et d'une règle syntaxique par laquelle ils sont combinés.* Kracht remet en cause cette définition, en soutenant que la compositionnalité ne peut pas être considérée pour une expression en tant que telle, mais seulement pour son analyse grammaticale et sémantique. En d'autres termes, un langage est compositionnel s'il possède une grammaire compositionnelle.

Baggio et al. (2012) rappellent les raisons pour lesquelles la compositionnalité est souhaitable dans l'analyse linguistique, en mentionnant: (i) la productivité (le nombre de phrases possibles est infini, alors que le cerveau humain n'a qu'une capacité limitée de stockage), (ii) la systématicité (l'humain est doté de compréhension par analogie), (iii) la méthodologie (le calcul sémantique est à mener de manière compositionnelle), (iv) la modularité (l'encapsulation d'informations dans la description de structures linguistiques est souhaitable).

Pagin & Westerståhl (2001b) mentionnent que la compositionnalité des langues naturelles n'est pas indiscutable pour plusieurs raisons, dont celle qui nous intéresse particulièrement dans cette thèse : l'existence de contre-exemples tels que les phrases de conviction (*belief sentences*), les citations (les deux remettent en cause le principe de substituabilité de synonymes), ainsi que les *idiomes*. Cette thèse s'intéresse notamment aux *unités* (ou *expressions*) *polylexicales* (UP), qui constituent une classe plus large que les idiomes et dont l'une des propriétés définitoire est la non compositionnalité ou une compositionnalité atypique.

L'utilité du principe de compositionnalité consiste notamment à permettre d'éviter l'explosion combinatoire des cas lexicalisés. Mes propres travaux fournissent un exemple de ce phénomène. Dans (Tallec et al., 2009) et (Tallec et al., 2010b) nous présentons le projet EmotiRob qui a

eu pour but la création d'un prototype de robot compagnon émotionnel pour enfants fragilisés. Le robot devait réagir via des expressions de visage au contenu linguistique des énoncés d'un enfant. Suite à la reconnaissance et la transcription de la parole, l'énoncé était soumis à l'analyse syntaxique en dépendances par le système *Emologus*. Le calcul de la valence émotionnelle de l'énoncé suivait le principe de compositionnalité. Nous avons admis que les mots du lexique de base peut être associés à des valeurs émotionnelles atomiques (inclues dans l'intervalle $[-2; 2]$) et que les prédicats modifient les valeurs émotionnelles de leurs arguments. Par exemple, le verbe *casser* inverse la valence de son argument alors que l'adjectif *mignon* la renforce. Des expériences avec un corpus de comptes enfantins annoté manuellement confirment nos hypothèses (Tallec et al., 2010a): la valence est déterminée de manière correcte pour 90% des énoncés. Des résultats semblables, à la hauteur de 87,9% d'exactitude, ont été obtenus par Neviarouskaya et al. (2010), qui considère une panoplie plus large de caractéristiques (le type de l'émotion, sa polarité, sa valence et son niveau de confiance). Notons également que certaines études dédiées aux unités polylexicales (Klebanov et al., 2013) démontrent leur degré élevé de compositionnalité émotionnelle malgré leur opacité sémantique.

La compositionnalité est au coeur des débats linguistiques depuis plusieurs décennies, notamment au sujet des unités polylexicales (*Multi-Word Expressions*, *MWEs*). Ces unités, définies plus largement dans le chapitre 3, incluent des objets très hétérogènes tels que les mots composés, les termes complexes, les entités nommées multi-mots, les constructions à verbe support, les idiomes, etc. Les définitions de ces notions et de leurs frontières sont des questions très controversées (Habert & Jacquemin, 1993; Downing, 1977; Fabre & Sébillot, 1996; Benveniste, 1974; Lyons, 1978).

La compositionnalité peut s'appliquer non seulement au domaine de la sémantique, mais aussi à la morphologie des unités polylexicales (Mel'čuk, 2010). Dans (Savary et al., 2007) nous nous penchons sur les problèmes de la *non-compositionnalité flexionnelle* des unités polylexicales en français, en polonais et en serbe. Un mot composé est considéré comme compositionnel lorsque ses propriétés flexionnelles peuvent être totalement déduites de ses composants et de sa structure syntaxique. Ainsi, par exemple le nom composé :

(1.1) *un perce-neige*

n'est pas compositionnel car il est au masculin alors que le seul substantif qu'il contient, *neige*, est au féminin.

La non-compositionnalité sémantique et morphologique est liée à l'idée de la *lexicalisation*. Si la signification, le référent ou la flexion d'une expression sont imprédictibles, cette expression est lexicalisée, c'est-à-dire doit être explicitement décrite dans un lexique afin de permettre son analyse appropriée. Dans les sections 3.3 et 3.6 je présente mes contributions à la description lexicalisée des unités polylexicales contiguës, qui consiste en un formalisme et son implantation pour la prise en compte des idiosyncrasies morphosyntaxiques.

Les débats sur la nature des unités polylexicales font souvent appel à la notion du (degré de) figement (Gross, 1988, 1990). Cependant, la deuxième caractéristique centrale de ces unités, contraire au figement, est celle de la variabilité linguistique. En effet, la plupart des UP sont partiellement figées et partiellement variables. La variabilité a été largement étudiée dans par la communauté de l'extraction terminologique, car près de 30% des termes apparaissant dans des corpus sont des variantes des termes contrôlés (contenus dans des listes et lexiques) (Jacquemin, 2001). Dans (Savary & Jacquemin, 2003) nous avons repris et peaufiné la définition d'une variante terminologique (Jacquemin, 2001), qui peut être :

- graphique : *behavioural model* → *behavioral model*,
- morphologique : *students union* →*student union*, *image converter* →*image conversion*,

- sémantique : *automobile cleaning* → *car washing*,
- syntaxique : *date of birth* → *birth date*, *processing of cardiac image* → *image processing*.

Nous avons ensuite effectué une étude contrastive détaillée des systèmes d'extraction terminologique dédiés à deux types d'applications:

- l'acquisition terminologique : ACABIT (Daille, 1994, 1996), ANA (Enguehard & Pantera, 1995), LEXTER (Bourigault, 1993, 1994, 1996), TERMINO (David & Plante, 1990a,b), TERMS (Justeson & Katz, 1995) et Xtract (Smadja, 1992),
- l'indexation par phrases (utilisant des EP comme termes) : CLARIT (Evans et al., 1991), COP (Metzler & Haas, 1989; Metzler et al., 1989, 1990), COPSY (Schwarz, 1989, 1990), l'indexeur de Fagan (Fagan, 1987), FASIT (Dillon & Gray, 1983), IRENA (Arampatzis et al., 1997, 1998), NPtool (Voutilainen, 1993), l'indexeur Sheridan/Smeaton (Smeaton & Sheridan, 1991; Sheridan & Smeaton, 1992), le générateur de variantes de Sparck Jones/Tait (Sparck Jones & Tait, 1984b,a), SPIRIT (Andreewsky et al., 1977) et TTP (Strzalkowski & Vauthey, 1992; Strzalkowski, 1994, 1995; Strzalkowski & Scheyen, 1996).

Nous nous sommes notamment intéressés à la manière et au degré de la prise en compte de la variation terminologique dans ces systèmes et nous avons décrit le système FASTR (Jacquemin, 2001), qui met ce phénomène au coeur de la reconnaissance des termes.

Il semblerait cependant qu'au moins jusqu'aux années 2000 l'intérêt de l'usage des UP et de l'analyse syntaxique pour les applications telles que la recherche d'information était très controversé (Brants, 2003). La mise à jour de cet état de l'art pourrait démontrer si dans ce domaine *le pendule est*, effectivement, *monté trop haut* (Church, 2011) et si un renouveau de la volonté de *cueillir des fruits accrochés plus en hauteur* apparaît dans la recherche fondamentale comme appliquée.

1.2 Unités polylexicales

Les propriétés le plus souvent évoquées dans diverses définitions des UP (Benveniste, 1974; Downing, 1977; Levi, 1978; Gross, 1990; Silberztein, 1993b; Gross, 1996; Cadiot, 1992; Sag et al., 2002; Derwojedowa & Rudolf, 2003) sont les suivantes :

- les UP sont composées d'au moins deux mots,
- elles se caractérisent par un degré de non-compositionnalité (ou idiosyncrasie) morphologique, distributionnelle ou sémantique,
- elles ont des référents uniques et constants.

Notons que les termes élémentaires utilisés dans ces définitions tels que mot, référence ou non-compositionnalité, sont eux mêmes controversés. C'est pourquoi dans nos travaux nous définissons la porté des UP de manière pragmatique: une UP est une séquence d'unités graphiques qui, pour des raisons propres à une (ou des) application(s) doit être listée, décrite et traitée comme une unité (Savary, 2005).

Les faits principaux concernant les UP sont:

- leur prédominance dans les langues naturelles (Gross & Senellart, 1998; Sag et al., 2002),
- leur comportement Zipfien (*data scarcity*),

- leur comportement idiosyncratique à différents niveaux de traitement linguistique: la segmentation (*bonshommes*, *aujourd'hui*), la morphologie (*perce-neige*, *grand-mères*), la syntaxe (*prendre une veste* vs. **la veste a été prise*), la sémantique (*prendre une veste* = subir un échec).

Les UP sont de nature très hétérogène, ce qui est reflété notamment par leurs différentes typologies (Sag et al., 2002; Mel'čuk, 2010).

Dans le chapitre 3.2 nous étudions l'état de l'art dans la représentation lexicale et le traitement automatique de UP. Nous rappelons notamment notre étude contrastive (Savary, 2008) des méthodes de description de UP par rapport à leurs propriétés flexionnelles (Courtois & , eds.; Silberztein, 1993a; Savary, 2000; Kyriacopoulou et al., 2002; Silberztein, 2005; Savary, 2008; Karttunen et al., 1992; Karttunen, 1993; Breidt et al., 1996; Oflazer et al., 2004; Alegria et al., 2004; Sag et al., 2002; Copestake et al., 2002; Villavicencio et al., 2004; Jacquemin, 2001). Suite à elle nous avons proposé des recommandations de meilleures pratiques telles que:

- la prise en compte d'une variété de langues en vue de l'universalisme du modèle,
- la description à deux couches (identification morphologique des composants, puis description de leurs combinaisons valables),
- le besoin de mécanismes d'unification pour la représentation compacte des paradigmes flexionnels,
- la numérotation des composants pour la représentation des variantes syntaxiques (ellipses, changements d'ordre etc.),
- le développement des plateformes lexicographiques pour l'automatisation de la description des UP, etc.

Certains travaux plus récents tels que (Itai & Wintner, 2013) semblent confirmer l'utilité de ces recommandations. L'approche de Grégoire (2010) va au delà de cet état de l'art : (i) en se consacrant à une panoplie large des UP non contiguës, notamment verbales, (ii) en introduisant des classes de flexion paramétrables pour limiter leur nombre, et (iii) en appliquant le lexique ainsi obtenu à l'analyse syntaxique. Cette proposition semble très prometteuse notamment dans le cadre du projet PARSEME décrit dans la suite de cette thèse.

La suite de l'état d'art des UP fait un panorama des méthodes existantes en extraction d'UP (Davis & Barrett, 2013; Pecina, 2010; Al-Haj & Wintner, 2010; Tsvetkov & Wintner, 2010; Morin & Daille, 2010; Delpech et al., 2012; Ramisch et al., 2010), leur identification dans le corpus (Vincze et al., 2013), leur annotation (Abeillé et al., 2003; Bejček & Straňák, 2010; Bejcek et al., 2011; Laporte et al., 2008a,b; Kaalep & Muischnek, 2008), ainsi que leur analyse syntaxique (Abeillé & Schabes, 1989; Sag et al., 2002; Copestake et al., 2002; Villavicencio et al., 2004; Attia, 2006; Nivre & Nilsson, 2004; Constant et al., 2012, 2013; Wehrli et al., 2010; Finkel & Manning, 2009a; Green et al., 2011, 2013). Ces analyses montrent notamment que, malgré la grande quantité des travaux consacrés à la problématique des UP, relativement peu de solutions existent pour les UP non contiguës.

La suite du chapitre 3 est consacrée à la description de *Multiflex*. C'est un formalisme et un outil pour la description lexicalisée des UP contiguës, qui permet la prise en compte à la fois de leur variabilité et de leur comportement idiosyncratique. Il se base sur une approche à deux couches (cf. plus haut). Premièrement, il admet que les mots simples peuvent être analysés et générés par un module morphologique externe. Ensuite, on spécifie comment combiner les formes fléchies des composants simples pour obtenir les formes fléchies des UP qui les contiennent. Les

variantes orthographiques et, partiellement, syntaxiques peuvent être décrites dans le même cadre.

Exemple 1.2 contient les variantes flexionnelles et syntaxiques du nom de personne polonais *Jan Rodowicz "Anoda"*, qui contient un prénom, un nom et un pseudonyme. Chaque forme est annotée avec :

- sa forme de base (lemme),
- ses traits flexionnels: singulier (sg), nominatif (nom), génitif (gen), genre masculin humain ($m1$), etc.,
- un trait pragmatique éventuel: forme officielle ($offic$), forme préférée en langage parlé ($spok$), forme neutre ($neut$), etc.

(1.2)

	Variante	Lemme	Traits
(PL)	*Jan Rodowicz „Anoda"*	Jan Rodowicz „Anoda"	sg:nom:m1:offic
	Jana Rodowicza Anody	Jan Rodowicz „Anoda"	sg:gen:m1
	Jan „Anoda" Rodowicz	Jan Rodowicz „Anoda"	sg:nom:m1
	J. Rodowicz „Anoda"	Jan Rodowicz „Anoda"	sg:nom:m1
	J. Rodowicz	Jan Rodowicz „Anoda"	sg:nom:m1
	„Anoda" Rodowicz	Jan Rodowicz „Anoda"	sg:nom:m1
	Rodowicz	Jan Rodowicz „Anoda"	sg:nom:m1:spok
	Rodowicz	Jan Rodowicz „Anoda"	sg:nom:m1:neut
	...		

Afin que la génération de cet ensemble complexe de formes soit possible, les composants simples (y compris les séparateurs) sont d'abord numérotés et analysés morphologiquement, comme dans la figure 1.1.

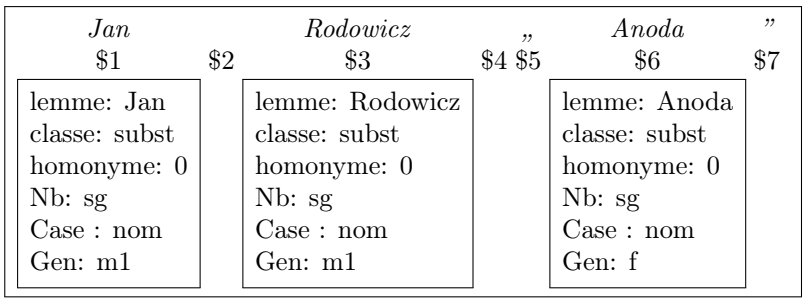


Figure 1.1: Identification morphologique des composants du nom de personne polonais *Jan Rodowicz „Anoda"*

A tout le nom composé on attribue ensuite le graphe flexionnel de la figure 1.2. La génération des variantes s'effectue en parcourant les différents chemins du graphe. Un chemin commence par la flèche la plus à gauche et se termine dans la boîte encerclée à droite. Chaque boîte sur le chemin décrit un composants (éventuellement vide). Des variables d'unification permettent d'assurer l'accord entre composants. Par exemple le chemin du milieu de la figure 1.2 produit le premier composant (*Jan*) décliné ($\langle \$1 : Case = \$c \rangle$), car la variable d'unification $\$c$ peut être instanciée avec n'importe lequel des 7 cas du polonais (décrits dans un fichier de configuration). Le composant 2 (espace) est ensuite recopié tel quel ($\langle \$2 \rangle$), tandis que le composant 3 (*Rodowicz*) est décliné à condition de s'accorder avec le composant 1, ce qui est assuré par la variable d'unification commune $\$c$. De la même manière, le composant 6 (*Anoda*) s'accorde

avec les deux noms précédents. Les équations morphologiques en dessous du chemin permettent d'obtenir les traits flexionnels de chaque forme composée. Ici, le trait pragmatique $Usage$ prend la valeur $offic$ ou vide ($\langle E \rangle$), le genre et le nombre sont hérités du premier composant ($Gen = \$1.Gen; Nb = \$1.Nb$), tel qu'il apparaît dans le lemme ($m1$ et sg) et le cas s'accorde avec celui du premier composant dans la forme fléchie correspondante ($Case = \$c$).

Les autres chemins du graphe fonctionnent de la même manière, tout en produisant des variantes graphiques et syntaxiques par l'ellipse et le ré-ordonnancement des composants. Au total, grâce à la factorisation due à l'unification et l'alternative, ce graphe permet d'obtenir les 126 variantes valables du nom.

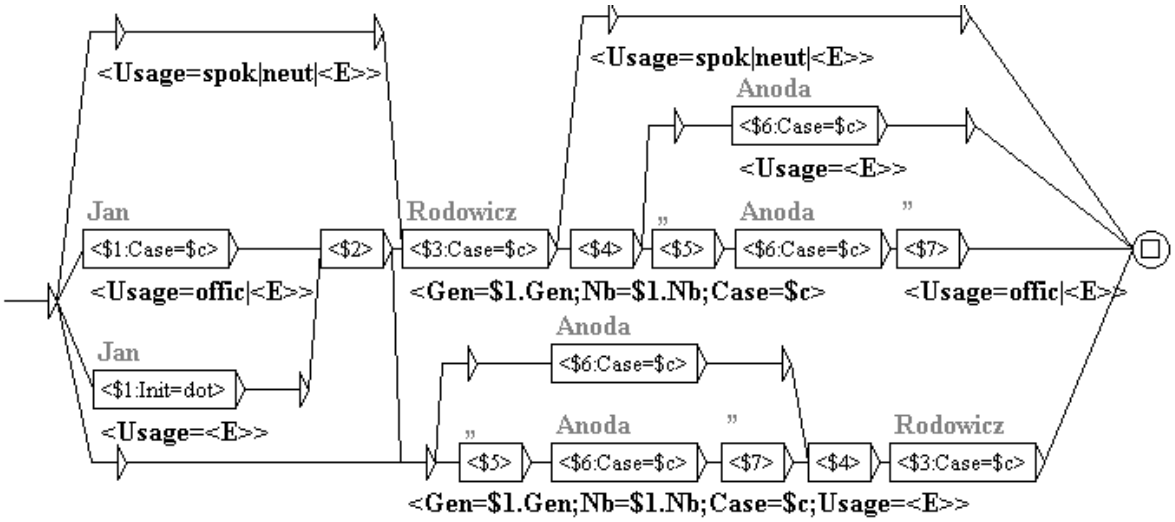


Figure 1.2: Graphe flexionnel pour le nom *Jan Rodowicz „Anoda”*

Le formalisme de Multiflex assure la représentation d'autres propriétés d'UP contiguës telles que l'exocentrisme (*perce-neige*), les accords irréguliers (*grands-mères*), la coordination (*Adam et Eve*), les fluctuations du genre (PL: *czerwony pająk*$_{m1|m2}$ 'araignée rouge'), les valeurs vides (PR: *ponto de água* 'aqueduc', **pontinho de água* 'small aqueduc'), le changement de tête (EN: *United Nations Organisation*, *United Nations*), l'omission ou l'insertion de séparateurs (SR: *radio aparat*, *radio-aparat*, *radioaparat*), les paradigmes défectifs (*wybory powszechne* 'élections nationales', **wybór powszechny* 'élection nationale'), l'insertion de composants externes (PL: *Mieszko I*, *Mieszko Pierwszy* 'Mieszko the First') et l'imbrication d'une UP dans une autre. Ce dernier phénomène peut être illustré par le nom de rue dans l'exemple (1.3), qui contient le nom de personne de l'exemple (1.2). Notons que ce dernier est ici représenté en tant que composant unique fléchi selon le graphe de la figure 1.2.

(1.3)

	Variante	Lemme	Traits
(PL)	*aleja Jana Rodowicza „Anody”*	aleja Jana Rodowicza „Anody”	sg:nom:f:offic
	al. Rodowicza	aleja Jana Rodowicza „Anody”	sg:nom:f:neut
	Rodowicza	aleja Jana Rodowicza „Anody”	sg:nom:f:spok
	aleja Jana Rodowicza Anody	aleja Jana Rodowicza „Anody”	sg:nom:f
	aleja J. „Anody” Rodowicza	aleja Jana Rodowicza „Anody”	sg:nom:f
	al. Jana Rodowicza	aleja Jana Rodowicza „Anody”	sg:nom:f
	J. „Anody” Rodowicza	aleja Jana Rodowicza „Anody”	sg:nom:f
	...		
	'avenue de Jan Rodowicz „Anoda”'		

La génération automatique des formes d'une UP revient à l'exploration de son graphe flexionnel en profondeur. La complexité en temps de cette opération est de $O(p \times v^{2 \times c \times w} \times s)$,

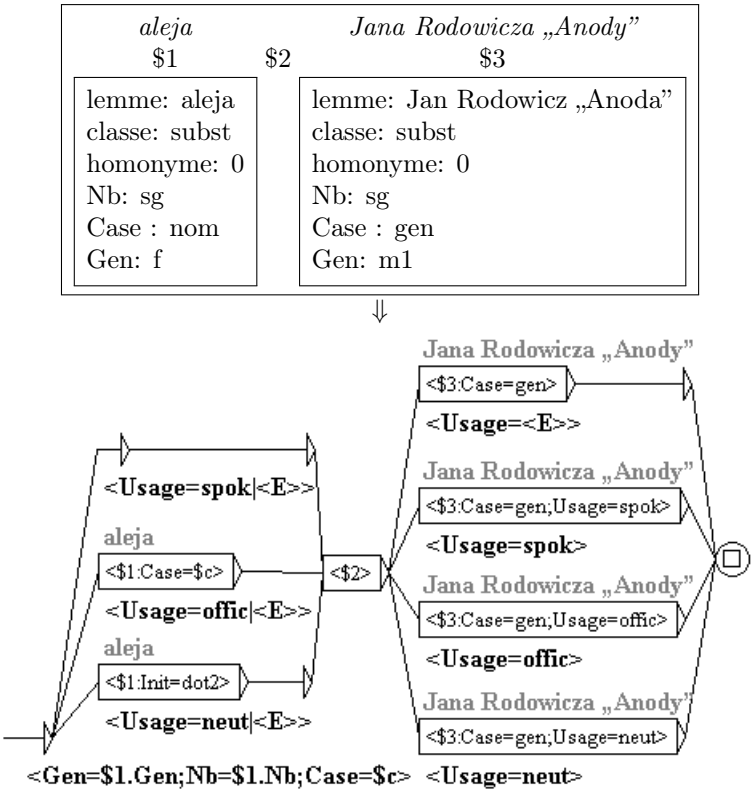


Figure 1.3: Nom de rue composé *aleja Jana Rodowicza „Anody”* contenant un nom de personne imbriqué en polonais

où p est le nombre maximal de chemins dans un graphe, v – le nombre maximal de valeurs flexionnelles (*sg*, *pl*, *nom*, *gen*, etc.) pour une catégorie flexionnelle (*Nb*, *Gen*, *Case*, etc.), c – le nombre maximal de catégories en lesquelles une classe (nom, adjectif, verbe, etc.) peut se fléchir, w – le nombre de composants de l'UP, et s – le coût maximal de génération d'une forme fléchie étant donné son lemme et ses traits flexionnels souhaités.

Différents aspects de Multiflex ont été décrits dans plusieurs publications. Dans (Savary, 2005) nous introduisons le formalisme de graphes flexionnels pour les UP contiguës, en prenant en compte l'unification et l'héritage. Dans (Savary et al., 2007) nous étudions la non-compositionnalité morpho-syntaxique et sa représentation par graphes en français, polonais et serbe. Dans (Savary, 2008) nous comparons le formalisme avec d'autres méthodes et outils dédiés à la description lexicale des UP. Dans (Savary et al., 2009) nous évoquons les spécificités du polonais, nous introduisons le mécanisme d'imbrication et nous évoquons l'interopérabilité de l'outil. Dans (Savary, 2009) nous décrivons l'implantation de Multiflex basée sur des outils à états finis et nous décrivons ses applications. Enfin dans (Graliński et al., 2010) nous effectuons une étude de l'usabilité du formalisme et de son interface graphique associée.

Multiflex, en tant qu'outil de description morpho-syntaxique des UP, est indépendant du module morphologique sous-jacent pour la morphologie des mots simples, à quelques conditions d'interopérabilité près: un modèle commun de la morphologie, une définition opératoire de l'unité graphique, et une génération à la demande de formes fléchies souhaitées. A ce jour, Multiflex possède une interface avec deux modules morphologiques différents. Premièrement, il collabore avec l'analyseur et le générateur morphologique multilingue du système Unitex[1] (Paumier, 2008). De ce fait il a été entièrement intégré sous Unitex, où il permet la flexion automatique de

[1] `http://www-igm.univ-mlv.fr/~unitex/`

Table 1.1: Dictionnaires électroniques d'UP produits avec Multiflex et ses prédécesseurs

Dictionnaire	**Langue**	**Types d'UP**	**Plateforme lexicogr.**	**Taille**			**Accessibilité**		
				Lemmes	**Graphes**	**Formes**	**Lemmes**	**Graphes**	**Formes**
DELAC anglais	anglais	noms généraux	Intex	60,000	NA	110,000	no	NA	LGPL-LR[2]
DELAC serbe	serbe	noms & adjectifs généraux	LeXimir	11,000	115	204,500	auprès des auteurs		
DELAC grec	Grec moderne	noms généraux du type A(A)N	Unitex	inconnue	inconnue	inconnue	auprès des auteurs		
SAWA	polonais	noms propres urbains	Toposław	9,000	450	309,000	CC-BY SA[3]		
SEJF	polonais	noms, adjectifs & adverbes généraux	Toposław	3,200	140	68,000	CC-BY SA		
SEJFEK	polonais	termes nominaux économiques	Toposław	11,000	290	146,000	CC-BY SA		

dictionnaires électroniques de mots composés (appelés des DELAC), qui sont ensuite appliqués à l'analyse morphologique de textes, tenant compte des UP. L'interface Multiflex-Unitex fait également partie d'une plateforme lexicographique serbe WS4LR (Krstev et al., 2006a), renommée en LeXimir (Krstev et al., 2013), qui possède notamment des fonctionnalités de prédiction automatique de graphes dont l'exactitude varie entre 58% et 86%. Deuxièmement, Multiflex offre une interface avec l'analyseur et le générateur morphologique du polonais, Morfeusz (Woliński, 2006), dans le cadre de la plateforme lexicographique Toposław (Marciniak et al., 2009b; Sikora & Woliński, 2009), qui contient notamment des modules de création, recherche, debugging et gestion automatisée de graphes.

Ces applications ont permis la création de plusieurs dictionnaires électroniques grammaticaux d'UP, résumés dans le tableau 1.1.

1.3 Entités nommées et au-delà

Les noms propres et, plus généralement, les entités nommées (EN) sont porteuses de charges sémantiques particulièrement élevées, car elles se réfèrent aux personnes, lieux, objets, concepts et évènements cruciaux pour la compréhension du texte. Leur rôle central en TAL est indéniable. Elles constituent de bons candidats pour des termes d'indexation et de catégorisation de documents. Elles sont soumises à des règles de traduction spécifiques. Elles jouent des rôles clefs dans l'extraction de l'information et les systèmes question/réponse. La modélisation et le traitement efficaces des EN nécessitent des ressources et outils complémentaires décrivant des phénomènes au niveau morphologique, syntaxique, sémantique et du discours.

Le chapitre 4 est dédié plus spécifiquement aux EN polylexicales. Nous démontrons que de telles EN dominent sur les EN uni-mot à la fois dans les dictionnaires électroniques spécialisés et dans les corpus. D'autres part, nous soulignons l'importance quantitative du phénomène de l'imbrication d'EN dans d'autres EN.

Dans la section 4.2 nous résumons l'état de l'art dans le traitement automatique des EN. Nous nous référons notamment à la tâche de l'annotation des EN en corpus, surtout lorsqu'elle est effectuée dans le cadre de modélisation linguistique à grande échelle et relativement indépendante des visées applicatives (Bejček & Straňák, 2010; Desmet & Hoste, 2010; Hinrichs et al., 2005a). Nous faisons ensuite un panorama du domaine de reconnaissance des entités nommées (REN).

[2] `http://infolingu.univ-mlv.fr/DonneesLinguistiques/Dictionnaires/telechargement.html`
[3] `http://creativecommons.org/licenses/by-sa/3.0/`

Les travaux les plus anciens et les plus répandus, souvent inspirés de la conférence MUC-1996 (Nadeau & Sekine, 2007), concernent les EN dans le sens des *signifiants* (de Saussure, 1916), qui pourraient, de manière plus appropriée, être désignées comme entités nommantes. De très nombreuses approches de ce type sont généralement classées en des méthodes à base de règles et dictionnaires, à base d'apprentissage automatique et hybrides Des dictionnaires spécialisés de noms propres, employés notamment dans des systèmes de ce premier type, sont de taille et nature assez variées (Wolinski et al., 1995; Gaizauskas et al., 1995; Wacholder et al., 1997; Mikheev et al., 1999; Farmakiotou et al., 2000; Friburger & Maurel, 2004; Freitas et al., 2010; Maurel et al., 2011; Krstev et al., 2011). Les nouveaux défis de la REN consistent à reconnaître non seulement les entités les plus larges, mais aussi imbriquées (Alex et al., 2007; Ramírez-Cruz & Pons-Porrata, 2008; Finkel & Manning, 2009c; Nouvel et al., 2013; Dinarelli & Rosset, 2012), et les catégoriser selon une typologie étendue à des dizaines de catégories, comme ceci a eu lieu lors de la campagne d'évaluation en français ESTER-2 (Galliano et al., 2009). La difficulté particulière provient aussi du fait d'appliquer la REN à des textes bruités, e.g. oraux, comme dans la campagne ETAPE[4].

L'intérêt plus centré sur les *signifiés* est apparu avec le programme Automatic Content Extraction (ACE) (Doddington et al., 2004) et portait sur toutes les *mentions* possibles des entités dans le texte, ce qui impliquait notamment la *résolution de coréférence*. Plus récemment, la Text Analysis Conference[5] (TAC) a introduit la tâche de *entity linking*, qui consiste en le rattachement des entités nommés du texte à des noeuds d'une ontologie externe, puis en la clusterisation des entités n'ayant pas d'équivalent dans l'ontologie afin d'assurer son enrichissement. Dans le stade ultime de cette évolution du domaine le rattachement des entités du texte se fait vers les entrées des ressources du web sémantique – les Linked Open Data (Bizer et al., 2009; Mendes et al., 2012; Suchanek et al., 2007; Hoffart et al., 2011), telles que le DBpedia, qui rajoute une couche ontologique formelle au-dessus des ressources collaboratives libres telles que le Wikipédia, le GeoNames, etc. Il est à souligner que les systèmes existants qui réalisent une telle *désambiguïsation d'EN* (Hachey et al., 2013) prennent rarement en compte les langues à flexion riche, et plus particulièrement ceux à déclinaison, ce qui réduit considérablement la nécessité du traitement de la variabilité morphologique des EN (Rizzo et al., 2012; Daiber et al., 2013). L'état de l'art de l'annotation et de la reconnaissance d'EN dans une telle langue, le polonais, est résumé dans la section 4.2.2 (Piskorski, 2005; Abramowicz et al., 2006; Marcińczuk & Piasecki, 2007; Lubaszewski, 2007; Mykowiecka et al., 2008; Lubaszewski, 2009; Graliński et al., 2009b,a; Marcińczuk & Piasecki, 2010; Marcińczuk & Piasecki, 2011; Broda et al., 2012; Nothman et al., 2013; Marcińczuk et al., 2013).

Je me réfère également à une étude de l'état de l'art présentée dans (Savary et al., 2013b). Elle contient une analyse contrastive de ressources lexicales et sémantiques d'EN telles que alignements WordNet/Wikipedia (Toral et al., 2008, 2012; Fernando & Stevenson, 2012; Nguyen & Cao, 2010), YAGO (Suchanek et al., 2007) et YAGO2 (Hoffart et al., 2011), Freebase (Bollacker et al., 2007), MENTA (de Melo & Weikum, 2010), DBpedia[6] (Bizer et al., 2009; Mendes et al., 2012) et JRC-NAMES (Steinberger et al., 2011).

Dans les sections suivantes je décris mes contributions dans le domaine de la création de ressources et outils linguistiques du polonais, à commencer par la couche d'annotation des EN dans le Corpus National du Polonais[7] (pol. *Narodowy Korpus Języka Polskiego*; *NKJP*). Ce corpus de 1.5 milliards de mots, contient un sous-corpus équilibré de 300 millions de mots (Przepiórkowski et al., 2012), ainsi que son sous-ensemble annoté manuellement de 1 million de mots.

[4] `http://www.afcp-parole.org/etape.html`
[5] `http://www.nist.gov/tac/about/index.html`
[6] `http://dbpedia.org`
[7] `http://nkjp.pl/`

Le corpus est annoté à plusieurs niveaux: la segmentation, la morphosyntaxe, les mots et les groupes syntaxiques (chunks), les entités nommées et les sens de mots. La couche des EN, dont j'ai dirigé la réalisation, a été décrite dans plusieurs publications, où ont été évoqués : (i) le schéma et les choix méthodologiques d'annotation (Savary et al., 2010), (ii) la construction des dictionnaires et grammaires d'EN pour la pré-annotation automatique (Savary & Piskorski, 2010, 2011), (iii) les méthodes et les outils pour l'annotation manuelle et l'adjudication (Waszczuk et al., 2010), (iv) l'accord inter-annotateur et la construction d'outils à base d'apprentissage pour l'annotation du corpus entier de 1.5 milliards de mots (Waszczuk et al., 2013). La documentation du guide d'annotation et des cas intéressants rencontrés est poursuivie dans (Savary et al., 2012a). Finalement, dans (Savary & Waszczuk, 2012) nous approfondissons l'analyse des outils pour la pré-annotation, l'annotation manuelle et l'annotation automatique.

La figure 1.4 présente la typologie d'EN utilisée pour l'annotation du corpus. Elle est complétée par une typologie orthogonale contenant les adjectifs relatifs aux personnes, locations et organisations (*warszawski* 'varsovien'), ainsi que les dérivations personnelles, i.e. gentilés (e.g. *warszawiak* 'un varsovien') et les dénominations de membres d'organisations. Les attributs accompagnant chaque EN annotée incluent notamment: les formes de base (*Stany Zjednoczone* pour *Stanów Zjednoczonych* 'Etats Unis'), les bases sémantiques de dérivation (*Stany Zjednoczone* 'Etats Unis' pour *amerykański* 'américain'), et les normalisations des expressions temporelles (*09:45:00* pour *za piętnaście dziesiąta* 'dix heures moins le quart').

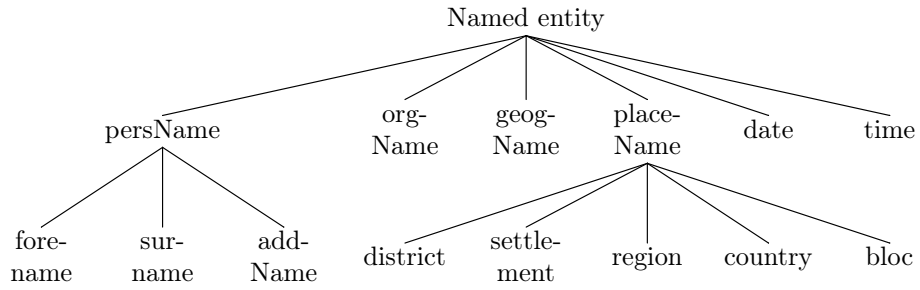


Figure 1.4: Hiérarchie des types d'EN utilisée dans le corpus polonais NKJP

La stratégie importante consiste à annoter non seulement les EN les plus larges, mais aussi toutes les EN imbriquées, comme dans les exemples (1.4)–(1.6).

(1.4) $[[Maria]_{\text{forename}}\ [Skłodowska]_{\text{surname}}\text{-}[Curie]_{\text{surname}}]_{\text{persName}}$

(1.5) $[ulica\ [[Mikołaja]_{\text{forename}}\ [Kopernika]_{\text{surname}}]_{\text{persName}}]_{\text{geogName}}$
RUE MIKOŁAJ$_{gen}$ KOPERNIK$_{gen}$
'rue Mikołaj Kopernik'

(1.6) $[[Wydział\ Prawa]_{\text{orgName}}\ [Uniwersytetu\ [Warszawskiego]_{\text{relAdj:settlement(Warszawa)}}]_{\text{orgName}}]_{\text{orgName}}$
FACULTÉ$_{nom}$ DROIT$_{gen}$ UNIVERSITÉ$_{gen}$ VARSOVIEN$_{gen}$
'Faculté de Droit de l'Université de Varsovie'

L'organigramme du processus de l'annotation, présenté dans le figure 1.5, inclut la pré-annotation automatique par la plateforme SProUT (Becker et al., 2002; Drożdżyński et al., 2004), qui offre : (i) un formalisme riche de grammaire de surface basé sur des outils à états finis, unification et cascades de règles, (ii) une consultation rapide de lexiques externes (gazetteers), (iii) une sortie XML dont les structures de traits utilisent une hiérarchie de types définie par l'utilisateur. Dans la section 4.4.1 nous décrivons l'adaptation et l'extension des lexiques et d'une grammaire polonaises pour la REN par SProUT, en vue de son adaptation à la pré-annotation

du corpus NKJP. Nous donnons également les résultats quantitatifs de la grammaire résultante et l'analyse de ses erreurs. Les résultats se résument en 3 caractéristiques :

- la précision et le rappel généraux varient de 68% à 78%, et de 35% à 39%, respectivement,
- les résultats sont, évidemment, meilleurs lorsque seulement les frontières, les types et les sous-types sont pris en compte que lorsque les autres attributs (lemmes, bases dérivationnelles, etc.) sont considérés; les différences entre ces deux scénarios d'évaluation varient de 2% à 13% de précision, et de 2% à 5% de rappel,
- les meilleurs résultats sont obtenus pour les expressions temporelles et les moins bons pour les noms d'organisations.

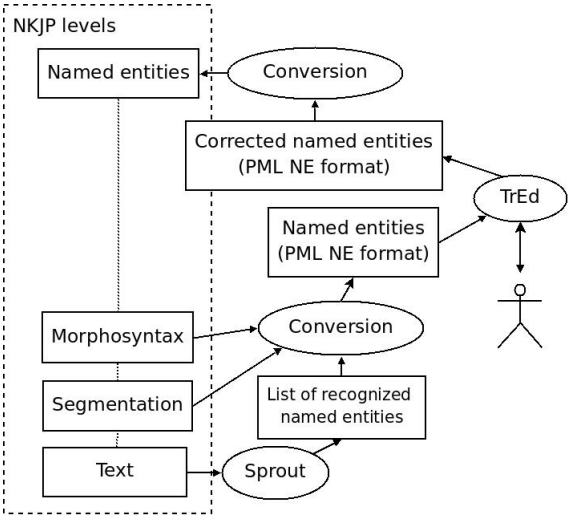


Figure 1.5: Flux de données dans l'annotation manuelle du sous-corpus NKJP de 1 million de mots

L'annotation manuelle, qui suit la pré-annotation (fig. 1.5), s'effectue via la plateforme TrEd[8] (Pajas & Štěpánek, 2008), adaptée aux besoins de NKJP par des macros, feuilles de style et raccourcis clavier. La figure 1.6 montre une copie d'écran de l'adjudication, effectuée par un annotateur expérimenté, suite à deux annotations indépendantes du même texte.

Des filtres adaptés assurent les conversions des formats des outils d'annotation entre eux, ainsi que vers le format final de NKJP (Przepiórkowski & Bański, 2009), qui est déporté (*stand-off*) et conforme au standard TEI P5 (Burnard & Bauman, 2008). La figure 1.7 montre une EN, contenant un adjectif relationnel, codée selon ce format.

Dans la section 4.3.3 nous décrivons les cas difficiles et les défis particuliers rencontrés lors de l'annotation. Ils concernent les phénomènes tels que:

- la coordination et le chevauchement des noms, en particulier noms de famille, dont un exemple est présenté dans la figure 1.7,
- les variantes elliptiques,
- les ambiguïtés d'imbrication,
- la métonymie et ses liens avec l'ellipse et l'imbrication,

[8] `http://ufal.mff.cuni.cz/~pajas/tred/`

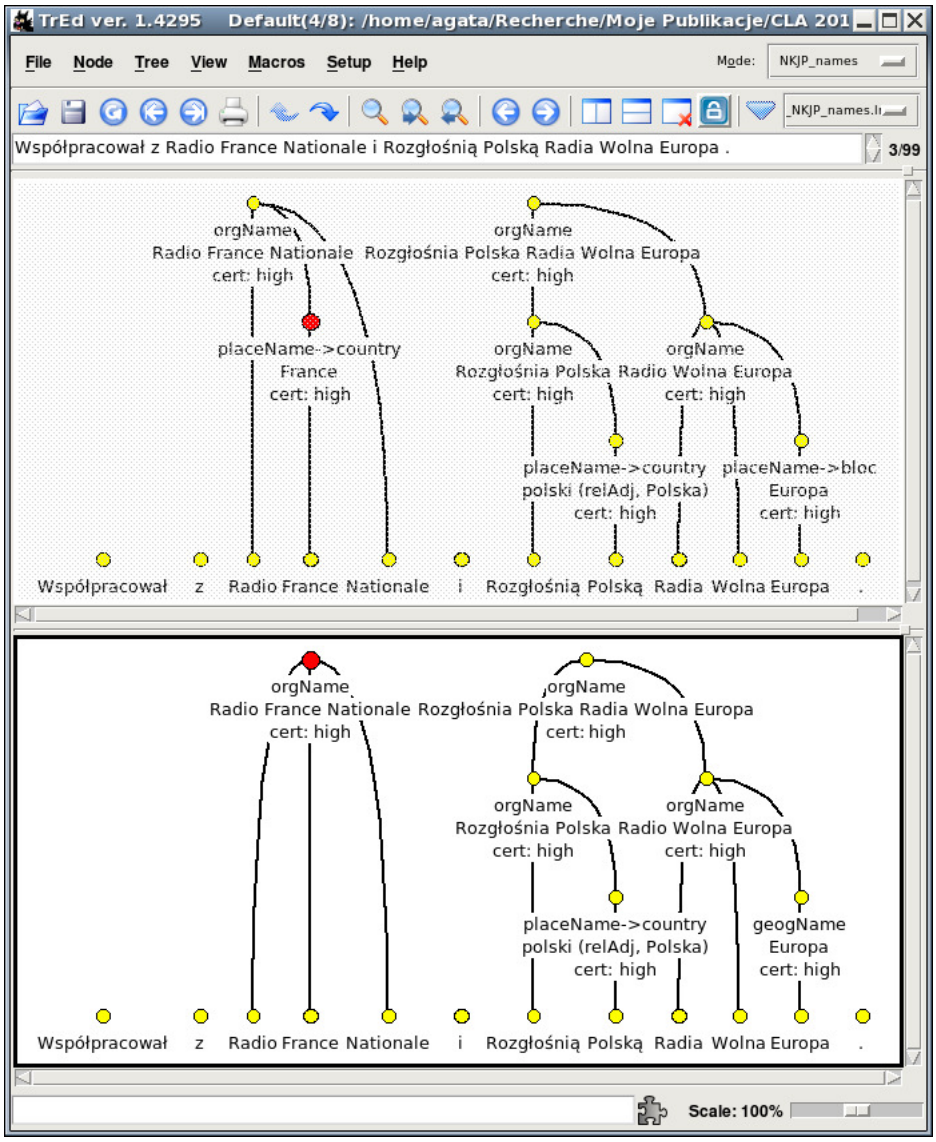
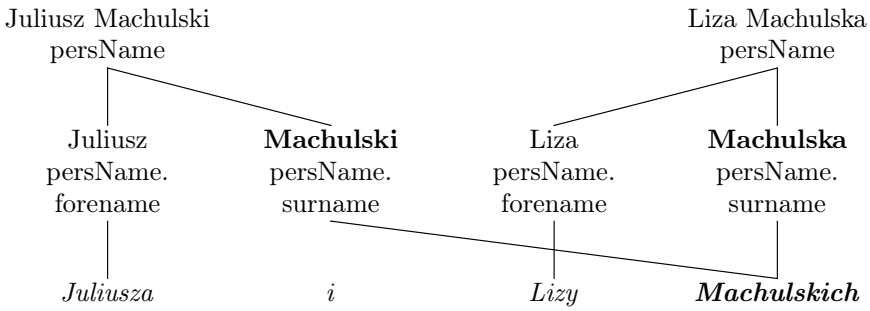


Figure 1.6: Adjudication dans TrEd pour la phrase avec des EN doublement imbriquées: 'Il a collaboré avec Radio France Nationale et la Station Polonaise de la Radio Europe Libre.'

- les ambiguïtés des bases dérivationnelles,
- les frontières gauches et droites incertaines.

(1.7)



JULIUSZ$_{sg:gen:m1}$ ET LIZA$_{sg:gen:f}$ MACHULSKI$_{pl:gen:m1}$
'de Juliusz et Liza Machulski'

```
<teiCorpus xmlns:xi="http://www.w3.org/2001/XInclude" xmlns="http://www.tei-c.org/ns/1.0">
<xi:include href="NKJP_1M_header.xml"/>
<TEI>
  <xi:include href="header.xml"/>
  <text><body>
  <p xml:id="named_1-p" corresp="ann_words.xml#words_1-p">
    <s xml:id="named_1.34-s" corresp="ann_words.xml#words_1.34-s">
      <seg xml:id="named_1.34-s_n2">
        <fs type="named">
          <f name="ne_type"><symbol value="orgName"/></f>
          <f name="orth"><string>Irlandzka Armia Republikańska</string></f>
          <f name="base"><string>Irlandzka Armia Republikańska</string></f>
          <f name="certainty"><symbol value="high"/></f>
        </fs>
        <ptr target="named_1.34-s_n3"/>                          <!-- Irlandzka -->
        <ptr target="ann_morphosyntax.xml#morph_1.2-seg"/>       <!-- Armia -->
          <ptr target="ann_morphosyntax.xml#morph_1.3-seg"/>     <!-- Republikańska -->
      </seg>
      <seg xml:id="named_1.34-s_n3">
        <fs type="named">
          <f name="derived">
            <fs type="derivation">
              <f name="derivType"><symbol value="relAdj"/></f>
              <f name="derivedFrom"><string>Irlandia</string></f>
            </fs>
          </f>
          <f name="ne_type"><symbol value="placeName"/></f>
          <f name="ne_subtype"><symbol value="country"/></f>
          <f name="orth"><string>Irlandzka</string></f>
          <f name="base"><string>irlandzki</string></f>
          <f name="certainty"><symbol value="high"/></f>
        </fs>
        <ptr target="ann_morphosyntax.xml#morph_1.1-seg"/>
      </seg>
    </s>
  </p>
  </body></text></TEI>
</teiCorpus>
```

Figure 1.7: Annotation au format TEI-P5 de l'EN *Irlandzka Armia Republikańska* 'Armée Républicaine Irlandaise'.

L'accord inter-annotateur des EN dans NKJP, tel que défini dans la section 4.3.4, varient entre 0.69 pour les noms d'organisation et 0.89 pour les noms de personnes.

Dans la suite du chapitre 4 nous décrivons notamment Nerf[9], un outil de REN employant l'apprentissage automatique à base des CRF, qui a été entraîné sur le corpus manuellement annoté et ensuite appliqué à l'annotation du corpus entier de 1.5 milliards de mots. Nerf, réalisé par Jakub Waszczuk, implémente la méthode d'annotation d'EN imbriquées nommée *joint label tagging* et introduite par Alex et al. (2007). Il obtient la précision générale de 0.83, le rappel de 0.76 et la F1-mesure de 0.79.

Ce même chapitre se poursuit par le résumé de nos travaux sur Prolexbase (Krstev et al.,

[9]Téléchargeable à `http://zil.ipipan.waw.pl/Nerf?action=AttachFile&do=view&target=nerf.dist.0.2.tgz`, sous licence GPL v3.

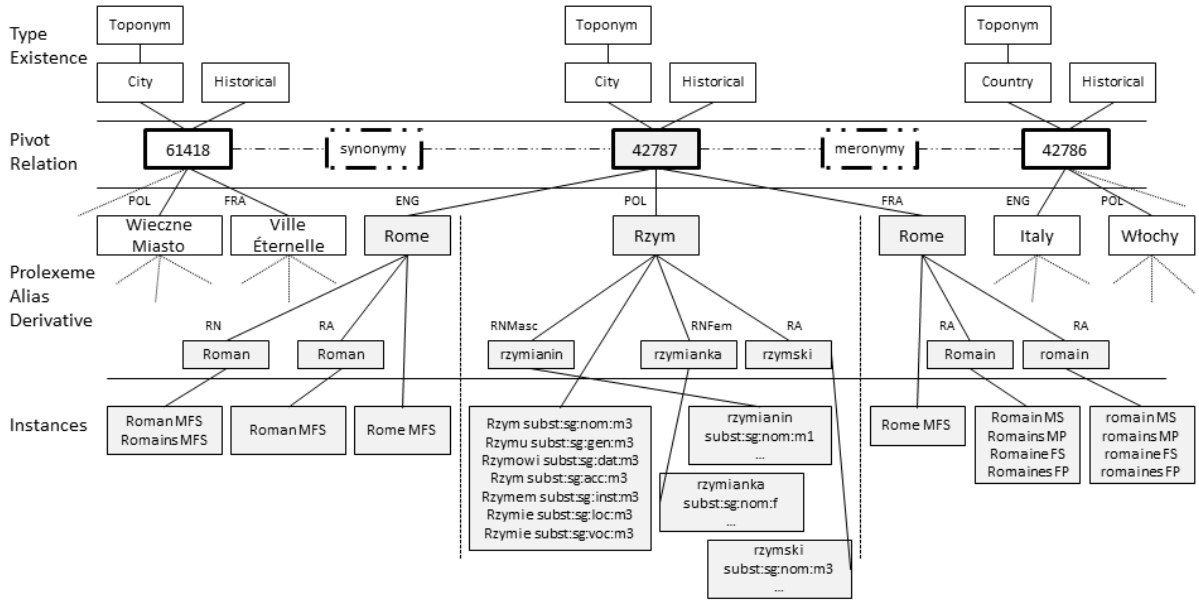


Figure 1.8: Extrait de Prolexbase avec quatre niveaux et trois lexèmes (appelés *prolexèmes*) en polonais, anglais et français.

2005; Tran & Maurel, 2006; Maurel, 2008), une base de données (ontologie au sens large) multilingue de noms propres, dont la richesse du modèle est illustrée par l'extrait de la figure 1.8. Nous proposons ProlexFeeder, un outil d'enrichissement semi-automatique de cette base à partir de ressources collaboratives libres en polonais, anglais et français: le Wikipédia et, à un moindre degré, le GeoNames, selon l'organigramme présenté dans la figure 1.9. Les enjeux majeurs de ce processus consistent en :

- L'alignement manuel des catégories de GeoNames et des types d'infoboxes du Wikipédia sur la typologie et les relations de Prolexbase. Par exemple la catégorie *Władcy Blois* 'comtes de Blois' est alignée avec le type *célébrité*, l'existence *historique*, la relation d'accessibilité avec le concept (appelé pivot) représentant la ville de Blois et le sujet *leader*.

- L'évaluation manuelle de la popularité des noms dans les 3 langues, basée sur la fréquence d'accès aux articles correspondants du Wikipédia.

- La prédiction de formes fléchies des noms polonais par les modules du système de traduction automatique Translatica (Jassem, 2004).

- La détection automatique des concepts déjà présents dans la base, pour éviter des doublons. Ceci a été réalisé par une fonction de similarité entre concepts basée sur leur lexèmes, variantes, types et liens URL. L'exactitude de la prédiction du bon pivot a atteint 97.2%.

- La correction et la validation manuelles des données extraites et pré-traitées automatiquement. Dans ce processus le traitement d'une entrée prenait 2 minutes en moyenne, la majeure partie de ce temps étant nécessaire à la correction de formes fléchies polonaises.

Le tableau 1.2 résume l'état de Prolexbase après la validation manuelle des données jugées les plus populaires. Une présentation plus détaillée de cette contribution est consultable dans (Savary et al., 2013a,b).

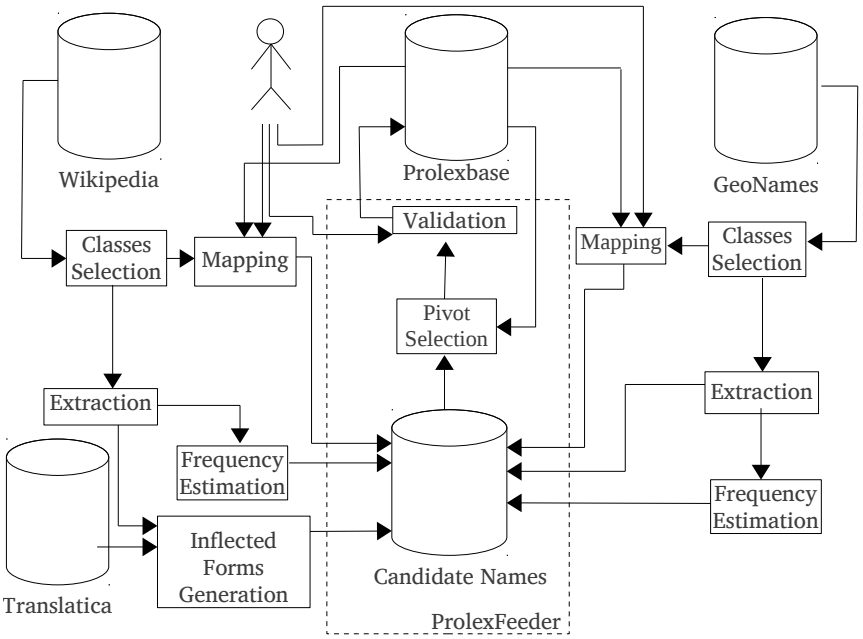


Figure 1.9: Organigramme de l'enrichissement de Prolexbase via ProlexFeeder.

Dans la dernière partie du chapitre 4 je présente mes travaux liés à l'annotation du Corpus Polonais de Coréférence (CPC)[10], qui complète le Corpus National du Polonais d'une nouvelle couche d'annotation. Avec ses 540,000 mots, la partie annotée manuellement du CPC est parmi les corpus les plus importants de ce type, avec Tüba/DZ (Hinrichs et al., 2005a) pour l'allemand, NAIST Text (Iida et al., 2007) pour le japonais, OntoNotes 2.0 (Pradhan et al., 2007) pour l'anglais, l'arabe et le chinois, le Prague Dependency Treebank (Nedoluzhko et al., 2009) pour le Tchèque et ANCOR (Muzerelle et al., 2013) pour le français.

L'annotation manuelle, précédée par la pré-annotation automatique, s'effectue à l'aide d'une version adaptée de MMAX2 (Müller & Strube, 2006). Elle est suivie de la révision des annotations par un deuxième annotateur. Une partie du corpus annotée par deux annotateurs en parallèle et révisée par un troisième expert, a permis le calcul de l'accord inter-annotateur.

Dans (Ogrodniczuk et al., 2013a), nous présentons les aspects majeurs de la portée et du schéma d'annotation, qui couvrent tous les groupes nominaux et pronominaux (incluant éventuellement des phrases relatives, parfois très complexes) et leurs ellipses. Contrairement à certaines approches, nous ne marquons pas de coréférence entre les mentions dont l'identité est identifiable grâce à la syntaxe, comme les appositions (*Jean Villain, père de 4 enfants*) et les prédicats (*Jean est un père*). En plus de la relation d'identité entre référents, nous incluons, à titre expérimental, la relation de la *presque-identité* (near-identity) proposée par Recasens et al. (2011). Comme évoqué dans (Ogrodniczuk et al., 2013b), cette relation est annotée dans notre corpus avec un accord inter-annotateur très faible, ce qui plaide pour sa complexité. Deux autres caractéristiques originales de notre schéma d'annotation consistent à : (i) indiquer la mention dominante, i.e. celle parmi les membres d'un cluster d'identité qui décrit le référent de la manière la plus précise, (ii) marquer les têtes sémantiques (qui se distinguent des têtes syntaxiques notamment dans les expressions numérales : *pięć kobiet* '⌈cinq femmes$_{pl:gen:f}$⌋$_{pl:nom:n}$').

Les défis particuliers dans la tâche de l'annotation sont liés aux mentions imbriquées, coordonnées et chevauchantes, qui exigent parfois la multiplication importante de mentions. Le

[10] `http://zil.ipipan.waw.pl/PolishCoreferenceCorpus`

Table 1.2: Etat actuel de Prolexbase. Les formes (instances) polonaises incluent seulement les formes fléchies des prolexèmes (et non pas de leurs variantes, appelées alias).

Pivots				
Tous	Toponymes	Anthroponymes	Ergonymes	Pragmonymes
73,405	81.3%	16.8%	1.4%	0.4%

Relations			
Toutes	Méronymie	Accessibilité	Synonymie
72,672	92.9%	5.3%	1.8%

	Pivots en relation de synonymie			**Pivots en relation de méronymie**			**Pivots en relation d'accessibilité**		
Tous	2,457		(3%)	65,768		(90%)	6,312		(9%)
Types	célébrité	1,325	(17%)	ville	48,110	(100%)	ville	2,214	(5%)
les plus	pays	390	(45%)	célébrité	7,053	(88%)	région	1,696	(40%)
fréquents	ville	157	(0.3%)	région	4,052	(97%)	célébrité	1,129	(14%)

Langue	**Prolexèmes**	**Alias**	**Dérivés**	**Instances**
PL	27,408	8,724	3,083	166,479
EN	19,492	14,039	94	18,575
FR	70,869	8,488	20,919	142,506

corpus dans son état final, d'après Ogrodniczuk et al. (2013c), contient plus de 180.000 mentions, 5.000 liens de presque-identité, 109.000 cluster uni-mention et près de 19.000 clusters contenant au moins deux mentions. Le corpus est distribué sous licence Creative Commons CC BY 3.0[11] et il est visualisable en ligne[12].

1.4 Méthodes à états finis pour les langages de mots et d'arbres

Les langages formels de mots (chaînes de caractères) et d'arbres sont un intérêt central en informatique, et ils sont souvent considérés en TAL comme approximations de langues naturelles. C'est pourquoi le chapitre 5 est consacré à mes contributions à ce domaine.

Je présente d'abord l'état de l'art de l'utilisation des méthodes à états finis en TAL à travers: (i) les expressions régulières (Justeson & Katz, 1995), (ii) les transducteurs à états finis (Kaplan & Kay, 1994; Laporte, 1997; Koskenniemi, 1983; Beesley & Karttunen, 2003; Roche & Schabes, 1997; Roche, 1997) et les cascades de transducteurs (Abney, 1996; Hobbs et al., 1997; Friburger & Maurel, 2001). Je fais ensuite référence au problème de recherche approximative de motifs (*approximate string matching*) (Hall & Dowling, 1980) basée sur les opérations élémentaires sur des lettres telles que l'insertion, la suppression et le remplacement d'une lettre, ou l'inversion de deux lettres adjacentes. La *distance d'édition* entre mots est ensuite définie comme le coût minimal d'une séquence d'opérations élémentaires transformant l'un des mots vers l'autre. Ce problème possède deux variantes majeures: la comparaison de mots (*string-to-string correction*) (Damerau, 1964; Levenshtein, 1966; Wagner & Fisher, 1974; Lowrance & Wagner, 1975; Du & Chang, 1992) et la correction d'un mot par rapport à un langage de mots (*string-to-language correction*). Une étude comparative de l'état de l'art dans ce dernier domaine, incluant une taxonomie de méthodes, leur implantation et évaluation dans un cadre commun, a été proposée par Boytsov (2011). Nous nous intéressons plus particulièrement à l'algorithme de Oflazer

[11] `http://creativecommons.org/licenses/by/3.0/deed.en_US`
[12] `http://glass.ipipan.waw.pl:11111/index.xhtml#/core/`

(1996), qui se sert de la représentation du langage sous forme d'automate à états finis (*finite-state automaton*, FSA). Il effectue le calcul de la distance d'édition en parcourant le FSA en profondeur et en maintenant une matrice d'édition dont les lignes correspondent aux caractères du mot corrigé et les colonnes aux transitions du FSA. Chaque fraction de la matrice est calculée une seule fois pour tous les mots ayant un préfixe commun. Une variante de cette méthode, proposée dans (Savary, 2001b), change l'ordre du parcours du FSA en poursuivant d'abord le plus long préfixe correct du mot à corriger.

Une extension du problème de la correction de mots est celui de la correction d'arbres. Des opérations élémentaires sur un arbre peuvent être assez variées et incluent généralement l'insertion ou la suppression d'un noeud (interne ou feuille) et le renommage d'un noeud. La distance entre deux arbres est définie comme le coût de la séquence minimale contenant de telles opérations. Ici également deux instances du problème existent : la comparaison d'arbres (*tree-to-tree correction*) (Selkow, 1977; Tai, 1979; Zhang & Shasha, 1989; David Barnard and Gwen Clarke and Nicholas Duncan, 1995) et la correction d'un arbre par rapport à un langage d'arbres (*tree-to-language correction*) (Bertino et al., 2004; Boobna & de Rougemont, 2004; Xing et al., 2006; Staworko & Chomicki, 2006; Tekli et al., 2007; Suzuki, 2007; Bertino et al., 2008; Staworko et al., 2008; Thomo et al., 2008; Svoboda, 2010; Svoboda & Mlýnková, 2011; Tekli et al., 2011).

Notre contribution principale liée au outils à états finis concerne ce denier domaine. Nous avons proposé une méthode de correction d'un document XML (vu comme arbre) par rapport à une DTD, qui étend deux algorithmes précédents : celui d'Oflazer (1996) pour la correction de mots par rapport à un FSA, et celui de Selkow (1977) pour la comparaison de deux arbres. L'idée générale peut être résumée par quelques principes fondamentaux :

- Les données du problème sont : l'arbre XML à corriger t, la DTD sous forme d'un schéma S, le seuil de correction th, et l'étiquette souhaitée c pour la racine de l'arbre corrigé.
- Le résultat incluent : (i) la liste de tous les arbres ayant la racine étiquetée par c et valides par rapport à S, dont la distance par rapport à t ne dépasse pas th, (ii) toutes les séquences d'édition possibles transformant t en un des arbres résultants, (iii) les coûts de ces séquences. Par exemple, pour l'arbre de la figure 1.10 et la DTD de la figure 1.11, les arbres corrigés résultants sont démontrés dans la figure 1.12, et leurs séquences d'édition correspondantes sont les suivantes : $\{\langle(relabel, 0, b), (delete, 0.1, /)\rangle, \langle(add, 3, c)\rangle, \langle(relabel, 2, c), (delete, 2.0, /)\rangle\}$.
- Les contraintes de la structure d'un document XML sont exprimées dans une DTD via des expressions régulières attribuées à des étiquettes. Afin qu'un document XML soit valide, il faut que, pour chaque noeud n, le mot formé par les étiquettes des fils de n soit inclue dans le langage décrit par l'expression régulière attribuée à l'étiquette de n.
- Chaque expression régulière présente dans une DTD est représentée sous forme d'un FSA parcouru selon les principes de l'algorithme d'Oflazer (1996).
- Lors de ce parcours, lorsque le renommage d'un noeud est supposé, il est nécessaire de considérer le changement potentiel de tout le sous-arbre attaché à ce noeud. Ceci implique la correction récursive, basée sur la distance entre arbres définie par Selkow (1977).

Notre algorithme est un résultat d'un travail de longue haleine, depuis sa conception et implantation dans un cadre incrémental (Cheriat et al., 2005; Bouchou et al., 2006b,a) jusqu'à sa redéfinition plus fondamentale, sa ré-implantation, et validation théorique et expérimentale. La publication majeure (Amavi et al., 2013) rassemble tous ses résultats finaux:

- Les définitions formelles des objets manipulés (un arbre XML, un schéma, un sous-arbre, une arbre partiel, un langage d'arbres) et leurs propriétés (validité, validité locale, validité

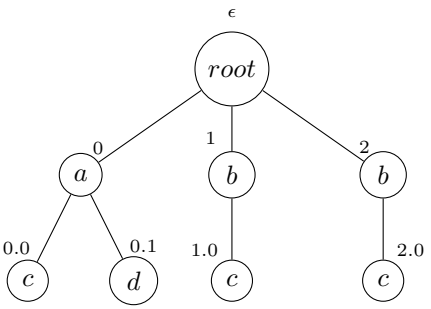


Figure 1.10: Un arbre XML à corriger.

Étiquette	Expression régulière
root	$b^*\|ab^*c$
a	cd
b	c
c	ϵ
d	ϵ

Figure 1.11: Une DTD

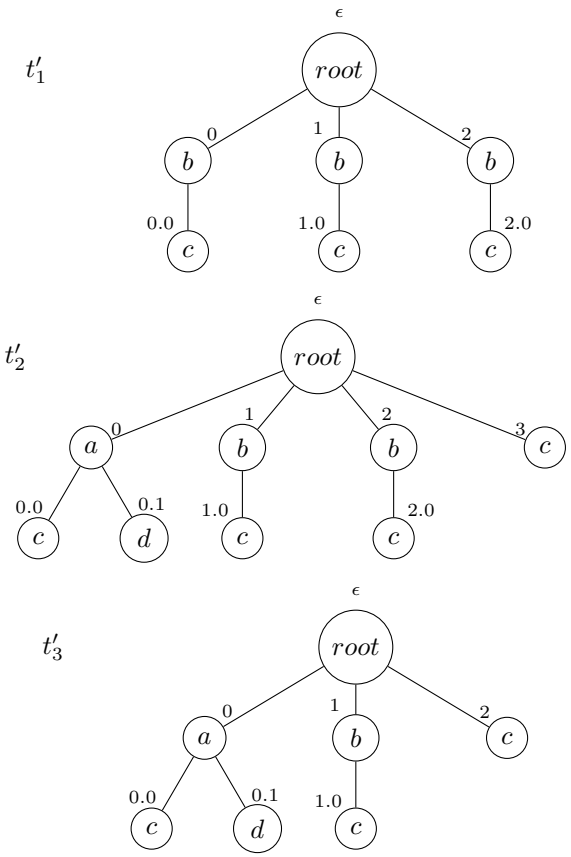


Figure 1.12: Trois corrections t'_1, t'_2 et t'_3 pour l'arbre t de la figure 1.10.

partielle), les opérations sur des noeuds (renommage, addition et suppression) et sur des sous-arbres (insertion et élimination), les séquences d'opérations, leur équivalence et leurs coûts.

- Les preuves de la terminaison, de la correction et de la complétude de l'algorithme.
- L'analyse de la complexité en temps qui est en $O((f_t+1)\times(f_S)^{|t|+th}\times 6\times|\Sigma|\times(|t|+th))^{th}$, où f_t signifie le fan-out maximum de t (le nombre maximum d'enfants d'un noeud dans t), f_S est le fan-out maximum des états dans le FSA du schéma S, $|t|$ est la taille de t (le nombre de ses noeuds) et $|\Sigma|$ est la taille de l'alphabet du schéma S.

- Les résultats des expériences effectuées sur un fichier du Corpus National du Polonais contenant des annotations d'entités nommées (cf. section 1.3). Ces résultats, obtenus suite à la variation des différents paramètres du problème (la taille du document, la valeur du seuil, le nombre et les positions des erreurs, la nature de la DTD), démontrent un comportement polynomial de l'algorithme malgré sa complexité théorique exponentielle.

- Une étude contrastive de l'état de l'art, qui prend en compte le choix des opérations élémentaires, les aspects de validité considérés (le bien formé, la validité de structure, et des attributs), les résultats produits (la distance d'édition, les corrections minimales ou dans un seuil, les séquences d'édition), le type du schéma (une DTD, un XML schéma, une DTD étendue) et son modèle (automate d'arbre, ensemble d'expressions régulières, hedge automaton, arbre ordonné, automate à pile, etc.), le modèle du document XML (un arbre, un mot d'étiquettes ouvrantes et fermantes), la complexité en temps et espace, existence des preuves, la nature et la disponibilité des données expérimentales, la disponibilité des implantations et des codes sources.

A la lumière de ce dernier élément, il apparaît que notre contribution est la première solution relativement complète du problème de la correction d'un arbre par rapport à un langage d'arbres. Non seulement nous calculons la distance d'édition entre un document et un schéma, mais nous fournissons également tous les arbres corrigés résultants, sans nous limiter aux solutions les plus proches de l'arbre initial. Ainsi, nous considérons qu'il s'agit d'un problème d'énumération plutôt que de décision, contrairement à ce qui a lieu dans beaucoup d'autres approches. Notre documentation est l'un des rares cas où les preuves de complexité, de correction et de complétude sont fournis. C'est aussi la seule contribution qui rend disponibles non seulement les exécutables et les sources, mais aussi le guide d'utilisateur et les données expérimentales. Par conséquence, il semble que c'est la seule approche reproductible. Finalement, nos codes sources sont les seuls à être distribués[13] sous une licence connue : la licence ouverte GNU LGPL v3.

Dans la suite du chapitre 5 j'évoque mes autres contributions à l'algorithmique des états finis, centrés sur le problème de la dynamicité des données, ce qui requiert des solutions *incrémentales*. Il s'agit premièrement de la validation et de la correction incrémentales d'un document XML par rapport à une DTD (Cheriat et al., 2005; Bouchou et al., 2006b,a), qui a motivé nos premiers travaux vers l'algorithme de correction décrit plus haut. Deuxièmement, nous avons proposé des solutions de construction incrémentale et pseudo-incrémentale d'automates pseudo-minimaux (Daciuk et al., 2005b). Une construction incrémentale minimise la partie de l'automate touchée par l'ajout d'un nouveau mot, ce qui est crucial notamment dans des applications en TAL où le vocabulaire varie fréquemment (e.g. en recherche d'information). Un automate pseudo-minimal possède une transition ou un état propre de chaque mot représenté (i.e. une transition/état appartenant seulement à ce mot). Cet élément propre peut être utilisé pour encoder des données spécifiques à un mot, par exemple sa valeur d'une fonction de hachage. Dans (Daciuk et al., 2005a) nous proposons des algorithmes de hachage parfait dynamique, i.e. tel que chaque mot du langage obtient une valeur unique et le rajout de nouveaux mot ne change pas la valeur de hachage des mots précédents.

1.5 Le cadre de travail et la direction de recherche

Cette dissertation est censée valider ma capacité à encadrer des travaux de recherche. C'est pourquoi je dédie le chapitre 6 à la description de mon expérience en la matière de :

[13] `http://www.info.univ-tours.fr/~savary/English/xmlcorrector.html`

- Collaborations extérieures internationales (en Pologne: Institut d'Informatique de l'Académie Polonaise des Science, IPIPAN, Varsovie; Université de Gdańsk, de Poznań et de Olsztyn; en Serbie: Université de Belgrade; en Russie: Université d'Etat de Tomsk), nationales (Université Paris-Est Marne-la-Vallée) et régionales (Université d'Orléans).

- Bibliométrie des mes 41 publications (depuis la thèse de doctorat), dont 9 articles dans des journaux internationaux à comité de lecture.

- Développement de logiciels (Multiflex et XMLCorrector).

- Montage de projets de recherche (1 projet COST, un projet Européen, un projet PHC EGIDE et une ANR), ainsi que participation aux projets en tant que leader de tâches (1 projet FEDER et un projet national), collaborateur (un projet PHC EGIDE, 3 projets nationaux, 2 projets régionaux) ou sous-traitant (1 projet Européen et un national).

- Encadrement de recherche (3 thèses de doctorat, 3 thèses de master).

- Évaluation de recherche en tant que : membre de comités scientifiques (1 revue, 3 numéros spéciaux de revue, 11 conférences et workshops), expert européen (évaluateur et reviewer), membre nommé de la section 27 du Conseil National des Universités, et membre de 3 jury de thèse.

- Organisation d'évènements, en tant que présidente du comité d'organisation de la conférence internationale CIAA-FSMNLP-2011 à Blois.

- Enseignement universitaire et relations internationales à l'IUT de Blois.

Concernant le montage et la gestion de projet, mon expérience principale concerne l'action COST[14] IC1207 PARSEME (PARsing and Multi-word Expressions)[15], dont j'ai été rédactrice de proposition et que je coordonne actuellement en tant que présidente du Comité de Gestion (Management Committee). Cette initiative rassemble une centaine de chercheurs de 28 pays majoritairement européens autour de quatre groupes de travail: (i) interface lexique/grammaire, (ii) analyse syntaxique symbolique des UP, (iii) parsing hybride des UP, (iv) annotation des UP dans des corpus arborés. Les activités financées par COST portent sur le fonctionnement collaboratif de ce réseau (réunions, ateliers, écoles d'été, missions courtes, dissémination, etc.).

1.6 Conclusions et perspectives

Ma dissertation se termine par les conclusions générales et les perspectives (chapitre 7). Ces dernières incluent:

- L'amélioration et l'extension des ressources et outils TAL existants, tels que Multiflex, Nerf et le corpus NKJP.

- L'intégration des ressources linguistiques fines dans les Linked Open Data, ainsi que le rapprochement du TAL, et notamment des acquis de la REN avec le web sémantique, dans le contexte de la désambiguïsation d'EN.

- Le parsing syntaxique des unités polylexicales, avec les défis définis dans le cadre de l'action COST PARSEME.

[14] `http://www.cost.eu/`, financé par European Science Foundation
[15] `http://www.parseme.eu`, `http://www.cost.eu/domains_actions/ict/Actions/IC1207`

- L'identification des UP dans des corpus arborés modélisée en tant que correction d'un arbre (un sous-arbre syntaxique extrait du corpus) par rapport à un langage d'arbres (l'ensemble de sous-arbres syntaxiques représentant une UP et ses variantes).
- Une taxonomie d'algorithmes de correction d'un arbre par rapport à un langage d'arbres, leur implantation et évaluation expérimentale dans un cadre commun.

Chapter 2

Composition and Variation – an Introduction

A large part of this thesis addresses some types of linguistic units which result from the composition of linguistic items and whose inherent properties are those of linguistic (orthographic, morphological, syntactic and semantic) variability.

Composing (or combining) linguistic items yields larger linguistic items (usually containing several words) whose central property is to be or not to be **compositional**. Let us briefly refer to some works in the domain of the philosophy and mathematics of the language that address the compositionality principle. According to Pagin & Westerståhl (2001a), compositionality is a key notion in linguistics, philosophy of language, logic, and computer science, but there are divergent views about its exact formulation, methodological status, and empirical significance. Many seminal contributions to this notion are attributed to Frege (Janssen, 2001), even if his idea of contextuality (a word has no meaning in isolation, but only in the context of a sentence) seems contradictory to his views on compositionality (we construct the sense of a sentence from the sense of its parts). As stressed by Kracht (2007), compositionality has not been thoroughly studied until the early 2000s. The generally admitted definition, after (Partee et al., 1990), is that *a compound expression is compositional if its meaning is a function of the meanings of its parts and of the syntactic rule by which they are combined.* Kracht points out that this definition is superficial in that (surface) expressions and their parts are usually ambiguous and that a meaning can only be assigned to *their analyses.* Consequently, compositionality is primarily a property of a grammar, and **a language is compositional if it has a compositional grammar**. Kracht also mentions that, in the literature, one analysis is often considered superior to another one on the grounds that it is compositional. He argues though that proving compositionality is hard due to the lack of standards as to the boundary between the syntax and the semantics.

Baggio et al. (2012) remind and refine the following reasons for **promoting compositionality** in linguistic analyses: (i) productivity (there are infinitely many sentences in any natural language, but the brain has only finite storage capacity), (ii) systematicity (the ability to understand certain utterances is connected to the ability to understand certain others), (iii) methodology (compositionality underlies the method for semantic calculus), (iv) modularity (information encapsulation at the level of the description of linguistic structure). They also argue that compositionality may imply a very large amount of rules dedicated to particular word combinations, thus it is an issue of balance between storage and computation: compositionality can often be rescued by increasing the demand on (brain) storage, whereas it must be abandoned under realistic constraints on storage.

It appears, however, that compositionality of a natural language is far from evident (or proven). The arguments against compositionality, as summarized by Pagin & Westerståhl

Table 2.1: Sample emotion predicate classes in Emologus

Class	**Example**	**Valency modification function**
Conserving	*aider* 'help'	$\forall_x Val(pred(x)) = Val(x)$
Inverting	*casser* 'break'	$\forall_x Val(pred(x)) = -Val(v)$
Positive shift	*mignon* 'cute'	$\forall_x Val(pred(x)) = max(Val(x) + 1, 2)$
Negative shift	*énérvé* 'stressed'	$\forall_x Val(pred(x)) = min(Val(x) - 1, -2)$
Minimum	*embrasser* 'kiss'	$\forall_{x,y} Val(pred(x, y)) = min(Val(x), Val(y))$
Multiplicative	*avoir* 'have'	$\forall_{x,y} Val(pred(x, y)) = Val(x) \times Val(y)$
Positive	*caliner* 'cuddle'	$\forall_{x,y} Val(pred(x, y)) = 2$
Negative	*dégoûter* 'disgust'	$\forall_{x,y} Val(pred(x, y)) = -2$
...	...	...

(2001b), include its vacuity, triviality, and superfluity, as well as – what is of major interest for this thesis – the fact that certain constructions are **counterexamples** which make the compositionality principle false. These problematic cases comprise belief sentences and quotations (both challenge the principle of substitutability of synonyms) as well as **idioms**. For instance, the meaning of the idiom *to kick the bucket* (i.e. to die) cannot be obtained by the same process as the one of interpreting the syntactically similar expression *to fetch the bucket*. The authors argue, however, that there are ways to incorporate idioms while preserving compositionality, in that different compositionality rules apply to idioms than to "regular" phrases.

In this thesis, I deal notably with Multi-Word Expressions (MWEs), which are larger classes than idioms but which are frequently defined under the premises of their non compositionality or *atypical* compositionality.

2.1 Compositionality of Emotion Expression

The hypothesis of linguistic compositionality, provided that it can be experimentally supported, is convenient for modeling and computation since it prevents a combinatorial explosion of lexicalized cases. As an example, let us consider the problem of emotion expression in linguistic utterances and its automatic detection and characterization.

In (Tallec et al., 2009) and (Tallec et al., 2010b) we present the EmotiRob project aiming at a prototype of an emotional companion robot for weakened children. One of its projected features is facial expression of simulated emotions as a reaction to an interaction with a child. Contrary to many other approaches in emotion detection, we assumed that polarity, also called valency (negative/positive/neuter) and intensity (moderate/strong) of an emotion conveyed by an utterance can be deduced from its propositional content, rather than from prosody only. We validated this hypothesis within *Emologus*, a spoken language understanding system, which proceeds in three steps: (i) chunking, (ii) building semantic relations between chunks (roughly, dependency parsing), (iii) contextual interpretation. The vocabulary of this prototype system is restricted to about 1,000 words from a corpus of child-invented tales collected in a primary school.

We admitted that emotion calculation is compositional: (i) basic lexical items have an atomic emotional value, included in the interval $[-2; 2]$, (ii) predicates can modify the emotional values of their arguments. Atomic emotional values were provided by psycholinguistic studies in children of ages 5 to 7. Emotion functions of predicates were determined by 5 adult annotators. Table 2.1 shows sample unary and binary predicate classes and their corresponding valency modification functions.

Given the atomic emotional values of lexical words and emotion predicate classes, the cal-

culation of the emotion associated to an utterance is performed compositionally. Consider the sentence in example (2.1). As a result of parsing in Emologus, the formula in example (2.2) is produced. Words *cochon* 'pig' and *ami* 'friend' have atomic emotional values 0 and 1, respectively. The unary predicate *petit* 'little' belongs to the positive shift class, i.e. composed with its emotionally neutral argument, it yields the emotional value 1. The binary predicate *avoir* 'have' yields a multiplication of the emotional values of *un petit cochon* 'a small piglet' and *amis* 'friends', which results in value 1. Finally, the unary operator *pas* 'not' inverts the value of its argument. As a bottom line, the emotional value of the whole sentence is -1.

(2.1) *Il etait une fois un petit cochon qui n'avait pas d'amis.*
Once upon a time, there was a little piglet who had no friends.

(2.2) (narrative (neg (to have [(subject: (pig [(size: little)])), (object: (friends))]))

In-domain evaluation (Tallec et al., 2010a) has shown that Emologus obtains a 90% accuracy in detecting the emotional value of an utterance. It significantly outperforms the baseline bag-of-words approach, which consists roughly in summing up the elementary emotional values of the words appearing in a given sentence, and which obtains a 68.8% accuracy on the same corpus. An error analysis shows that Emologus never assigns an emotional value whose valency is opposite to the expected one. Note, however, that the sub-language studied in EmotiRob is restricted to a domain with almost inexistent language resources, and with a relatively short vocabulary containing few compounds and multi-word expressions. A large-scale validation would be needed in order to study the influence of such non-compositional phenomena on the performances of the compositional emotion detection.

A validation of an approach similar to ours in the related domain of attitude (affect, judgment and appreciation) detection in adults is presented by Neviarouskaya et al. (2010). Here, the attitude detection operates on: (i) affect categories (anger, guilt, joy, etc.), (ii) polarity (positive, negative, neuter), (iii) intensity (between 0 and 1), and (iv) confidence level. A core lexicon of attitude-conveying terms (*unfriendly*, *desire*, etc.) is annotated with affect category, polarity and intensity. A closed list of modifiers and functional words (*slightly*, *hardly*, *never*, *without*, *increase*, etc.) is assigned attitude modification operators, similarly to predicates in our approach. Modal operators (*arguably*) are attributed the related confidence values. Verbs are classified with respect to their influence on attitude conveyed by a sentence (e.g. *to defend* belongs to the 'preservation' class). Finally, compositional attitude calculus is based on rules of polarity reversal, aggregation, propagation, domination, neutralization, and intensification, at various grammatical levels (similar to our valency modification rules). These rules are applied to the output of dependency parsing. An evaluation on a 1000-sentence manually annotated corpus shows the overall top-level (when polarity only is accounted for) accuracy of 0.879. These results, comparable to Emologus performances, confirm that a compositional rule-based calculus of emotion/attitude can yield relatively reliable results. Interestingly enough, some studies show that even semantically opaque linguistic units such as Multi-Word Expression (MWEs), to which the majority of this thesis is dedicated, show a relatively high degree of compositionality with respect to their emotional profile (Klebanov et al., 2013).

2.2 Compositionality of Multi-Word Expressions

The compositionality issues lie at the heart of linguistic debates since several decades, notably with respect to units crossing words boundaries, which are generally designated as **Multi-Word Expressions** (**MWEs**). They include a wide range of heterogeneous objects such as compounds,

complex terms, multi-word named entities, light verbs, idioms, etc. I define this notion more precisely in Chapter 3.

For instance, a rich discussion concerning the frontiers of the **nominal composition** (Habert & Jacquemin, 1993) took place at the end of the past century. Some linguists stated that nominal compounds result from the application of the compositionality principle to nominal phrases (Downing, 1977; Fabre & Sébillot, 1996) while others, conversely, view nominal compounds as semantically or referentially non-compositional structures (Benveniste, 1974; Lyons, 1978).

The idea of compositionality of MWEs can be extended to other areas than the semantics alone. Mel'čuk (2010) defines the semantic and **morphosyntactic compositionality** of the linguistic signs, where a sign is composed of a *signifié* (meaning), a *signifiant* (a string of phonemes or characters) and morphosyntactic properties (part of speech, inflectional features, etc.). A complex linguistic sign is compositional if both its *signifié* and its morphosyntactic properties result from a straightforward (proper to its syntactic structure) combination of those of their components. Thus, compositionality is a binary property, it cannot be partial.

In Savary et al. (2007) I address notably the **inflectional compositionality** and non-compositionality of compounds in French, Polish and Serbian (cf. Section 3.4). Compounds are said to be inflectionally compositional if their inflectional properties can be fully deduced from the properties of their respective constituents and of their syntactic structure. For instance the regular plural formation of *Noun-Noun* compounds in English consists in putting their final nouns in the plural form. Compound (2.3) is compositional in this sense while (2.4) is not.

(2.3) *chief justice, chief justices*

(2.4) *lord justice, lord justices, lords justice, lords justices*

In English, such examples belong to a closed list and are of relatively little quantitative importance. Since French presents a richer inflectional morphology, inflectional irregularities within compounds are frequent. For instance, the class of French *Verb-Noun*-type compounds contains numerous examples in which the gender and number of the whole structure cannot be deduced from those of its constituents. For instance the French compound:

(2.5) *un perce-neige* 'a snowdrop'

is masculine although the noun *neige* is feminine. Here again, while *Verb-Noun* composition is productive in French, the resulting compounds remain inflectionally non-compositional.

In Slavic languages, the difficulties with the inflection of compounds may be even more important due to declension and a complex gender, number and animateness cross-dependencies within nouns and adjectives. For instance (Czerepowicka & Kosek, 2011), the *Adj-Noun* compound in example (2.6) is in masculine human gender although its nominal component *pająk* 'spider' has masculine animate gender. Thus, this compound is said to be exocentric since it contains no headword from which its gender could be deduced.

(2.6) *czerwony pająk* 'lit. a red spider = ex-communist'

The semantic or inflectional non-compositionality of compounds is closely connected to the idea of **lexicalization**: if an expression has a meaning, a reference or inflectional properties that are not totally deducible from its components, this expression is lexicalized, i.e. has to be explicitly mentioned and described in a lexicon in order for it to be processed appropriately. In Section 3.3 and 3.6 I describe my contributions to the lexical description of contiguous multi-word expressions, including a formalism and a tool meant for taking their morphosyntactic idiosyncrasies into account.

2.3 Linguistic Variability — Central Challenge in NLP

Linguistic debates on operational definitions that allow to distinguish WMEs from the regular phrases frequently refer to the idea of "frozenness", i.e. the fact of blocking the linguistic transformations that are usually allowed for a syntactic structure under study. For instance, if components of the expressions *cross-roads* or *to kick the bucket* are replaced by their synonyms, as in *cross-routes* and *to hit the container*, the idiomatic sense is lost.

While keeping in mind this inflexibility of MWEs, one should not underestimate their remaining degree of variability: some "regular" transformations are prohibited in a MWE but some others are allowed. On the basis of this observation, Gross (1988) introduces the idea of a *degree of frozenness* in nominal compounds: the more transformations typical for a certain syntactic structure are blocked in a nominal compound having this structure the more this compound is frozen. He further shows (Gross, 1990) how this degree can be handled operationally within the *lexicon-grammar*[1] approach. Note that this idea of a partial frozenness can be opposed to the "absolute" compositionality as understood by Mel'čuk (2010).

The flexibility of MWEs is also largely addressed in the seminal paper by Sag et al. (2002), in which it becomes one of the main defining criteria for a MWE typology, including fixed, semi-fixed and syntactically-flexible expressions (cf. Section 3.1).

The variability of some classes of MWEs was also addressed by the community of computational terminology. Jacquemin (2001) shows that up to 30% of terms in a corpus are variants of those appearing in controlled lists which is an important challenge to many NLP applications. In (Savary & Jacquemin, 2003) we provide a contrastive state of the art study in rule-based and hybrid term extraction with a special impact on how well the existing methods account for linguistic variability of complex (multi-word) terms. We adapt and refine the definitions proposed by Jacquemin (2001). Namely, a terminological variation is a transformation of a controlled multi-word term that satisfies the following three conditions:

1. All "content" words (i.e. words other than prepositions, determiners, etc.) of the controlled term are preserved by the transformation or transformed into any of the 3 types of variants listed in point 2.

2. Content words of the variant may be graphically modified, and morphologically or semantically related to those of the controlled term, which yields:

   - **graphical variants**, e.g. *behavioral model* → *Behavioral model*, *lookup* → *Look-up*[2],
   - **morphological variants**, e.g. *students union* →*Student union*, *image converter* →*Image conversion*,
   - **semantic variants**, e.g. *genetic disease* → *Hereditary disease*, *automobile cleaning* → *Car washing*,

3. Words may be inserted or deleted and the order of words (or of their variants) may be modified but the dependency relations existing between content words of the original term must be preserved. Such word insertions/deletions or word order modifications yield **syntactic variants**, e.g. *date of birth* → *Birth date*, *processing of cardiac image* → *Image processing*.

[1] A lexicon-grammar is a table whose first column contains compounds under consideration and columns represent linguistic transformations typical for its syntactic structure; a cell in line i and column j is checked if compound i admits transformation j.

[2] Terms on the left-hand side of arrows are variants, while those on the right-hand side, spelled with initial capitals, are controlled terms, i.e. terms listed in a lexicon.

Different types of variations may co-occur, for example *diseases are familial* and *transmissible neurogenerative diseases* are morphological, syntactic and semantic variants of *Genetic disease*.

In (Savary & Jacquemin, 2003) we further study four subdomains of term extraction: (i) controlled phrase indexing (with initial data), (ii) free phrase indexing (without initial data), (iii) thesaurus enrichment (corpus-based terminology with initial data), and (iv) term acquisition (corpus-based terminology without initial data).

The **term acquisition** systems under study are: ACABIT (Daille, 1994, 1996), ANA (Enguehard & Pantera, 1995), LEXTER (Bourigault, 1993, 1994, 1996), TERMINO (David & Plante, 1990a,b), TERMS (Justeson & Katz, 1995) and Xtract (Smadja, 1992). Three of them apply to French, three to English, one to Malgasy and one is language-independent. All systems use tagging, morphological analysis or stemming, and all but one rely on syntactic patterns followed or preceded by statistical filtering. The linguistic variation is taken into account to a rather limited extent, except in ACABIT, where a good coverage of syntactic variants is handled by syntactic transformation rules. The same paper also gives a more in-depth description of FASTR (Jacquemin, 2001), a shallow parser based on unification grammar and meta-grammar, specifically dedicated to the recognition, normalization and acquisition of compound terms and their variants in English and French.

The paper further presents a contrastive state-of-the-art of **phrase indexing** (indexing using multi-word terms) systems: CLARIT (Evans et al., 1991), COP (Metzler & Haas, 1989; Metzler et al., 1989, 1990), COPSY (Schwarz, 1989, 1990), the Fagan indexer (Fagan, 1987), FASIT (Dillon & Gray, 1983), IRENA (Arampatzis et al., 1997, 1998), NPtool (Voutilainen, 1993), the Sheridan/Smeaton indexer (Smeaton & Sheridan, 1991; Sheridan & Smeaton, 1992), the Sparck Jones/Tait variant generator (Sparck Jones & Tait, 1984b,a), SPIRIT (Andreewsky et al., 1977), and TTP (Strzalkowski & Vauthey, 1992; Strzalkowski, 1994, 1995; Strzalkowski & Scheyen, 1996). All of these tools but one (SPIRIT for French) apply to English. Most of them rely on morphological analysis, stemming or part-of-speech tagging, as well as shallow or deep parsing. Almost all systems account for some types of term variation via variant conflation (attaching document variants to the query terms) or variant generation (straightforward text match of expanded query terms). A common technique in variant conflation is to transform query and/or document terms into binary head-modifier relations (*the efficiency of these four sorting algorithms* → *algorithm efficiency + sorting algorithm*) which are then used as indexation terms. Usually, only noun phrases are addressed by this process although verbal and adjectival phrases may be equally informative (*index a document* → *document indexation*). An update of this state-of-the-art study is worth while, especially in the context of extending syntactic parsing to wide classes of Multi-Word Expressions, notably verbal ones (cf. Sections 3.2.5 and 7.3). Such an update would be also interesting with respect to the impact of using MWEs and deep linguistic analysis (including parsing) on the quality of information retrieval (IR). Namely, until the 2000s the usefulness of such linguistically motivated techniques was highly controversial (Brants, 2003). Nowadays, it appears that the *the pendulum might have swung too far* (Church, 2011) and a renewed interest in *higher hanging fruit to pick* appears both in fundamental and in applied research.

Chapter 3

Multi-Word Expressions

Multi-word expressions (MWEs) encompass a bunch of hard-to-define and controversial linguistic objects (Habert & Jacquemin, 1993; Corbin, 1992). Their numerous linguistic and pragmatic definitions (Benveniste, 1974; Downing, 1977; Levi, 1978; Gross, 1990; Silberztein, 1993b; Gross, 1996; Cadiot, 1992; Sag et al., 2002; Derwojedowa & Rudolf, 2003) invoke three major points:

- they are composed of two or more words,
- they show some degree of morphological, distributional or semantic non-compositionality (or idiosyncrasy),
- they have unique and constant references.

However, the basic notions (a word, a reference, the non-compositionality) and measures (degree of non-compositionality), used in those definitions are themselves controversial.

A basic fact about MWEs is that their are **prevalent** both in corpora and in lexicons of a natural language. For instance, Gross & Senellart (1998) showed that more than 40% of all tokens in a one-year corpus of the French journal *Le Monde* belong to multi-word units or expressions, and should not be analyzed individually. Sag et al. (2002) cite some studies considering the number of multi-word expressions as high as the one of single words, and argue that these figures are an underestimate, especially in terminological sublanguages.

Another important characteristics of MWEs is that, like most other linguistic units, they are subject to **sparseness** problems. In (Savary, 2000) I show that 85% of all graphically distinct compound noun forms appear less than twenty times in a one-year corpus of the *Herald Tribune*. Baldwin & Villavicencio (2002) experimented with a random sample of two hundred English verb-particle constructions and showed that as many as two thirds of them appear at most three times in the Wall Street Journal corpus. These facts are particularly challenging for corpus-based statistical methods for MWE extraction and identification (cf. Sections 3.2.2 and 3.2.3).

The difficulty in the automatic treatment of MWEs lies in their **idiosyncratic** behavior which can occur **at different levels** of traditional language processing chains: segmentation, morphology, syntax, semantics, etc. At the segmentation level, MWEs can form single tokens: *passer_by*, (FR) *bon_hommes* 'fellows'; include separators (non-alphabet letters): (FR) *aujourd'hui* 'today', *λ-calculus*; cross token boundaries: *personal computer*; or embrace discontinuous sequences tokens: *put* sth. *off*. At the morphological level they can be exocentric: (FR) *perce-neige* '[pierces$_{3pers.sing}$-snow$_{sing.fem}$]$_{sing.masc}$ = snowdrop'; have irregular inflection: (FR) *grand-mères* 'grand$_{sing.masc}$-mothers$_{pl.fem}$'; or have defective paradigms: (PL) **wybór powszechny* 'general election', *wybory powszechne* 'general elections'. At the syntactic level they have irregular structures: *all of a sudden*; or they block some transformations typical

for their (regular) structures: **the bucket was kicked by him.* At the semantics level, they show a varying degree of non-compositionality: *to spill the beans* = to reveal a secret.

3.1 Heterogeneous Nature of Multi-Word Expressions

The seminal paper by Sag et al. (2002) distinguishes four main types and six subtypes of MWEs:

1. **Fixed expressions** defy conventions of grammar and compositional interpretation but undergo no internal variation (*by and large*, *in short*, *ad hoc*, *Alta Vista*). They can be automatically processed as *words with spaces*.

2. **Semi-fixed expressions** have a fixed word order but undergo some degree of lexical variation. Since they hardly admit insertions of external elements they can be treated as full complex phrases with a single part of speech (MW nouns, MW adverbs, MW adjectives, MW verbs, etc.). They are subdivided into:

   - Semantically **non-decomposable idioms**, such as *to kick the bucket* (to die), *to shoot the breeze* (to have an informal conversation). They have opaque semantics and (therefore) admit no syntactic variability (**the bucket was kicked*), even if they inflect (*kicked the bucket*).
   - **Compound nominals**, with non-decomposable semantics, e.g. *part of speech* (grammatical category), *attorney general* (senior legal officer).
   - **Proper names**, which have constant referents but hardly any meaning, thus cannot be seen as semantically compositional. They are however subject to complex syntactic transformations: *the San Francisco 49ers*, *those San Francisco 49ers*, *San Francisco 49ers*, *the 49ers*, *the league-leading 49ers*, *those 49ers*, *49ers*, etc.

3. **Syntactically-flexible expressions** exhibit a large syntactic variability and are therefore subject to discontinuity problems (thus can be treated neither as words with spaces, nor as full complex phrases). They include:

   - **Verb-particle constructions**, either semantically idiosyncratic (*brush up on*) or compositional (*break up*). They can admit insertions of largely unrestricted NP arguments (*to call* one's friend *up*) or adverbs (*fight* bravely *on*). They defy a compositional approach due to idiosyncrasies (*call/ring/phone/telephone* vs. *call/ring/phone/ *telephone up*).
   - **Decomposable idioms**, whose semantics can be deduced from their components provided that the components themselves are interpreted in an idiomatic way, e.g. *to spill the beans* can be analyzed as made up of *spill* in the sense "to reveal" and *the beans* in the sense "the secret".
   - **Light verb constructions** (LVCs) also known **support-verb-nominalisation** (SVN) constructions. They are combinations of a verb and a noun, in which the former has lost its meaning to some degree and the latter is used in one of its original senses (*have lunch*, *give a try*, *make a decision*, *make use*). The syntactic head in an LVC is the (light) verb, while its semantic head is the noun. They are highly idiosyncratic since it is hard to predict which light verb combines with a given noun.

4. **Institutionalized phrases** are semantically and syntactically compositional, but statistically idiosyncratic. E.g. *traffic lights* means the same as *intersection regulator*, but the former has been conventionalized and not the latter. Such phrases admit a wide range of syntactic transformations.

Mel'čuk (2010) defines a different typology (in French) shown in Fig. 3.1. It is mainly based on the idea of selection constraints and semantic non-compositionality, and it attaches a lesser importance to the problems of morphological or syntactic (in)flexibility. MWEs are called **phrasemes**, i.e. non-free phrases in which the choice of at least one component is constrained by the other component(s). Such a constrained component cannot be replaced by all of its synonyms, even if it can be replaced by some of them, e.g. *to call/ring/phone/*telephone* somebody *up*. The following types and subtypes of phrasemes are distinguished:

1. **Pragmatic phrasemes**, or **pragmatemes**, in which the constraints on choice of components occur in the process of translating a given conceptual representation (CR) into a chosen semantic representation (SemR). For instance, the situation when an author wishes to put an impact on a certain part of a citation is represented by different meanings in different languages: (FR) *C'est moi qui souligne* 'I underline', (EN) *Italics/Emphasis mine*, (DE) *Hervorhebung des Autors* 'shift by the author', etc.

2. **Semantic phrasemes**, in which the constraints occur in the process of translating the semantic representation (SemR) into the expression itself. They are further subdivided into:

   (a) **Semantically compositional phrasemes**, including:

   - **Clichés**, which are lexically, inflectionally and syntactically totally fixed but semantically compositional, as in: *in other words* (rather than: **with different lexemes*), (FR) *Ce qu'il fallait démontrer* 'what was to be demonstrated' (rather than **Ce qu'il était nécessaire de prouver* 'what was necessary to be proved').
   - **Collocations**, which contain one unconstrained and one constrained component and are semantically compositional: (FR) *agile comme un singe* 'as agile as a monkey', *aimer à la folie* 'to love to madness', *décerner un prix* 'to award a price'.

   (b) **Semantically non-compositional phrasemes**, also called **locutions**. In a locution each component is freely chosen but the meaning is non-compositional, as in (FR) *vache à lait* 'a milk cow = an exploited person', (FR) *jeter l'éponge* 'to throw the sponge = to abandon, to throw the towel', etc. A locution can show a different degree of transparency with respect to the meaning of its components:

   - In a **quasi-locution** the meanings of the components are combined but the semantic head is missing. For instance, in (FR) *donner le sein à X* 'give the breast to X = to breastfeed X' the main semantic component 'to feed a baby' is expressed by none of the component words.
   - A **semi-locution** includes the meaning of only a part of its components and its semantic head is missing. E.g. the meaning of (FR) *fruits de mer* 'sea fruit = seafood' contains the sense of *de mer* 'of sea' but not of *fruit*, and no component represents the meaning of 'animals other than fish'.
   - A **complete locution** includes the meaning of none of its components, as in (FR) *faire table rase* 'to make a shave table = to account for no preceding activity or events', (FR) *en tenue d'Adam et Eve* 'in Adam's and Eve's dress = naked'.

The two very different typologies cited above show that MWEs are large groups of linguistic units of a very diverse nature and properties. Consequently, their description and automatic processing usually follow versatile rules. In most of my contributions to this domain I define the scope of MWEs pragmatically: we consider a MWE as a sequence of graphical items which, for

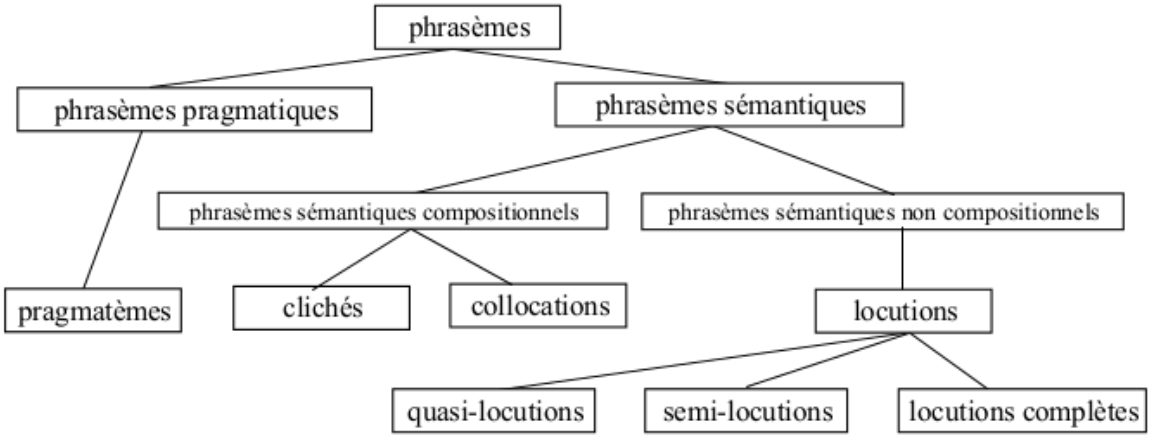


Figure 3.1: A typology of MWEs by Mel'čuk (2010).

some application-dependent reasons, has to be listed, described (morphologically, syntactically, semantically, etc.) and processed as a unit (Savary, 2005).

3.2 Lexical Representation and Automatic Processing of Multi-Word Expressions – State of the Art

Multi-Word Expressions (MWEs) have been subject to extensive linguistic studies for several decades. Their complex behavior makes them one of the major challenges in natural language processing – it's *pain in the neck*, as the seminal paper by Sag et al. (2002) puts it. Formal description and automatic processing of MWEs now has a growing international community gathered, notably, around the Multi-Word Expression Workshop[1] organized on an almost yearly basis since 2003. Special issues on MWEs were edited in several journals (Villavicencio et al., 2005; Rayson et al., 2010; Ramisch et al., 2013a,b; Szpakowicz et al., 2013).

In this section I review some research approaches to the lexical description of MWEs, to their automatic extraction and identification in corpora, as well as their links with treebank annotation and parsing.

3.2.1 Lexical Description of Multi-Word Expressions

The fact that MWEs are word sequences with unpredictable properties is often related to the idea of **lexicalization**: the unpredictable, related to an individual word combination, undeducible from the general grammar rules of the language, has to be encoded in a lexicon. This task is not as simple as its formulation suggests. Firstly, it assumes that the properties to be tested and their testing procedures, as well as the notion of unpredictability itself, are well defined. Secondly, it requires appropriate description formalisms, which should ideally express the degree of lexicalization and be application-independent. Thirdly, such descriptions should be as factorized as possible, i.e. only the unpredictable properties should be expressed in the lexicon, while the regular ones should be referenced at the grammar level.

Since answering all of these challenges at once is complex, simpler, more easily achievable goals have been defined in the MWE community. One of them is to simply collect lists of MWEs, which can be further extended with their parts of speech and other data necessary for

[1] `http://multiword.sourceforge.net/PHITE.php?sitesig=CONF`

their identification in text. This implies taking their variability, notably at the level of inflection, into account.

In (Savary, 2008) I provide a contrastive state of the art study of different approaches to the lexical description of (especially contiguous) MWEs, notably with respect to inflection. I address:

- the Paris school of **DELA**[2] **electronic dictionaries** (Courtois & , eds.): the French DELAC[3], (Silberztein, 1993a), the English DELAC (Savary, 2000), the Greek DELAC (Kyriacopoulou et al., 2002), the NooJ DELAC (Silberztein, 2005) and Multiflex applied notably to Serbian (cf. Section 3.3);
- the **two-level morphology** approaches using the finite-state lexicon compiler, *lexc*, accompanied by the regular expression compiler, *xfst* by Beesley & Karttunen (2003): the French lexc module for compounds (Karttunen et al., 1992; Karttunen, 1993), IDAREX for German (Breidt et al., 1996) and a multi-word processor for Turkish (Oflazer et al., 2004);
- a **relational database** approach: HABIL for Basque (Alegria et al., 2004);
- **unification grammar and meta-grammar** approaches: the LinGO project for English (Sag et al., 2002; Copestake et al., 2002; Villavicencio et al., 2004), and FASTR for English and French (Jacquemin, 2001).

This study allowed me to put forward recommendations concerning the best practices for lexical description of contiguous MWEs, which can contribute to the ongoing work on standards, such as Calzolari et al. (2002) and ISO/TC 37/SC 4 (2007):

1. **A variety of natural languages should be taken into account during elaboration of standards**. The predominating position of the English language has prevented the NLP research from a full appreciation of the importance of morphological phenomena in multi-word units. Taking into account lesser studied, often inflectionally rich, languages, such as Slavic or concatenative languages, should lead to more universal models, platforms and standards.

2. For instance, the study of these languages calls for the **necessity of a unification mechanism** for a compact description of agreement rules between components, as well as of huge inflectional paradigms.

3. If we wish to provide a reusable and universal morphological resource of MWEs, it is important to **keep in mind at least the two most general linguistic applications: the morphological analysis and generation**. In particular, it should be possible not only to identify a MWE in a corpus but also to annotate it with morphological features necessary for further processing stages. Approaches like IDAREX, which do not allow for annotation, seem satisfactory only for a limited number of applications (e.g. concordancers).

4. On the very basic graphical level, the NLP community is still far from reaching a consensus on what should be considered as an elementary indivisible unit. Morphological analyzers of simple words differ at this point, even with respect to the same natural language. However, defining the graphical frontier between lexical units is necessary, as it influences the way

[2]DELA stand for *LADL's electronic dictionary*, where LADL is the name of the central laboratory having proposed the methodology.

[3]DELAC stands for the *LADL's electronic dictionary for compounds*.

how multi-word units are defined and processed. I think that **the definition of a lexical unit should be flexible, and adaptable to each new language or application**. In particular, it should be possible to describe squeezed compounds (e.g. *passers_by*) as sequences of simple words. Conversely, sequences containing blanks (e.g. *a priori*) should be describable as indivisible tokens. Moreover, it should be possible to view separators, punctuation marks, digits, etc. as full members of MWEs and allow to describe their absence, presence and variation.

5. For an efficient human usage and treatment, non-abstract lemmas[4] of MWEs should be offered to lexicographers. Since a lemma of a MWE may contain simple words that are not lemmas themselves, avoiding abstract multi-word base forms requires the **annotation of simple components with their own base forms and features**, as in the English DELAC, Multiflex and HABIL.

6. **The extensiveness of orthographic, morphological, syntactic and semantic variation calls for a common descriptive framework** in which all those types of variations could be taken into account. Here again, lesser studied languages, such as Turkish, reveal new types of morphosyntactic variants such as duplications.

7. In order to express omissions, insertions and order changes, it is necessary to **refer to the position of a single component in a compound**. In the existing approaches that may be done either by numbering lexical items (as in IDAREX, Multiflex, HABIL and FASTR), or by regular expressions that identify token frontiers (as in lexc).

8. Most often, morphological forms that simple words take within MWEs, are subsets of the inflectional paradigms of these words. Thus, it seems most natural to admit a **'two-layer' approach**[5]:

   - Describing the morphology of simple words as individual units.
   - Describing multi-word units as morphologically and syntactically conditioned compositions of simple words and other lexical items, such as separators, digits, etc.

   Approaches, such as NooJ, in which this postulate is not assumed, suffer from a too high degree of redundancy in component morphology description.

9. Studies on the morphological treatment of simple words have been developed for decades and resulted in a large number of formalisms and tools in various languages. Rather than impose a uniform framework both for simple words and MWEs, it seems reasonable to encourage **modularity and interoperability**. Thus, a morphological module for MWEs should be able to interact with any such module for simple words, provided that some interface constraints have been properly defined and respected.

10. In order to reach large-scale dimensions in MWE resources, **tools for automated lexicon enrichment** (as opposed to lexicon construction from scratch) should be integrated into the descriptive process. Such tools should allow to assign inflection rules to MWEs semi-automatically (Krstev et al., 2006a; Marciniak et al., 2009b; Sikora & Woliński, 2009; Krstev et al., 2013). They might be based on rule and corpus mining.

[4] An abstract MWE lemma stems from lemmatizing each constituent word individually, as in *mémoire vif* 'live$_{sing.masc}$ memory$_{sing.fem}$ = random access memory'. A non-abstract lemma takes the syntactic structure of a MWE into account, e.g. *mémoire vive* 'live$_{sing.fem}$ memory$_{sing.fem}$'.

[5] Not to be confused with Koskenniemi's two-level morphology.

11. Non-contiguous MWEs, as well as their sense computation, remain a challenge. Studies dedicated to multi-word expressions should focus as much on their morphological constraints as on their semantic complexity.

Since the publication of the (Savary, 2008) paper, several new approaches to the lexical description of MWEs have been published, including two prominent ones: for Dutch (Grégoire, 2010) and for Hebrew (Itai & Wintner, 2013). Both of them seem to confirm the above recommendations but they go beyond contiguous MWEs.

The MWE lexicon for Hebrew (which is a highly inflectional language) (Itai & Wintner, 2013) contains over 3,700 entries and follows the two-layer approach (cf. recommendation 8 above). MWE components are identified via pointers towards their entries in the morphological lexicon of simple words and their inflectional properties. Agreement between components and feature inheritance from headwords are expressed by a unification-like mechanism, similar to the one in Multiflex (cf. Section 3.3). Syntactic variants, including word order change and ellipsis, are described due to component numbering. Irregular inflection via non-standard prefixes is signaled by special attributes. Similarly to HABIL (Alegria et al., 2004), open slots can be marked for unconstrained modifiers, i.e. only the number of inserted external elements can be specified but not their morphological, syntactic or semantic properties. This makes the MWE lexicon well adapted to morphological analysis but its integration into deep parsing would require an additional expressive power.

DuELME (Grégoire, 2010) is a lexicon of Dutch multi-word, notably verbal, expressions. It contains about 5,000 entries. Candidate MWEs are extracted from a corpus by pattern-based methods and divided by a decision-tree classifier into probable true and false positives. Their variants in the corpus are analyzed in order to detect their unpredictable properties. Pre-selected MWE candidates are then validated and described with the two-layer approach. Firstly, the lemmas of the lexically fixed components are identified (the morphological features of these components are stated in external parameters) and some restrictions for the non fixed components are expressed, e.g. animate object, admitted pronominalization, modal verbs going with the head component (*have* or *be*), possible adjectival modifiers, and restriction to negated use only. Secondly, the MWE is assigned a pattern. Patterns are represented as *parameterized equivalence classes* which reflect the syntactic structure of MWEs. A sample class is: *expression headed by a verb, taking a direct object consisting of a fixed determiner and a modifiable noun*, whereas an external parameter states if the object noun is in singular or in plural. Parameters allow to prevent the explosion of the number of classes. The DuELME formalism is meant to be theory- and implementation-neutral and its applicability to a particular dependency parser has been demonstrated. I think that this description framework is promising in that it applies to the lexical description of verbal MWEs and offers an abstract formalism, which can potentially be compiled into different parsing frameworks.

3.2.2 Multi-Word Expression Extraction

In order for a MWE lexicon (may it be a plain list of keyword phrases or a more complex database) to be corpus-based and created in a maximally efficient way, **MWE extraction** (MWEE) techniques have been developed since the nineties. The MWEE task consist usually in obtaining a list of word combinations which have the MWE status independently of a particular context, i.e. they appear in a corpus at least once as MWEs. Initially, the most studied subtask was *terminological extraction*, and already then hybrid (both knowledge-based and data-driven) approaches proved the most efficient (Smadja, 1992; Daille, 1996). In (Savary & Jacquemin, 2003) we provide a contrastive state of the art study in rule-based and hybrid term extraction with a special impact on how well the existing methods account for linguistic variability of

complex (multi-word) terms (cf. Section 2.3).

Further on, the MWEE task was extended to more generally defined classes of (especially contiguous) MWEs, such as named entities, compounds or light verb constructions, and is now a relatively well-studied problem. In many systems, it is represented as an instance of an n-gram classification problem. N-grams are extracted from a corpus, filtered using heuristics and assigned feature vectors. These features rely of various association measures (AM), which, roughly, capture the degree to which the components of an n-gram appear together more often than expected by chance. The appropriateness of different AMs for the MWE was largely studied. E.g. Davis & Barrett (2013) use pointwise mutual information (PMI) to predict the acceptability of SVN constructions (*take a walk, give a presentation*). Pecina (2010) shows that combining various AMs in Czech binary collocation extraction can result in a mean average precision of up to 86.3% (collocations are understood in this study as binary word combinations which do not necessarily correspond to valid syntagmatic groups). Feature vectors may also be enriched with linguistically-motivated features. For instance, Al-Haj & Wintner (2010) analyse nominal compounds in Hebrew and use their idiosyncratic linguistic properties as features of an SVM classifier, additionally to classical AM measures.

MWEE can benefit from a multilingual context. Since MWEs are usually at least partly semantically non-compositional they are usually not translated word by word. This fact is leveraged by Tsvetkov & Wintner (2010), who perform MWEE in parallel Hebrew-English word-aligned texts. Word-to-word aligned word pairs are removed and the remaining sequence pairs are fed to AM calculations, in order to retain the best MWE candidates. Morin & Daille (2010), conversely, rely on compositional translation of two-word terms in order to extract bi-lingual dictionaries from a French–Japanese comparable corpus[6]. They use morphologically-based stripping-recoding rules in order to capture translation equivalents with non-equivalent syntactic structures (*apport nutritif* 'nutritive intake' vs. *nutrition intake*). In (Delpech et al., 2012), these ideas are extended to extracting compositional translations of squeezed terms (e.g. *cardiotoxicity*) into multi-word terms (*toxixité cardiaque* 'cardiac toxicity') and their variants (*toxicité pour le coeur* 'toxicity for the heart') from English-French and English-German comparable corpora, which yields an average precision of 91% on the best candidate translation.

As in most data-driven approaches, while relying on a limited corpus for AM calculation, one faces the problem of data sparseness: most data appear very rarely or never in the corpus. Ramisch et al. (2010) address this issue in that the web is exploited as a corpus. They propose estimation and normalization measures to cope with two problems arising in this approach: (i) the size of the web is unknown but it is necessary for statistical measures, (ii) web crawlers do not respect the Zipfian phenomenon for the most frequent keywords.

3.2.3 Multi-Word Expression Identification

A task similar to MWEE is **MWE identification** (MWEI) in a running text, i.e. deciding if a given (more or less contiguous) sequence of tokens is a MWE in a given context. Additional challenges with respect to MWEE result notably from MWE ambiguity, i.e. the fact that a sequence of tokens corresponding to a MWE in one context may be a non-MWE in another context. Moreover, true MWE rather than collocations are often sought for in this task, i.e. detecting the MWE boundaries is an inherent part of the task.

An important subtask, intensively studied for over two decades, is the *named entity recognition* (NER), discussed in detail in Section 4.2 but other MWE subclasses were addressed in different identification frameworks as well. For instance, Vincze et al. (2013) show that in

[6]A comparable corpus, usually easier to obtain than a parallel corpus, is a collection of texts which are no translations of one another but concern the same topic and were produced in similar communication conditions.

Hungarian and English most LVCs are contiguous bigrams or trigrams[7] and build a CRF-base sequential tagger for their identification. They show that syntax-based features (e.g. links from a dependency treebank) are crucial in English MWEI, while they are less influential in a morphologically rich language such as Hungarian, due to the fact that the Hungarian morphology encodes a lot of (morpho)syntactic information. This fact is an interesting argument towards leveraging such languages in the NLP community. The same authors show that efficiency of MWEI is domain dependent (cross-domain experiments always yield weaker results than in-domain ones).[8]

3.2.4 Annotating Multi-Word Expressions in Corpora

MWEs, as particularly interesting linguistic objects, have obviously attracted attention of corpus linguistics. Modeling their behavior in annotated corpora, and prominently in treebanks, has been undertaken in various languages and linguistic frameworks. Abeillé et al. (2003) describe the **French Treebank (FTB)**, a theory-neutral, automatically pre-annotated and manually corrected 1-million word French corpus of newspaper texts, annotated for morphology, syntax and functional relations. Contiguous compounds, detected mostly according to linguistic tests proposed by Gross (1996), compound proper names and temporal expressions are represented as 3-level flat substructures (i.e. their categories, morphological tags of their constituents, and the constituent words themselves).

Bejček & Straňák (2010) discuss another large treebank, the **Prague Dependency Treebank** of Czech (Böhmová et al., 2003), annotated at 3 layers: morphological, analytical (accounting for syntax) and tectogrammatical (accounting for functional relations), in which MWEs are annotated by identifying the corresponding subtrees of the 3rd layer and then replacing these (monosemic) subtrees by single nodes. An associated MWE lexicon stores previously found subtrees for further annotation automatization. In (Bejcek et al., 2011) it is further shown how tectogrammatical dependency subtrees unify different morphosyntactic variants of the same MWE. It is also argued that elements elided in MWEs (e.g. due to coordination) should be restored in deep syntactic trees.

In (Laporte et al., 2008a,b) another manually annotated French corpus, distributed under an open license, is described. It contains about 38,000 words and about 9,500 MWEs of two categories: compound adverbs (annotated for flat syntactic structure) and compound nouns (annotated for flat syntactic structure and inflectional features).

Kaalep & Muischnek (2008) discuss a 300,000-word Estonian corpus annotated for morphology and (mostly binary) multi-word verbs (MWVs): phrasal verbs, idioms, light verbs constructions, and verb infinitive constructions, most of them concerned by inflection, word order variation and discontinuities. The annotation scheme is linear: each token obtains its morphological tag, and – if it belongs to a MWV – it is assigned this MWV's citation form and an indication if the other component of the same MWV appears to the left or to the right. In total, 8,200 instances of tagged MWVs occur, i.e. every fifth predicate is represented by a MWV.

Named entities, many of them composed of several tokens, have been specifically addressed in many other annotation efforts whose partial state-of-the-art in presented in Section 4.2.1.

3.2.5 Parsing and Multi-Word Expressions

Multi-Word Expressions truly cross boundaries between traditional layers of linguistic processing, notably between lexicon, syntax and semantics. Even if some idiosyncrasies of MWEs call

[7] Split LVCs account for up to 21% text occurrences.

[8] Domain-dependence has always been obvious e.g. in terminological extraction but might be partly surprising in seemingly universal constructions such as LVCs.

for their lexical description, other regular properties make them resemble well-formed syntactic structures. Therefore, one of the main challenges is the most appropriate integration of MWE processing within parsing. This problem has been studied within different theoretical linguistic frameworks.

Abeillé & Schabes (1989) show how French MWEs can be integrated in parsing with **Lexicalized Tree Adjoining Grammars** (LTAGs). An LTAG grammar contains a finite set of *elementary* (initial or auxiliary) *trees* each of which has at least one lexicalized element (called the *head*). MWEs are represented as special kinds of elementary trees in which heads are made out of several lexical items that need not be contiguous. One of them serves as an index. During parsing, a sentence can be derived by combining elementary trees via *substitution* (inserting an elementary tree at a non-terminal leaf) or *adjunction* (inserting an elementary tree at a non-terminal internal node), which yields a *derived tree* (the syntactic structure of the sentence) and a *derivation tree* (showing which elementary trees have been combined and how). While parsing ambiguous MWEs (e.g. *He kicked the bucket.*), the idiomatic and the literal readings obtain the same derived trees but the derivation trees differ. Accordingly, the idiomatic semantics stems from direct attachment of lexical items in the elementary trees, while the literal compositional semantics is a product of substitution (of non-terminal nodes with lexicon items). Linguistic transformations such as passivization, interrogative clauses, etc., are handled in MWEs as in compositional phrases by grouping different elementary trees into tree families. The parsing process consists of two steps. Firstly, the elementary trees corresponding to literal and idiomatic readings are selected (provided that all their lexicalized items are present in the sentence). This reduces the search space of potential elementary trees to be combined. Then, the syntactic analysis is pursued as in the usual case. At this stage, an idiomatic reading can be rejected if the syntactic dependencies or the unification constraints are violated.

LTAGs show several advantages with respect to parsing MWEs. Firstly, unification constraints on feature structures attached to tree nodes allow one to express dependencies between arguments at different depths in the elementary trees, as in NP_0 *vider DET sac* 'to express one's secret thoughts', where the determiner DET embedded in the direct object must agree in person and number with the subject NP_0. Secondly, the *extended domain of locality* offers a natural framework for representing two different kinds of **discontinuities**. Namely, *discontinuities coming from internal structure* are directly visible in elementary trees (e.g. via the NP_1 non-terminal node in NP_0 *takes* NP_1 *into account*) and are handled in parsing mostly by substitution. *Discontinuities coming from insertion of modifiers* (e.g. *a bunch of NP*, *a* <u>*whole*</u> *bunch of NP*) are invisible in elementary trees but are handled in parsing by adjunction.

Sag et al. (2002); Copestake et al. (2002); Villavicencio et al. (2004) address the representation of English MWEs within a **Head-Driven Phrase Structure Grammar** (HPSG), with a particular impact on semantics. A simplex entry in an HPSG lexicon consists of a triple: orthography, type and semantic predicate. Constraints on the type are expressed as a typed feature structure (TFS), while semantics is represented by an atomic predicate. The Minimal Recursion Semantics formalism allows to combine elementary semantic predicates of a semantically compositional expression (e.g. *to spill water*). Partly compositional semantics is treated by paraphrasing. For instance, the lexemes *to spill* and *beans* are linked both to their literal meanings and to their paraphrases *to reveal* and *secret*, used for calculating the semantics of the idiom *to spill the beans*. Finally, MWEs with opaque semantics (e.g. *to kick the bucket*) are represented, similarly to simplex words, with separate semantic predicates having no links to the semantics of the component words.

Attia (2006) handles MWEs in Arabic via a finite-state machinery and the **Lexical Functional Grammar** (LFG). Fixed (*in a nutshell*) and adjacent semi-fixed (*traffic light*) MWEs are first processed by a composition of finite-state lexical transducers (Beesley & Karttunen,

2003) which simultaneously divides one-word phrases into components (e.g. *andtominister* → *and@to@minister*) and joins MWEs into words with spaces (e.g. *minister@foreign* → *minister foreign* 'foreign minister'). The latter are then handled at the syntactic parsing stage as single tokens (they obtain single nodes in the C-structures and single feature structures in the F-structure). Syntactically flexible MWEs are handled by the grammar only. For instance, *Noun Adj* compounds in Arabic typically allow for an insertion of a genitive modifier between the noun and the adjective: *bike fiery* 'motorbike', *bike the-boy the-young the-fiery* 'the young boy's motorbike'. Lexical rules express modifier selection for MWE headwords, e.g. a rule states that if *bike* is modified by *fiery* its meaning is *motorbike*, otherwise its meaning is *bike*. Such a MWE with an inserted genitive is handled by usual grammar rules as syntactically compositional but as semantically non-compositional, due to the lexical selection rules. As a result, the nodes representing *bike* and *the fierry* are separated in the C-structure by the nodes for *the-boy the-young* but the meaning of *bike* in the F-structure is *motorbike* (`TRANS=motorbike`) instead of *bike*. Lexical selection rules also cover phrasal verbs, e.g. a rule states that the object of *rely* has to be preceded by the preposition *on*. This shows strong links between LFG lexical rules and valence dictionaries.

Interestingly enough, the proposals from the LTAG (Abeillé & Schabes, 1989) on the one hand, and from the HPSG (Sag et al., 2002; Copestake et al., 2002; Villavicencio et al., 2004) and the LFG (Attia, 2006) communities are complementary in that, while addressing the lexical encoding of MWEs for unification grammars, the former stress mainly the syntactic and the latter the semantic (non-)compositionality issues. Schuler & Joshi (2011) argue, however, that Tree Adjoining Grammars and all other **tree-rewriting (or substitution) systems (grammars)** are a more natural candidate for modeling MWEs than **string rewriting systems**, such as the HPSG or categorial grammars, since they can model entire fragments of phrase structure trees as elementary blocks. The accuracy of tree-substitution grammars (TSGs) for MWE processing was also confirmed within a probabilistic paradigm by Green et al. (2013), who show that a TSG parser performs better than a Probabilistic Context Free Grammar (PCFG) parser on both French and Arabic contiguous MWEs.

Among the works which evaluate parsing performances on real-size data, the crucial issue discussed with respect to parsing and MWE identification is the relative order of these two tasks in NLP processing chains. Some proposals include the identification of MWEs **before** the parsing. For instance, Nivre & Nilsson (2004) train a probabilistic dependency parser on two versions of a Swedish corpus annotated for contiguous MWEs. The results show that taking MWEs into account increases both the robustness of the parser (the percentage of sentences receiving a projective dependency graph) and its accuracy, both inside the MWEs and in the surrounding structures. Parsing errors that can be eliminated through the MWE pre-identification concern notably: (i) MWEs with irregular syntactic structures (e.g. the compound preposition *as regards* which tends to be analyzed as a clause), (ii) prepositional phrase attachments (e.g. in *is run as a rule by the commune*, where the *PP by the commune* tends to be attached to the head of the compound adverb *as a rule*).

Other approaches perform MWE identification **after** parsing, e.g. by re-ranking the parser's outputs in that MWE-oriented interpretations are promoted (Constant et al., 2012).

Wehrli et al. (2010) argue that the results of both the MWE identification and of parsing, as well as of further parsing-based applications, e.g. in machine translation, are always enhanced when both tasks are performed **simultaneously**. MWE identification, applied in this Chomskian grammar-based approach to French, is based on two rules: (i) components of MWEs are marked in the lexicon with a special lexical feature, (ii) when a noun N (e.g. *record*) having this feature appears in a (partial) parse tree each of its governing nodes G (e.g. *break*) is checked to see if an attested MWE can be formed out of G and N. If so then this analysis is given high

priority over competing analyses.

In the probabilistic framework, joint parsing and contiguous MWE identification has been addressed, notably, in several Stanford tools trained on MWE-annotated corpora. It these corpora, parse trees are augmented so as to contain special non-terminal nodes representing MWEs. Consequently, MWEs are represented directly in the resulting grammar. Finkel & Manning (2009a) show that parsing performed jointly with (nested) named entity recognition (cf. p. 78) by a Conditional Random Field (CRF)-based PCFG parser increases the performance of both tasks in English. Green et al. (2011, 2013) use French and Arabic treebanks in which MWEs are represented by flattened structures headed by special non-terminals (MWN, MWA, MWP, etc.). Probabilistic CFG and TSG parsers trained on these corpora achieve over 36% improvement in MWE identification over n-gram surface statistics tools.

Finally, Constant et al. (2013) show that different strategies of (contiguous) MWE identification can be combined with parsing within the same framework. In the first stage they pre-identify MWEs by a CRF sequential tagger. Then, they perform joined lexical segmentation and POS tagging, taking knowledge on MWEs into account. Finally they parse the resulting word lattices (representing ambiguous segmentations) with a PCFG-based parser. The results show that the parsing quality improves dramatically for the Oracle segmentation and tagging of a sentence (i.e. the closest to the gold standard).

3.3 Multiflex — a Multilingual Tool for Describing the Morphosyntax of Multi-Word Units

As defined at the beginning of this chapter, Multi-Word Expressions (MWEs) encompass a wide range of heterogeneous syntactic structures with unpredicted properties. One of the major challenges in the automatic processing of these linguistic objects is their spanning possibly non-contiguous sequences of text tokens. In this chapter I will mainly address **Multi-Word Units** (**MWUs**) defined as contiguous MWEs. They encompass a number of hard-to-define (Habert & Jacquemin, 1993) linguistic objects: compounds, complex terms, multi-word named entities (addressed in more details in Chapter 4), multi-word lexemes, institutionalized phrases, etc.

As discussed in Section 2.3, MWUs show an important degree of flexibility at different levels: orthographic (*head word* vs. *headword*), inflectional (*gentleman farmer* vs. *gentlemen farmers*), syntactic (*birth date* vs. *date of birth*), and semantic (*hereditary disease* vs. *genetic disease*). Unfortunately, the flexibility of compounds is hard to represent precisely and exhaustively within general grammar-based models due to idiosyncrasy. For instance, *chief justice* and *lord justice* are morphosyntactically similar structures, but their plural formation is different: *chief justices* vs. *lord justices, lords justice* or *lords justices*.

One of my main contributions to automatic processing of MWUs is ***Multiflex***, a formalism and a tool that copes with flexibility and idiosyncrasy of MWUs by a fully lexicalized two-layer approach. First we admit that inflected forms of single words can be analyzed and generated by an external morphological module. Then we specify how to combine inflected forms of single components in order to obtain an inflected form of a MWU. For instance, in order to obtain the plural forms of *battle cry*, *battle royal* and *battle of nerves*, we need to be able to generate the plurals of *battle*, *royal* and *cry*. Then we need to say how these different forms combine: *battle cries*, *battle royals*, *battles royal*, *battles of nerves*, but not **battles cries*, **battles royals* or **battles of nerve_*. At the same time we take into account possibly many orthographic and syntactic variations. The description is done via a used friendly graph-based graphical interface.

Different aspects of Multiflex have been described in several publications. In (Savary, 2005) I introduce the first outline of the inflection graphs formalism for compounds, taking unification and value inheritance into account. In (Savary et al., 2007) we study the linguistic phenomenon of

morpho-syntactic non compositionality and its representation with Multiflex graphs. In (Savary, 2008) I compare the formalism with other tools dedicated to similar problems. In (Savary et al., 2009) we discuss the specific issues of the Polish version of Multiflex, its advanced facilities such as handling embedded MWUs, as well as some the interoperability issues. In (Savary, 2009) I describe the finite-state machinery behind Multiflex and discuss its applications. Finally, in (Graliński et al., 2010) we perform a usability study of the formalism and of the associated graphical user interface. This chapter is a comprehensive description of the present state of Multiflex from a multilingual point of view. Its universality is argued via the study of linguistic properties of MWUs in several European languages and their representation within Multiflex, namely in English (EN), German (DE), French (FR), Portuguese (PR), Polish (PL) and Serbian (SR). Morphological models of all these languages are described and the influence of different modeling rules on the subsequent graphs is discussed.

3.3.1 Linguistic Prerequisites

Before the inflection of a MWU can be described by Multiflex some initial information is required with respect to the language studied, and to the internal structure of the multi-word lemma.

Morphological Model

The general morphological model of a given natural language is first given by a list of all existing morphological elements: (i) **categories** (number, gender, etc.) and **features** admitted by each category (singular and plural for number, etc.), (ii) **classes** (noun, adjective, etc.), categories in which a class inflects, and those that are fixed for the class. Figure 3.2 shows extracts of models for five European languages: two Germanic (English and German), two Romance (French and Portuguese) and one Slavic (Serbian). For instance in English the number (*Nb*) category admits two values: singular (*s*) and plural (*p*). Nouns (*noun*) inflect for number *(Nb,⟨var⟩)*, adjectives (*adj*) for degree *(Deg,⟨var⟩)*, and adverbs are uninflected. In Serbian the gender category (*Gen*) admits three values: masculine (*m*), feminine (*f*), and neuter (*n*). There are three numbers: singular (*s*), plural (*p*) and paukal (*w*, dedicated to two, three or four objects). Adjectives (*adj*) inflect for number *(Nb,⟨var⟩)*, gender, case, and degree (*Comp*), while nouns (*subst*) inflect for number, case and gender, and have a fixed animateness *(Anim,⟨fixed⟩)*. Empty morphological values (⟨*E*⟩) are admitted. For instance, in Portuguese the degree (*Gr*) is unmarked if it is positive, and marked if it is diminutive (*D*), augmentative (*A*), or superlative (*S*). The labels chosen for identifying classes, categories and values are arbitrary and may vary with the underlying morphological module for simple words (cf section 3.3.3).

The morphological model for Polish, presented in Figure 3.3, includes a rather rich set of morphological categories, as well as additional graphical and user-defined facilities. This fine-grained tagset (Przepiórkowski & Woliński, 2003) admits 12 inflectional categories, 38 values, and 35 classes. It results from a critical review of existing tagsets in morphologically rich languages, in view of a better tagset interoperability. The main criteria for delimiting grammatical classes are morphological (how a given form inflects; e.g., nouns inflect for case, but not for gender) and morphosyntactic (in which categories it agrees with other forms; e.g., Polish nouns do not inflect for gender but they agree in gender with adjectives and verbs). Semantic criteria (e.g. being a proper or a common name) are eschewed. In this way, some traditionally admitted inflection classes such as 'pronoun' are eliminated and replaced by a finer set of **flexemic classes** (Bień, 1991). For instance, a **pronoun non-3rd person** (*ppron12*, e.g. *ja* 'I') has a number and person value and inflects for case (*mnie*, 'me$_{dative}$'), gender (*ja [byłam/byłem]*, 'I [was]$_{fem/masc}$') and accentability (*mi*, 'me'). A **pronoun 3rd person** (*ppron3*, e.g. *on* 'he') inflects for number (*oni* 'they'), case (*jego* 'him$_{genitive}$'), gender (*ona* 'she'), accentability (*go*

	CATEGORIES	CLASSES
English	Nb: s, p	noun: (Nb,⟨var⟩)
	Deg: ⟨E⟩, C, S	adj: (Deg,⟨var⟩)
		adv:
German	Nb: e, m	noun: (Nb,⟨var⟩),(Gen,⟨fixed⟩),
	Gen: M, F, N, U	(Cas,⟨var⟩)
	Cas: n, a, d, g	adj: (Cas,⟨var⟩),(Gen,⟨var⟩),
	Det: x, y, z, u	(Nb,⟨var⟩),(Det,⟨var⟩),
	Deg: p, k, s	(Deg,⟨var⟩)
French	Nb: ⟨E⟩,s,p	noun: (Nb,⟨var⟩),(Gen,⟨var⟩)
	Gen: ⟨E⟩,m,f	adj: (Nb,⟨var⟩),(Gen,⟨var⟩)
	Tense: W,P,Y,G,K,I,J,S,T,F,C	v : (Tense,⟨var⟩),(Pers,⟨var⟩),
	Pers : ⟨E⟩,1,2,3	(Nb,⟨var⟩),(Gen,⟨var⟩)
Portu-	Nb: s,p	noun: (Nb,⟨var⟩),(Gen,⟨fixed⟩),
guese	Gen: m,f	(Gr,⟨var⟩)
	Gr: ⟨E⟩,D,A,S	adj: (Nb,⟨var⟩),(Gen,⟨var⟩),(Gr,⟨var⟩)
Serbian	Nb: s,p,w	noun: (Nb,⟨var⟩),(Case,⟨var⟩),
	Case: 1,2,3,4,5,6,7	(Gen,⟨var⟩),(Anim,⟨fixed⟩)
	Gen: m,f,n	adj: (Nb,⟨var⟩),(Case,⟨var⟩),
	Anim: v,q,g	(Gen,⟨var⟩),(Anim,⟨var⟩),
	Comp: a,b,c	(Comp,⟨var⟩),(Det,⟨var⟩)
	Det: d,k,e	

Figure 3.2: Extracts of morphological models of five European languages in *Multiflex*

'him') and post-prepositionality (*niego* 'him'), and has a person value. Different flexemes of the same verb can also have very different inflectional behavior. Adverbial participles (*pcon*, *pant*, e.g. *czytając* 'reading', *przeczytawszy* 'having read') are uninflected and have aspect; non-past forms (*fin*, e.g. *czytam* 'I read') inflect for number and person and have aspect; gerunds (*czytanie* 'reading') inflect for number, case and negation and have gender and aspect, etc. Multiflex allows us to straightforwardly express such rigorous delimitation of classes according to the criteria of morphosyntactic homogeneity. Note also the richness of the gender category (*Gen*), which admits nine values[9]: 3 masculine (*m1*, *m2* and *m3*), one feminine (*f*), two neuter (*n1* and *n2*), and three plural (*p1*, *p2* and *p3*) genders. Adjectives (*adj*) inflect for number *(Nb,⟨var⟩)*, gender, case, and degree, while nouns (*subst*) inflect for number and case, and — contrary to what is admitted in Serbian — have a fixed gender *(Gen,⟨fixed⟩)*.

The application of Multiflex to the description of Polish urban proper names (cf. Section 3.5) required the introduction of two types of non-inflectional categories. **Graphical categories** handle morpho-graphical issues, such as letter case (*LetterCase*), initialisms and acronyms (*Init*). For instance, features $first\_upper$ and *dot*3 applied to the lemma *generał* 'general' produce the uppercase initial *Gen.* used e.g. in abbreviated street names. **User-defined categories** (*EXTRA CATEGORIES*) can bee freely chosen for a particular application, for instance to distinguish official (*offic*), neutral (*neut*) and spoken (*spok*) variants.

Segmentation of a Compound Lemma

The segmentation of a multi-word lemma into elementary lexical units is delegated to an underlying module for single words. In particular, the role of non-alphabetical characters, as well as the possibility of dividing a contiguous sequence of letters into several units, may depend on the language studied, as well as on the morphological model chosen. For instance, in Figure 3.4

[9]Other approaches to Polish morphology estimate the number of genders at five (Przepiórkowski, 2004), six (Szpakowicz, 1986; Vetulani et al., 1998), eight (Woliński, 2001) or eleven (Jassem, 1996).

CATEGORIES		EXTRA CATEGORIES		GRAPHICAL CATEGORIES		CLASSES	
Nb :	sg, pl	Usage:	⟨E⟩,	LetterCase:	same,	subst:	(Nb,⟨var⟩),(Case,⟨var⟩),
Case:	nom, gen, dat,		offic,		all_lower,		(Gen,⟨fixed⟩)
	acc, inst, loc,		neut,		all_upper,	depr:	(Nb,⟨var⟩),(Case,⟨var⟩),
	voc		spok		first_upper,		(Gen,⟨fixed⟩)
Gen:	m1, m2, m3, f,				...	adj:	(Nb,⟨var⟩),(Case,⟨var⟩),
	n1, n2, p1, p2,			Init:	⟨E⟩,dot,no_dot,		(Gen,⟨var⟩),(Deg,⟨var⟩)
	p3				dot2,no_dot2,	ppron12:	(Nb,⟨fixed⟩),(Case,⟨var⟩),
Pers:	pri, sec, ter				dot3,no_dot3,		(Gen,⟨var⟩),(Pers,⟨fixed⟩),
Deg:	pos, com, sup				...		(Accent,⟨var⟩)
Asp:	imperf, perf					ppron3:	(Nb,⟨var⟩),(Case,⟨var⟩),
Neg:	aff, neg						(Gen,⟨var⟩),(Pers,⟨fixed⟩),
Accent:	akc, nakc						(Accent,⟨var⟩),(Postprep,⟨var⟩)
Postprep:	praep, npraep					fin:	(Nb,⟨var⟩),(Pers,⟨var⟩),
Accom:	congr, rec						(Asp,⟨fixed⟩)
Agglt:	nagl, agl					ger:	(Nb,⟨var⟩),(Case,⟨var⟩),
Vocal:	wok, nwok						(Gen,⟨fixed⟩),(Asp,⟨fixed⟩),
							(Neg,⟨var⟩)
						pcon:	(Asp,⟨fixed⟩)
						pant:	(Asp,⟨fixed⟩)
						...	

Figure 3.3: Extended morphological model for Polish in *Multiflex*

spaces and punctuation characters are considered as components on their own. The component *bonhomme* is seen either as a unique token (*$1*), since it contains no separator, or as a double token (*$1* and *$2*), because it inflects like a typical French *Adj Noun* compound: *bonshommes de neige*. The sequence *1920* is considered either as one lexical unit (*$7*), since it consists of a contiguous sequence of digits, or as a four-unit compound (*$7* through *$10*), because it corresponds to a complex numeral.

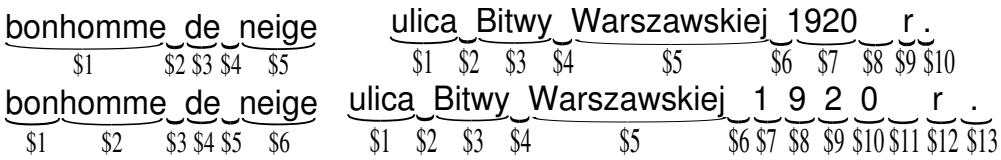


Figure 3.4: Two possible segmentations of compound lemmas *bonhomme de neige* 'snowman' in French and *ulica Bitwy Warszawskiej 1920 r.* 'Warsaw Battle 1920 Street' in Polish

Annotation of a Compound Lemma

Once the lemma has been segmented, it is necessary to provide the morphological annotation for each unit which can possibly inflect during the inflection of this compound lemma. The annotation contains the unit's lemma, morphological features and any other information necessary to generate other inflected forms of the same unit. For instance in Figure 3.5 the unit *vive* is the feminine singular form of the lemma *vif*, whose inflection paradigm is identified by code *A38*. The identifiers for inflectional categories and values in an annotation are those defined in the language model in Figure 3.2.

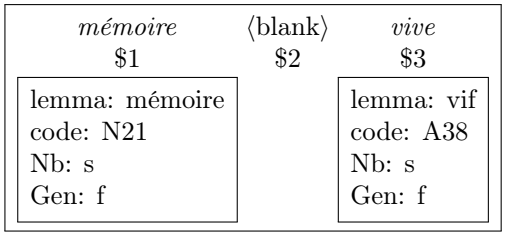


Figure 3.5: Lemma annotation for *mémoire vive* 'random access memory' in French

3.3.2 The Formalism

The Multiflex' formalism proposed in (Savary, 2005) was enhanced with new features and operators. The inflectional paradigm assigned to a multi-word lemma is represented by a graph which shows how single units are combined in order to inflect the MWU. This description is generation-oriented, i.e. is used to generate the set of the variants and inflected forms of a MWU, which can then, e.g., be straightforwardly searched for in a text.

Invariable Inflection

In the simplest case, no constituent of a multi-word lemma varies while the lemma is inflected. Example (3.1) shows a French compound, for which the singular and the plural form are identical.

(3.1)

	Variant	Lemma	Features	
(FR)	*porte-serviettes*	porte-serviettes	ms	'towel hanger' (masc. sing.)
	porte-serviettes	porte-serviettes	mp	'towel hangers'(masc. pl.)

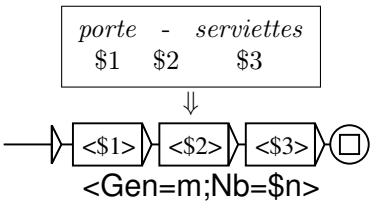


Figure 3.6: Lemma annotation and inflection graph for *porte-serviettes* 'towel hanger' in French

Figure 3.6 shows the annotation of this lemma, here a trivial one, as well as provides it with an inflection graph, here containing one path. A path consists of edges and boxes. It starts with the leftmost edge and ends with the final encircled box. The morphological information contained in the boxes refers to the constituents of the multi-word lemma. The information placed under a box refers to the morphological description of the resulting inflected form. In Figure 3.6 the three boxes refer to three constituents, here: *porte*, hyphen, and *serviettes*, with no morphological details, which means that they need to be recopied as such from the compound lemma. The information under the box is a set of category-value equations. A category may be assigned a fixed value from this category's domain. Here, category *Gen* is assigned the value *m*, which means that each generated form is in masculine gender, independently of the gender of its constituents. A category may also be assigned a *unification variable*. The variable may take any value from the category's domain. Here, variable *$n* takes any value listed for the category *Nb* in Figure 3.2 for French, i.e. singular (*s*) or plural (*p*). A complete exploration of the graph results in the set of all inflected forms of the MWUs, here the two forms listed in (3.1), annotated by their lemmas, inflectional classes and inflectional features.

Head Inflection and Value Inheritance

Most often, in order to inflect a MWU we need to inflect at least its headword, as in example (3.2). In the corresponding graph in Figure 3.7 a unification variable *$n* is assigned to the number of unit *$1*. It means that the first component may freely take both the singular and the plural number. Note that the description under the path contains two types of assignment. The latter (*Nb=$n*) indicates that the number of the multi-word unit varies but is determined by the same variable as the one of its first constituent (*$n*). The former (*Gen=$1.Gen*) mentions that the gender has a fixed value inherited from the first constituent's gender, as it appears in the compound lemma (here *f*).

(3.2)

	Variant	Lemma	Features	
(FR)	*machine à laver*	machine à laver	fs	'washing machine'
	machines à laver	machine à laver	fp	

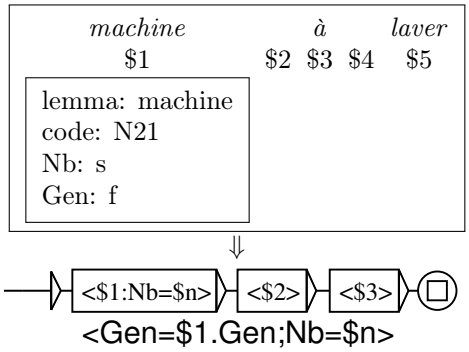


Figure 3.7: Inflecting the headword in *machines à laver* in French

Machine à laver inflects basically in the same way as many other French compounds of type *Noun Prep Noun* which are not necessarily of feminine gender. For instance *bonhomme de neige* (with its first segmentation in Figure 3.4) inflects in number also by putting its headword *bonhomme* into plural. All such examples may share the same inflection graph from Figure 3.7 due to the inheritance equation *Gen=$1.Gen*. Each time a graph is applied to a compound, the gender is inherited from the first constituent: feminine for *machine à laver, corde à sauter*, etc., and masculine for *bonhomme de neige, verre à vin*, etc.

Agreement

Many multi-word units are morphosyntactically compositional in the sense that they possess a regular syntactic structure, in which the headword inflects and its modifiers agree with it according to syntagmatic rules typical for the given language. For instance example (3.3) shows a Serbian compound adjective in which both adjectival components inflect and agree in number (*Nb*), case (*Case*), gender (*Gen*), animateness (*Anim*), degree (*Comp*), and determinedness (*Det*).

(3.3)

	Variant	Lemma	Features	
(SR)	*sam samcit*	sam samcit	aks1mg	'lit. alone small-alone= all alone'
	sama samcita	sam samcit	aks2mg	
	same samcite	sam samcit	aes2fg	
	...			

These 6 categories of inflection are expressed in the graph in Figure 3.8 by six unification variables *$n*, *$c*, *$g*, *$a*, *$m*, and *$d*. Each variable can take any value of the corresponding category domain. However it is common for components *$1* and *$3*, which means that its value must be unified for these two components in each multi-word inflected form. The resulting paradigm counts 77 inflected forms, which can all be described by a unique path. Without unification the graph would have to contain 77 separate paths.

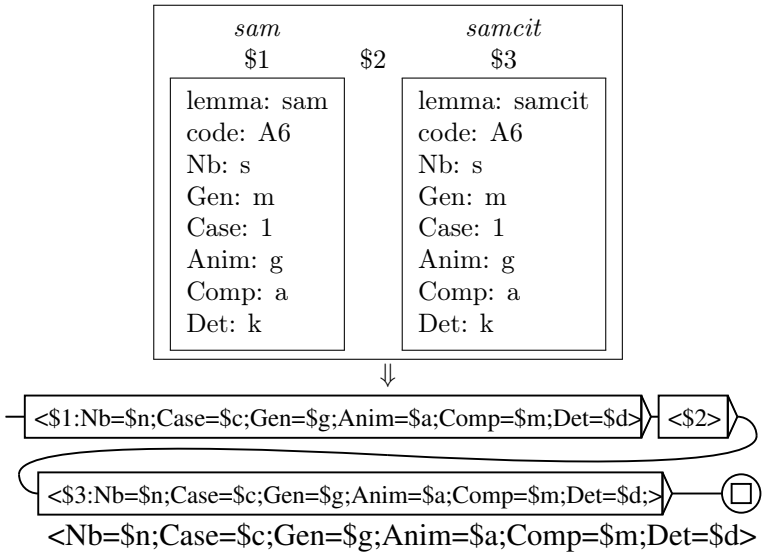


Figure 3.8: Agreement expressed by unification in *sam samcit* 'completely alone' in Serbian

Insertions, Omissions and Order Change

Numbering components within a MWU allows to express their omission, insertion and order change, which are frequent sources of orthographic and syntactic variation. For instance, Figure 3.9 shows a Serbian compound in which the hyphen may freely be replaced by a blank space or omitted, which results in squeezing the two remaining components into a contiguous sequence of letters.

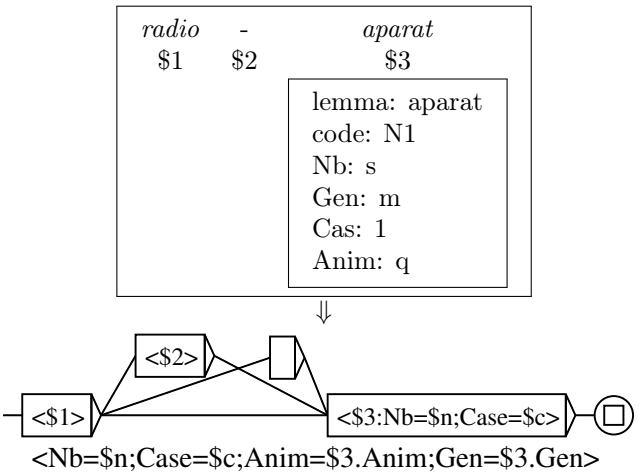


Figure 3.9: Lemma annotation and inflection graph for *radio-aparat* 'radio set' in Serbian

Thus, for each of the 16 possible inflection tags, three orthographic variants co-exist, as in

example (3.4) for masculine (*m*) singular (*s*) dative (*3*) non-animate (*q*). The whole paradigm counts 48 inflected forms, some of which are mentioned in (3.4).

(3.4)

	Variant	Lemma	Features	
(SR)	*radio-aparatu*	radio-aparat	ms3q	'radio set'
	radioaparatu	radio-aparat	ms3q	
	radio aparatu	radio-aparat	ms3q	
	...			

Components may change order within a variant for two reasons: (i) the language in question allows a relatively free word order, (ii) a head modifier takes a different grammatical structure. Such transformations are frequently accompanied by component omission or/and insertion. In example (3.5) the space between two nominal components may be freely omitted, which is represented by the middle path in the graph in Figure 3.10. Moreover, as shown in the lower path, component *$1* (here: *birth*) may shift to the final position, and then a new fixed unit *of* must be inserted.

(3.5)

	Variant	Lemma	Features
(EN)	*birth place*	birth place	s
	birthplace	birth place	s
	place of birth	birth place	s
	birth places	birth place	p
	birthplaces	birth place	p
	places of birth	birth place	p

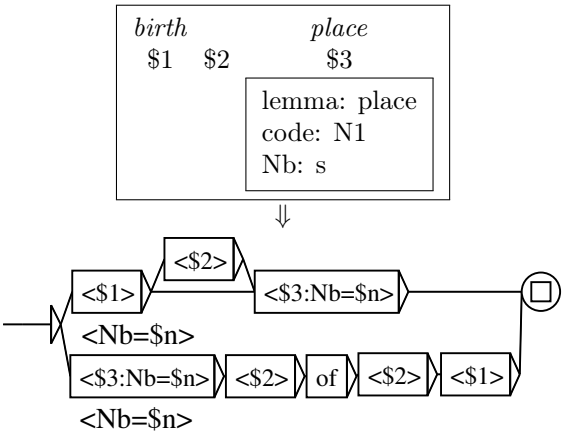


Figure 3.10: Syntactic variants of *birth place*

Empty Values

Empty morphological values, marked in the language files in Figure 3.2 by the ⟨*E*⟩ symbol, allow to model two kinds of situations: (i) a certain value is implicit if no other value of the same category appears, (ii) a value is irrelevant for a subset of inflected forms.

The former case is illustrated by Figure 3.11 and Figure 3.12. Both headwords *café* 'coffee' and *ponto* 'bridge' are nouns, thus according to Figure 3.2, they may inflect in number (*Nb*) and gradation (*Gr*: *cafezinho* 'small coffee', *pontinho* 'small bridge') and they have a fixed gender (*Gen*). Since many nouns and adjectives do not actually admit gradation, the positive gradation value is usually implicit, and only the diminutive (*D*), augmentative (*A*) and superlative (*S*) are

explicitly marked. Thus, if no gradation value appears for a noun or an adjective, it is supposed to be positive.

When a headword admits gradation the MWU containing it may do alike, as in example (3.6), but this is not necessarily the case, as in (3.7).

(3.6)

	Variant	Lemma	Features	
(PR)	*café com leite*	café com leite	ms	'coffee with milk'
	cafezinho com leite	café com leite	msD	'small coffee with milk'
	cafés com leite	café com leite	mp	
	cafezinhos com leite	café com leite	mpD	

(3.7)

	Variant	Lemma	Features	
(PR)	*ponto de água*	ponto de água	ms	'aqueduct'
	**pontinho de água*	ponto de água	msD	'*small aqueduct'
	pontos de água	ponto de água	mp	
	**pontinhos de água*	ponto de água	msD	

The corresponding inflection graph in Figure 3.11 allows for gradation due to the unification variable *$gr* which is propagated to the whole compound. The graph in Figure 3.12 mentions no information on gradation, which means here that this category is implicitly fixed to the positive value. The resulting inflected forms are consequently marked for diminutive and unmarked for the positive gradation value, as specified in (3.6) and (3.7).

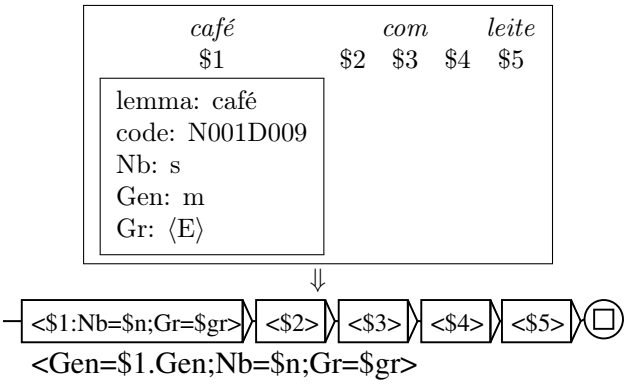


Figure 3.11: Empty gradation value in inflection of *café com leite* 'white coffee' in Portuguese

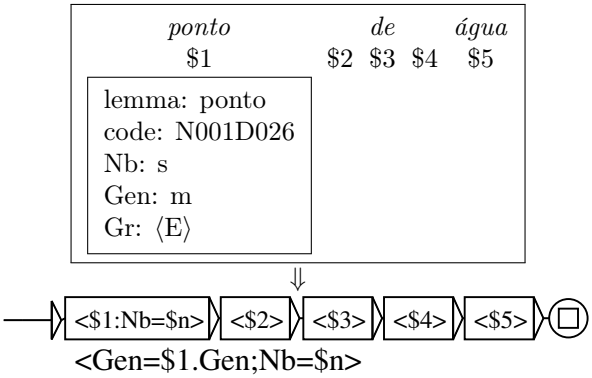


Figure 3.12: Empty gradation value fixed for *ponto de água* 'aqueduct' in Portuguese

A feature may be irrelevant for a certain subset of inflected forms if the morphological model admits inflection paradigms which do not respect the Cartesian-product rule. In other

words, some inflection class inflects in several categories but some combinations of values are not allowed. In the morphological model of French shown in Figure 3.2 the past participle (K) forms of verbs (e.g. *vu, vus, vue*, and *vues* for the verb *voir*) are included in the inflection paradigms of those verbs. Thus, each verb is supposed to inflect in tense (*Tense*), person (*Pers*), number (*Nb*) and gender (*Gen*). Since the participle forms actually inflect like adjectives, the person category is not relevant for them, while the gender category is relevant only for them. Moreover, the infinitive (W) is a particular form for which only the tense inflection is relevant. Due to all those restrictions the number, gender and person category are required to admit empty values.

Figure 3.13 shows a compound verb in French in which the prefix *sous* is uninflected and inseparable from the head verb *entendre*. While exploring the corresponding graph Multiflex generates all possible combinations of values for tense, person, gender and number, i.e. $11 * 4 * 3 * 3 = 396$ combinations, however only 51 of them are allowed by the single words' module for *entendre*. In particular, example (3.8) lists the resulting infinitive (W), one indicative (I) and four participle (K) forms, together with their annotations.

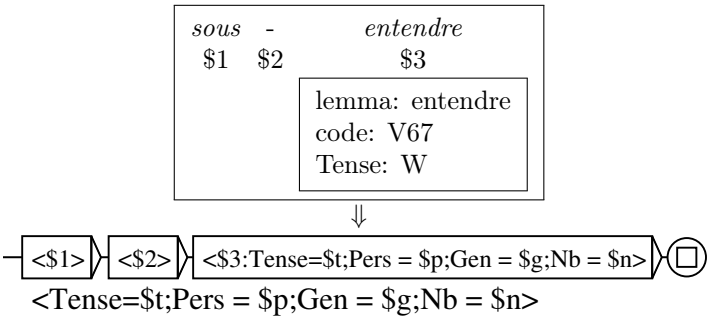


Figure 3.13: Lemma annotation and inflection graph for *sous-entendre* 'have sth in mind' in French

(3.8)

	Variant	Lemma	Features	
(FR)	*sous-entendre*	sous-entendre	W	'have sth in mind'
	sous-entendais	sous-entendre	I1s	
	sous-entendu	sous-entendre	Kms	
	sous-entendus	sous-entendre	Kmp	
	sous-entendue	sous-entendre	Kfs	
	sous-entendues	sous-entendre	Kfp	
	...			

Head Shifting

In MWUs of a more complex structure, ellipsis is a frequent syntactic transformation. With no change in meaning some non-essential modifiers or complements may be omitted. Sometimes the headword itself is left out, and then other components usually shift to the head position, as in example (3.9).

(3.9)

	Variant	Lemma	Features	
(DE)	*Organisation der Vereinten Nationen*	Organisation der Vereinten Nationen	neF aeF deF geF	'United Nations Organisation'
	Vereinte Nationen	Organisation der Vereinten Nationen	nmF amF	'United Nations'
	Vereinten Nationen	Organisation der Vereinten Nationen	nmF amF dmF gmF	'United Nations'

In the corresponding graph in Figure 3.14 the upper path covers all forms in which the headword *Organisation* is the only one to inflect, while the lower path describes the elliptic variant with the final component *$7* taking the role of the headword. Thus, the initially singular-only lemma gets transformed into a plural variant, as shown in (3.9). Note that on the lower path the adjectival component *$5* inflects freely in determinedness. Therefore, in the complete paradigm each elliptical variant appears twice in nominative and accusative: with the adjective's determined and undetermined form. Both variants obtain the same morphological tag.

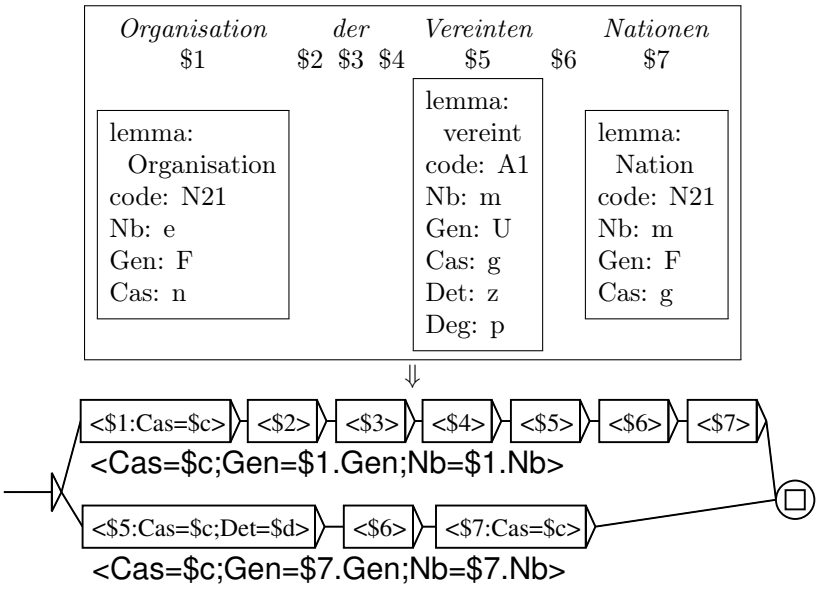


Figure 3.14: Head shifting in *Organisation der Vereinten Nationen* 'United Nations Organisation' in German

Graphical Variation and User-Defined Categories

Names of persons are subject to rich morphosyntactic variation. As shown in example (3.10), a given name (*Jan*) can be transformed to an initial or omitted, and a surname (*Rodowicz*) can be preceded or followed by a pseudonym (*„Anoda"*) with or without quotes. Additionally, user-defined categories (cf. Section 3.3.1) can distinguish particular forms e.g. for pragmatic purposes. Here, the full form *Jan Rodowicz „Anoda"* is marked as the official form (*offic*), while the surname *Rodowicz* alone is preferred for generation in neutral contexts (*neut*) and in speach (*spok*).

(3.10)

	Variant	Lemma	Features
(PL)	*Jan Rodowicz „Anoda”*	Jan Rodowicz „Anoda”	sg:nom:m1:**offic**
	Jan Rodowicz Anoda	Jan Rodowicz „Anoda”	sg:nom:m1
	Jan „Anoda” Rodowicz	Jan Rodowicz „Anoda”	sg:nom:m1
	J. Rodowicz „Anoda”	Jan Rodowicz „Anoda”	sg:nom:m1
	J. Rodowicz	Jan Rodowicz „Anoda”	sg:nom:m1
	„Anoda” Rodowicz	Jan Rodowicz „Anoda”	sg:nom:m1
	Rodowicz	Jan Rodowicz „Anoda”	sg:nom:m1:**spok**
	Rodowicz	Jan Rodowicz „Anoda”	sg:nom:m1:**neut**
	...		

The corresponding graph in Figure 3.15 uses the graphical category *Init* with its value *dot* in order to cover the initial *J.* and the pragmatic features appear under relevant paths (factorized by the alternative operator '|'). The resulting inflection paradigm contains 126 variants.

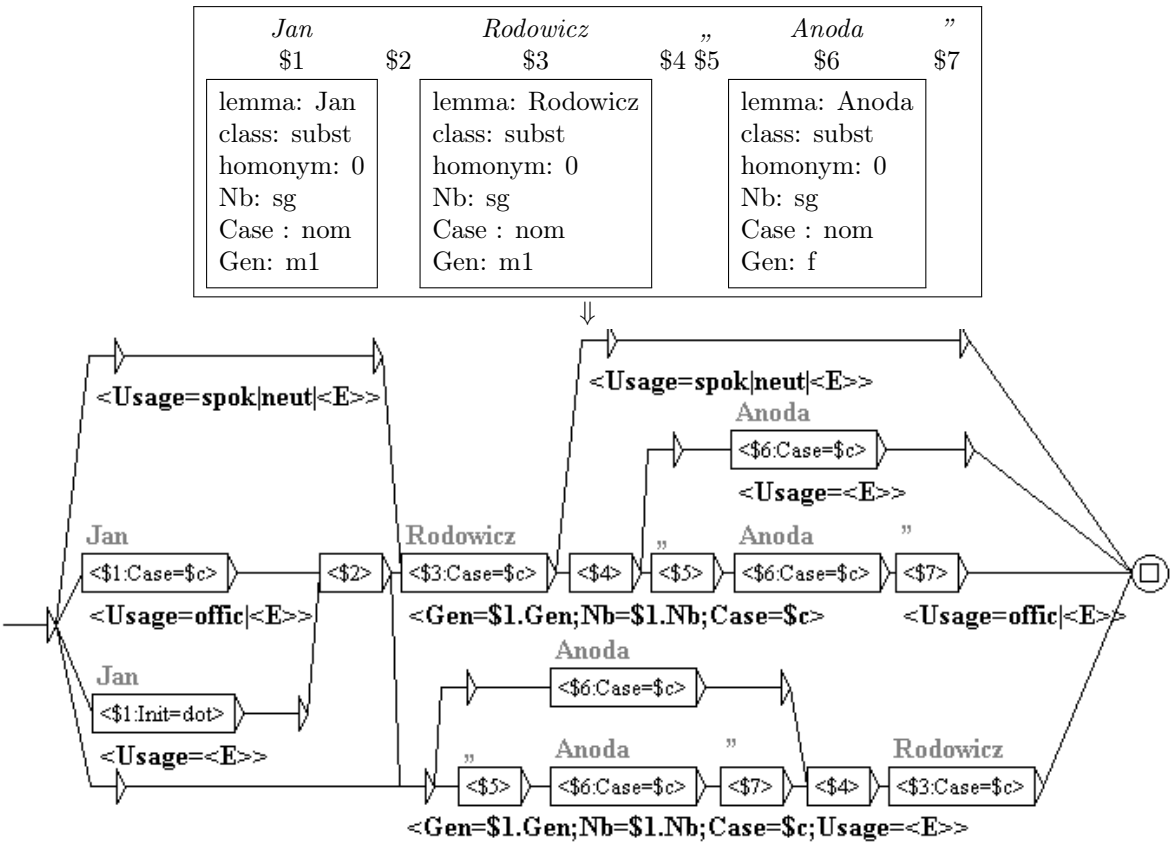


Figure 3.15: Graphical and user-defined features in *Jan Rodowicz „Anoda”* in Polish

Inserting External Elements

A MWU may be subject to insertion of external items which do not appear in its lemma. In the simplest case, a graphical variation may require an insertion of a new word separator, as shown in Figure 3.9, p. 52. In more complex situations, insertions of external lexemes may occur in nominal, adjectival and adverbial compounds mainly in stylistically marked cases, as in example (3.11). Since this process seems rather productive, an exhaustive representation of such insertions at the level of individual MWUs is not the best solution.

(3.11) *były to bajońskie, że tak powiem, sumy* 'these were gargantuan, so to say, amounts'

It is however sometimes necessary to introduce a particular external element in a particular MWU variant, for instance in one of the pragmatic variants addressed in the previous section. If inflected forms and variants are produced for the sake of future speech generation (as was the case in the SAWA e-dictionary of urban proper names discussed in Section 3.5) numerals appearing in Arabic or Roman digits should preferably be spelled out, as in example (3.12).

(3.12)

	Variant	Lemma	Features
(PL)	*Mieszko I*	Mieszko I	sg:nom:m1:offic
	Mieszko ***Pierwszy***	Mieszko I	sg:nom:m1:**spok**
	Mieszko I	Mieszko I	sg:nom:m1:neut
	Mieszko	Mieszko I	sg:nom:m1
	Mieszka I	Mieszko I	sg:nom:m1:offic
	Mieszka ***Pierwszego***	Mieszko I	sg:nom:m1:**spok**
	Mieszka I	Mieszko I	sg:nom:m1:neut
	Mieszka	Mieszko I	sg:nom:m1
	...		

The current morphological generators, however, fail to spell out the numeral like *pierwszy* 'the first' given the Roman numeral 'I'. This problem can be overcome by inserting a graph box with a new external lemma rather than a reference to a exiting constituent number, as in the upper path in Figure 3.16. Its morphological features, as in the case of a regular component, can still be instantiated ($Deg = pos$), inherited from ($Gen = \$1.Gen$) or unified with ($Case = \$c$) another MWU-inherent component.

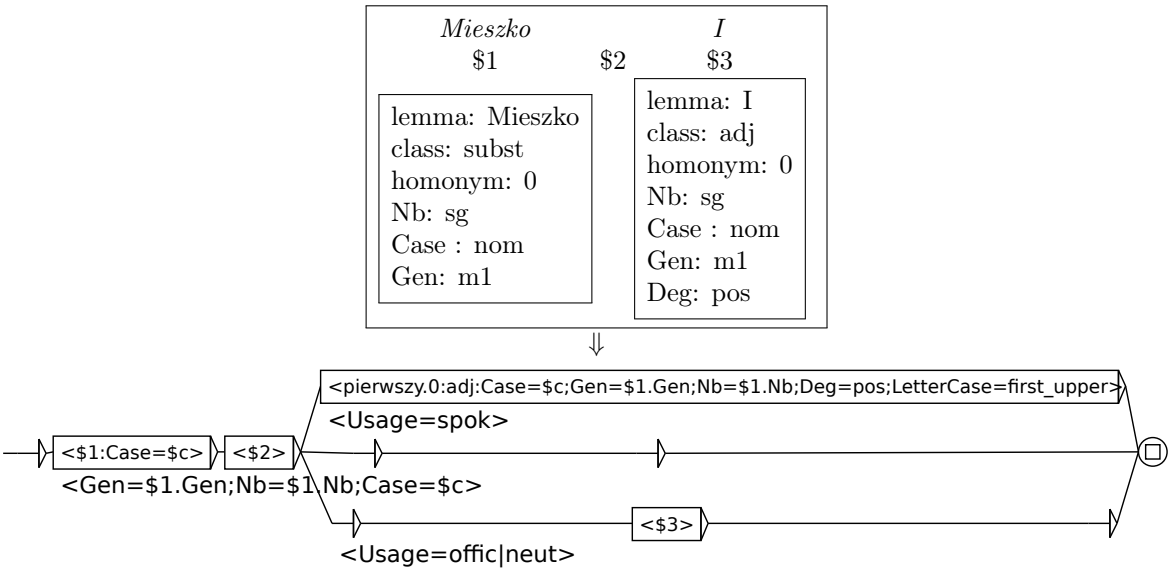


Figure 3.16: Inserting external elements in *Mieszko I* 'Mieszko the First' in Polish

Nesting

Some types of MWUs, in particular named entities and complex terms, often contain nested structures that are attested MWUs themselves. Example (3.13) shows an avenue name of type *Noun* $Noun_{gen}$, in which the patronym shown in Figure 3.15 occupies the genitive complement position. The headword *avenue* 'street' itself can be abbreviated (*al.*) or omitted. Thus, the total number of elliptical variants comes up to 48 in each of the 7 cases. Despite this complex

paradigm, when embedding is accounted for, as in Figure 3.17, the corresponding inflection graph is relatively simple. The patronymic complement is seen as a unique (although multi-word) constituent, with the sole constraint of being in genitive of the corresponding pragmatic variant. This graph combined with the one in Figure 3.15 results in the set of 336 inflected forms and variants.

(3.13)

	Variant	Lemma	Features
(PL)	*aleja Jana Rodowicza „Anody”*	aleja Jana Rodowicza „Anody”	sg:nom:f:**offic**
	al. Rodowicza	aleja Jana Rodowicza „Anody”	sg:nom:f:**neut**
	Rodowicza	aleja Jana Rodowicza „Anody”	sg:nom:f:**spok**
	aleja Jana Rodowicza Anody	aleja Jana Rodowicza „Anody”	sg:nom:f
	aleja J. „Anody” Rodowicza	aleja Jana Rodowicza „Anody”	sg:nom:f
	al. Jana Rodowicza	aleja Jana Rodowicza „Anody”	sg:nom:f
	J. „Anody” Rodowicza	aleja Jana Rodowicza „Anody”	sg:nom:f
	...		
	'Jan Rodowicz „Anoda” Avenue'		

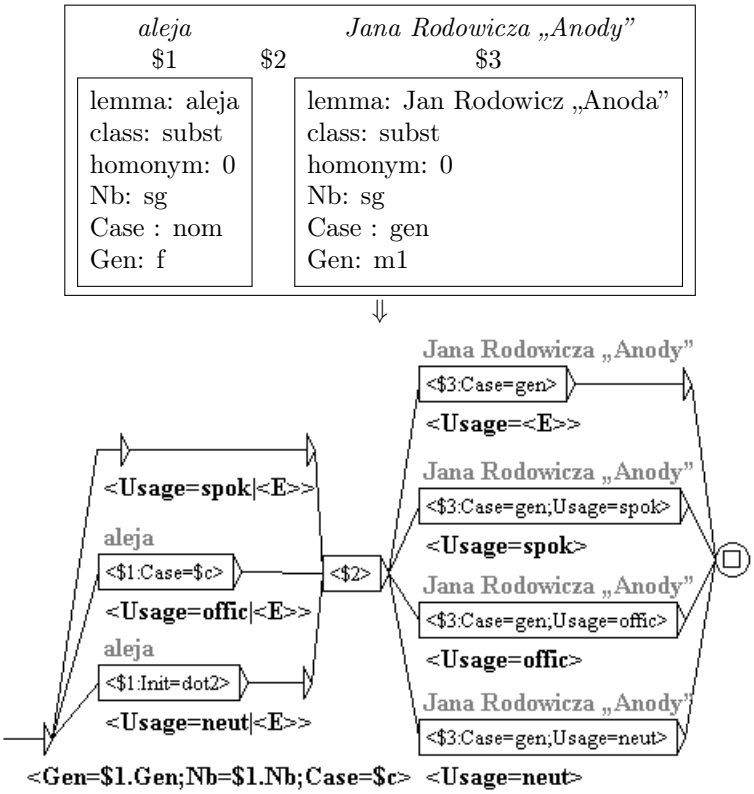


Figure 3.17: Embedded compounding in *aleja Jana Rodowicza „Anody”* in Polish

Expressing the embedding process explicitly is usually more elegant and convenient than providing a flat representation. If the same toponym were seen as a flat sequence of words it would contain nine constituents. Its inflection graph would show a complex set of paths combining full and elliptical versions of both the head part and the modifier. Moreover, the 126 patronym variants, some of which are shown in example (3.10), would have to be redundantly represented in each graph referring to the same patronym, e.g. *Nagroda im. Jana Rodowicza „Anody”* 'Jan Rodowicz „Anoda” Price'.

3.3.3 Interoperability

The morphosyntactic description of MWUs in Multiflex is based on a 'two-layer approach'. Single words are described first, then each inflected multi-word form is seen roughly as a particular combination of the inflected forms of its components. Numerous morphological models and tools for single words have been developed for decades. Rather than impose its own uniform model, Multiflex is designed so as to be able to collaborate with any external morphological module for single words, further called the *underlying module*, as soon as three interface constraints are observed.

Firstly, the underlying module and Multiflex must share the same morphological model (up to identifier replacement), described as in Figure 3.2. Different models are possible for the same natural language. For instance, the model admitted in the previous sections for French suggests that participle forms, inflecting in gender, belong to the inflectional paradigms of verbs, as discussed in section 3.3.2. In a different language model, the past participle form could be seen as a result of a derivational process (producing an adjective from a verb). Then the gender category would no longer be relevant for verbs and would not need to admit an empty value.

Secondly, the underlying module should provide a clear-cut definition of a token boundary. Multiflex imposes virtually no constraint on this definition. For instance, all four segmentations in Figure 3.4 can be admitted. Figure 3.18 shows the annotations and inflection graphs for two possible segmentations of the French compound discussed in section 3.3.1. Both segmentations result in the same inflectional paradigm shown in example (3.14).

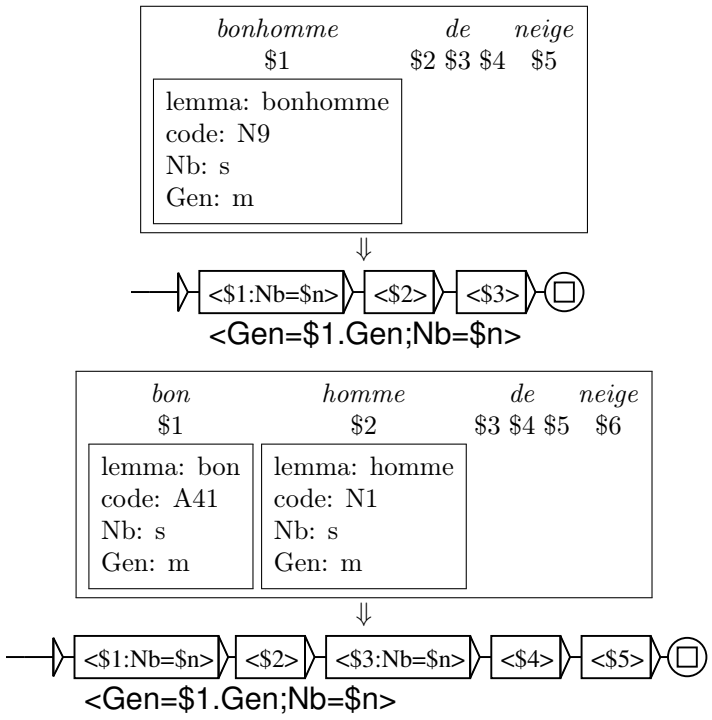


Figure 3.18: Two possible token boundaries and the corresponding lemma annotations and inflection graph for *bonhomme de neige* 'snowman' in French

(3.14)

	Variant	Lemma	Features	
(FR)	*bonhomme de neige*	bonhomme de neige	ms	'snowman'
	bonshommes de neige	bonhomme de neige	mp	

Thirdly, the underlying module should generate on demand particular inflected forms for the tokens it has itself defined. More precisely, given a lemma and a morphological tag, it should

generate all inflected forms of the lemma corresponding to the tag. For instance, if the first segmentation in Figure 3.4 is admitted, the underlying module should be able to provide the plural form *bonshommes* of the lemma *bonhomme*. With the second segmentation, the plural forms *bons* and *hommes* of lemmas *bon* and *homme* need to be available. The two corresponding descriptions of this compound are shown in Figure 3.18.

3.3.4 Complexity

The machinery behind Multiflex is based on a depth-first search (DFS) exploration of the minimal finite-state machine compiled from an inflection graph. Recall that the classical DFS algorithm complexity in time[10] is $O(|E|)$, where E is the set of graph's edges (Aho et al., 1980). Each node has to be visited only once and the results of processing a node are stored in case it is reached again via different incoming edges.

Multiflex boxes (which replace edges in the graph implementation stemming from the Unitex system[11]), however, have a rather rich semantics, including unification. Consider a node n that has previously been visited via a path p_1 and is being revisited via a different path p_2. The effects of unification performed while traversing p_1 may be quite different from those obtained via p_2. Thus, the constraints imposed on the partial forms to be generated after n may be different in the context of p_2 than they were in the context of p_1. It is therefore hard to follow the classical DFS approach. We explore, instead, each individual path completely, regardless if some of its parts have previously been visited.

Let

- E be the MWU entry to be inflected,
- w – the number of components in E (including separators),
- p – the number of paths in the inflection graph,
- c – the maximum number of categories a class can inflect for,
- v – the maximum number of values in a the domain of an inflectional category,
- s – the maximum cost of generating a simple word form given its lemma and the desired inflection features.

Firstly, recall that Multiflex' formalism allows external units (not appearing in the graph's lemma) to be inserted within a variant but the number of such elements is restricted in practice to less than w. Thus, the length of each path in a graph is $2 \times w$ at most.

Secondly, note that whenever E contains nested MWUs, the exploration of its graph is equivalent to flattening its structure and to replacing boxes of its graph with the graphs assigned to the nested components (provided that unification variables are appropriately renamed and components renumbered).

Each single word unit in E, and each of its external units, can be concerned by at most c unification variables instantiated to no more than v values. Thus, each unit has to be inflected into v^c forms at most. All relevant forms of all components in E and of all its external units have to be combined, which leads to $v^{2\times c\times w}$ combinations at most. Thus, the time complexity of the graph exploration is of $O(p \times v^{2\times c\times w} \times s)$.

In practice, the values of the above parameters are largely constrained. As far as the three Polish MWU e-lexicons discussed in Section 3.5 are concerned, w is no larger than 12, p does not

[10] Whenever the set of edges is larger than the set of vertices, which is usually the case

[11] http://www-igm.univ-mlv.fr/ unitex/

exceed 70, c is bounded by 5, and v does not exceed 11. Parameter s depends on the underlying morphological module for simple words.

3.3.5 Applications

Presently, Multiflex has been successfully interfaced with two underlying morphological modules for simple words stemming from *Unitex* and *Morfeusz*.

Unitex (Paumier, 2008) is a MWU-aware multilingual corpus processor containing DELA-type (Courtois & , eds.) modules for over a dozen European and Asian languages. Models of all languages but Polish in Figure 3.2 stem from this tool. Multiflex is fully integrated in this software as a module for an automatic generation of electronic lexicons of compound inflected forms (the so-called DELACF) which are matched against a corpus during the process of morphological analysis. The integration was performed in the framework of the French Outilex project. Both tools are distributed under the **LGPL license**, which allows in particular their free distribution and modification. Multiflex interfaced with submodules of Unitex is also a part of an encoding support software ***WS4LR*** (Krstev et al., 2006a), later renamed as ***LeXimir*** (Krstev et al., 2013), developed for Serbian but applicable to other languages with DELA-like electronic lexicons. It allows an automated controlled encoding of various linguistic resources such as morphological dictionaries, aligned corpora, wordnets, etc. It contains, notably, facilities for rule-based automatic Multiflex graph prediction, which speed up the lexicographer's work: 58% to 86% of new incoming MWUs are automatically assigned to correct inflection graphs.

Morfeusz (Woliński, 2006) is a morphological analyzer of Polish based on a large inflectional dictionary represented as a relational database (Woliński, 2009). It has been enlarged with a generation module in view of its interfacing with Multiflex. The Multiflex-Morfeusz integration (Savary et al., 2009) was achieved within the French-Polish *Polonium* [12] project and further enhanced within a nationally funded Polish project, a spin-off of the European LUNA[13] project aiming at spoken dialog corpus annotation. One of the crucial challenges identified by LUNA is the rich morphosyntactic variability of proper names in spoken dialogs (Mykowiecka et al., 2008). A dictionary creation tool ***Toposław*** (Marciniak et al., 2009b; Sikora & Woliński, 2009) containing the Multiflex-Morfeusz suite addresses this issue. Multiflex graphs allow to conflate orthographic, inflectional, syntactic and partly semantic variants within one paradigm, as illustrated in example (3.13) and in Figure 3.17. The morphological annotation of MWU's components is automated, search, matching and debugging functions facilitate graph management, and versatile MWU filters provide convenient ways of verifying and transforming multiple lexicon entries at a time. Additionaly, lexicon entries can be assigned to a customisable domain-dependent ontology of concepts. Toposław was intensively used in lexicographic work resulting in three e-dictionaries, SAWA, SEJF and SEJFEK (see Sction 3.5).

The precision and interoperability of Multiflex allows it to answer many other potential needs including: (i) development and enrichment of multilingual linguistic resources, (ii) MWU-aware morphosyntactic analysis of texts, (iii) enhancement of various NLP applications, such as information extraction, text classification, question answering, machine translation, etc., due to its ability to conflate different surface realizations of the same underlying concept.

3.4 Morphosyntactic Non-Compositionality of MWUs

In (Savary et al., 2007) we study phenomena of morphosyntactic non-compositionality in MWUs in French, Serbian and Polish, and of their representation in Multiflex' formalism. They are due

[12] `http://www.info.univ-tours.fr/~savary/Polonium/Polonium.html`
[13] `http://ist-luna.eu`

to:

- Exocentricity, i.e. missing headword, as in example (3.1).
- Agreement irregularities, as in (3.15), where the modifier *grand* may or may not be inflected for number, and may disagree in gender with the head noun.

(3.15)

	Variant	Lemma	Features	
(FR)	*grand-mère*	grand-mère	fs	'grandmother'
	grand-mères	grand-mère	fp	
	grands-mères	grand-mère	fp	

- Defective paradigms, as in (3.16), where only the plural forms exist although the headword admits a singular form

(3.16)

	Variant	Lemma	Features	
(PL)	*zimne nogi*	zimne nogi	pl:nom:f	'lit. 'cold legs' = a dish consisting of meat and jelly'
	**zimna noga*	zimne nogi	sg:nom:f	

- Coordinated structures, as in (3.17), where the compound's inflection features may or may not be inherited from one of its constituents, and both constituents agree in case but not necessarily in gender.

(3.17)

	Variant	Lemma	Features	
(SR)	*alfa i omega*	alfa i omega	s1f	'alpha and omega'
(PL)	*Adam i Ewa*	Adam i Ewa	pl:m1:nom	'Adam and Eve'

Further examples of non-compositionality managed within Multiflex come the lexicographic work on Polish general-language compounds, which led to the construction of *SEJF*, an e-dictionary of nominal, adjectival and adverbial compounds (cf Section 3.5). Czerepowicka & Kosek (2011) showed some interesting problems including defective paradigms, free word order, foreign words and gender fluctuation. Example (3.18), according to these authors, shows an idiomatic compound *czerwone pająki* 'red spiders' in a grammatically compositional context: the compound inherits its masculine animate gender from its head noun *pająk* 'spider'. In example (3.19), however, the compound takes its *natural gender* (masculine human) even if *pająk* never appears in this gender as an individual word.

(3.18) *[. . .] ustroju narzuconego części Europy przez* ***czerwone pająki*** *z Brukseli [. . .]*
'the regime imposed to a part of Europe by [red spiders]$_{m2}$ (post-communists) from Brussels'

(3.19) *[. . .] głosowałem na tych* ***czerwonych pająków*** *[. . .]* 'I voted for those [red spiders]$_{m1}$'

No entirely satisfactory solution seems to exist for cases of this type. The noun *pająk* could be included as a separate entry in the e-dictionary of simple words with human masculine gender and a restriction to the idiomatic usage only. A closed list of such cases already appears in Morfeusz, e.g. *dziadek* 'grandpa' is represented by two entries: a regular masculine human entry, and a masculine inanimate entry restricted to the idiom *dziadek do orzechów* 'literally: nut grandpa = nutcracker'. The authors of SEJF have chosen an intermediate solution requiring no intervention into Morfeusz. It exploits the syncratic nature of masculine forms, as shown in Figure 3.19. While the upper path generates all compositional forms with the compound's gender equal to *m2* (masculine animate), the bottom path represents the idiosyncratic accusative plural

form in exceptionally non-compositional gender *m1* (masculine human). This artifice is possible due to the fact that plural forms are identical in masculine animate genitive and in masculine human accusative.

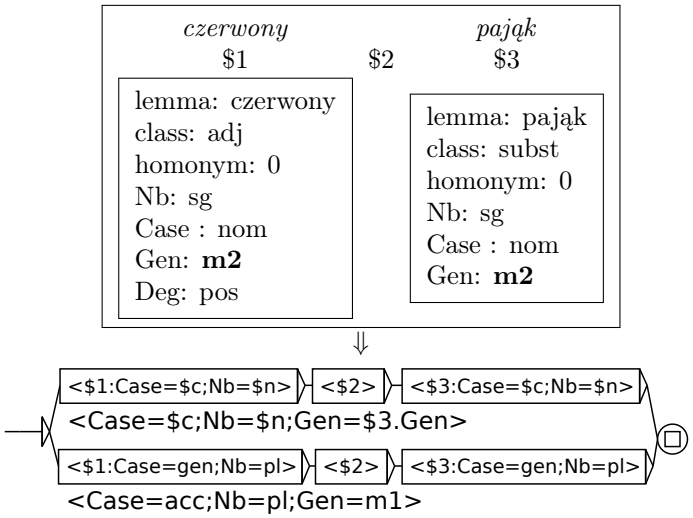


Figure 3.19: Gender fluctuation in *czerwony pająk*, literally 'red spider' = 'post-communist'

Some proper nouns and named entities are also concerned by specific non-compositionality issues. For instance appositions and coordinations belong to frequent constructions with complex agreement rules. Example (3.20) discussed in Section 3.5 is a classical apposition (with an additional acronym), in which both nouns agree in number and in case but not in gender. If this agreement pattern is considered as characteristic to appositions, many productive structures among names of institutions have to be considered non-compositional due to non-agreement in case, as in examples (3.21)–(3.22).

(3.20) *Bank*$_{\text{sg:nom:}\mathbf{m3}}$ *BPH* [*Spółka Akcyjna*]$_{\text{sg:nom:}\mathbf{f}}$,
Banku$_{\text{sg:gen:}\mathbf{m3}}$ *BPH* [*Spółki Akcyjnej*]$_{\text{sg:gen:}\mathbf{f}}$, ... 'Bank BPH Joint Venture'

(3.21) *Alianz*$_{\text{sg:nom:}\mathbf{m3}}$ *Polska*$_{\text{sg:nom:}\mathbf{f}}$, *Alianzu*$_{\text{sg:}\mathbf{gen}\text{:}\mathbf{m3}}$ *Polska*$_{\text{sg:}\mathbf{nom}\text{:}\mathbf{f}}$, ...
'Alianz Poland' (a bank name)

(3.22) *Widzew*$_{\text{sg:nom:}\mathbf{m3}}$ *Łódź*$_{\text{sg:nom:}\mathbf{f}}$, *Widzewa*$_{\text{sg:}\mathbf{gen}\text{:}\mathbf{m3}}$ *Łódź*$_{\text{sg:}\mathbf{nom}\text{:}\mathbf{f}}$, ... (a soccer club name)

3.5 Electronic Lexicons of Multi-Word Units

My studies on computational morphology of MWUs resulted in the construction of several electronic lexicons in three languages summarized in Table 3.1.

My first motivation for an inflection tool for compounds came from the multilingual corpus processor **Intex** (Silberztein, 1993a), and led to a prototype described in (Chrobot, 1998a). This prototype was successfully applied to the creation of two DELA-type electronic lexicons of English compounds: a lexicon of **general English compounds** with about 60,000 lemmas and 110,000 inflected forms (Savary et al., 1999), and a terminological **lexicon of complex terms in computer science** with 58,000 lemmas, 109,000 inflected forms (Chrobot, 1999). The first of these resources was distributed with *Intex* and its enhanced version NooJ and its enhanced open-source equivalent Unitex (cf. Section 3.3.5). The second resource was used in translation aid software *LexProCD Databank* within a prototype of a rule-based term extraction module (Chrobot, 1998b; Savary, 2001a).

Later on, Multiflex was interfaced with Unitex and embedded in LeXimir (see Section 3.3.5) and has been systemtically used to create an **MWU general-purpose electronic dictionary for Serbian** (Krstev et al., 2006b, 2011, 2013). This resource currently contains about 11,000 nominal and adjectival lemmas (including over 1,000 proper names) assigned to 115 inflection graphs, and yields over 204,000 inflected forms[14].

The Multiflex-Unitex suite was also applied to the construction of a similar resource for modern Greek (Foufi, 2013). The resulting e-dictionary comprises nominal lemmas of type *Adjectif Noun* and *Adjectif Adjectif Noun*. One of the interesting issues addressed here is the frequency of elliptical variation which transforms MWUs into single words, as well as additional ambiguity of simple words arising from this variability.

The Multiflex-Morfeusz framework embedded in Toposław (see Section 3.3.5) has been used for three different language resources for Polish described below.

SAWA[15], the **Grammatical Lexicon of Warsaw Urban Proper Names** (**S**łownik elektroniczny n**A**zewnictwa **WA**rszawy) (Marciniak et al., 2009a; Savary et al., 2009) is an electronic lexicon containing about 9,000 proper names of places related to the Warsaw transportation system, i.e. names of streets, squares, monuments, buildings, bus, tram and subway stops, etc., as well as names of persons to whom some objects (notably streets) are dedicated. A large majority (about 98%) of these names are MWUs, while only 2% correspond to simple words (e.g. *Bemowo*). Stylistically marked names such as (3.23), as well as previous names, notably those used before 1989, as in (3.24), are also included. The morphosyntax of names is described by over 450 Multiflex graphs, which allow an automatic generation of about 300,000 variants. Except for inflectional and syntactic variants, also pragmatic variants, necessary for text generation, are represented. For instance, example (3.25) shows an official variant used in official lists and documents, the neutral variant preferred in text generation and the spoken variant preferred for speech generation. The dictionary has been developed within a nationally funded Polish project, a spin-off of the European LUNA project (cf Section 3.3.5). The resource should further serve as a front end for a dialog system containing a model of Warsaw topography (streets, places, monuments, etc.) and transport (bus-stops, underground stations, etc.).

(3.23) Popular name: *Czterech Śpiących* 'The Four Sleeping ones'
Official name: *Pomnik Braterstwa Broni* 'the Monument of the Brotherhood in Arms'

(3.24) Former name: *aleja Świerczewskiego* 'Świerczewski Avenue'
Present name: *aleja Solidarności* 'Solidarity Avenue'

(3.25) Official variant: *ulica Bitwy Warszawskiej 1920 r.* 'Warsaw 1920 Battle Street'
Neutral variant: *ulica Bitwy Warszawskiej* 'Warsaw Battle Street'
Spoken variant: *ulica Bitwy Warszawskiej tysiąc dziewięćset dwudziestego roku* 'Warsaw nineteen twenty Battle Street'

SEJF[18], the **Grammatical Lexicon of Polish Phraseology** (**S**łownik **E**lektroniczny **J**ednostek **F**razeologicznych) (Graliński et al., 2010; Czerepowicka, 2011; Czerepowicka & Kosek, 2011; Czerepowicka, submitted) is an e-dictionary containing multi-word units of the general (non terminological) Polish language. It comprises about 3,200 multi-word lexemes, with the following distribution:

[14] According to a personal communication with Cvetana Krstev.
[15] `http://zil.ipipan.waw.pl/SAWA`
[16] `http://infolingu.univ-mlv.fr/DonneesLinguistiques/Dictionnaires/telechargement.html`
[17] `http://creativecommons.org/licenses/by-sa/3.0/`
[18] `http://zil.ipipan.waw.pl/SEJF`

Table 3.1: Electronic dictionaries of MWUs produced with Multiflex and its predecessors

Dictionary name	Language	MWU types	Lexicogr. framework	Dictionary size			Source code availability		
				Lemmas	Graphs	Forms	Lemmas	Graphs	Forms
English DELAC	English	general-purpose nouns	Intex	60,000	NA	110,000	no	NA	LGPL-LR[16]
Serbian DELAC	Serbian	general-purpose nouns, adjectives	LeXimir	11,000	115	204,500	from authors		
Greek DELAC	modern Greek	general-purpose A(A)N nouns	Unitex				from authors		
SAWA	Polish	urban proper names	Toposław	9,000	450	309,000	CC-BY SA[17]		
SEJF	Polish	general-purpose nouns, adjectives and adverbs	Toposław	3,200	140	68,000	CC-BY SA		
SEJFEK	Polish	economic nominal terms	Toposław	11,000	290	146,000	CC-BY SA		

- over 2,100 nominal compounds (e.g. *bajońskie sumy*, literally: 'Bayonne quantities' = 'gargantuan sum'),

- over 440 adjectival compounds (e.g. *prosty jak strzała* 'as straight as an arrow', *wprost proporcjonalny* 'directly proportional'),

- over 600 adverbial compounds (e.g. *chcąc nie chcąc*, literally: 'wishing, not wishing' = 'willy-nilly'),

- 43 others (e.g. *ni z gruszki, ni z pietruszki*, literally: 'neither from a pear, nor from parsley' = 'irrelevantly').

These lemmas, together with their associated 160 graph-based inflection paradigms, yield about 68,000 corresponding inflected forms. Some interesting non-compositionality issues addressed in this lexicographic study are discussed in Section 3.4.

SEJFEK[19], the **Grammatical Lexicon of Polish Economic Phraseology** (**S**łownik **E**lektroniczny **J**ednostek **F**razeologicznych z **EK**onomii) (Graliński et al., 2010; Savary et al., 2012b) is an electronic lexicon containing multi-word nominal terms of Polish economic and financial terminology. It contains over 11,000 multi-word nominal lexemes (e.g. *aktywne ryzyko płynności* 'active liquidity risk'), over 146,000 corresponding inflected forms (e.g. *aktywnego ryzyka płynności*), and 305 Multiflex inflection graphs. The high number of graphs results from a big variety of syntactic structures typical for technical terms, as well as from their high degree of variability (acronyms, ellipses, word order change, restrictions in number inflection, etc.). Example 3.26 shows a sample inflection paradigm of the lemma *Bank BPH Spółka Akcyjna* 'Bank BPH Joint Venture' containing an embedded lemma *spółka akcyjna* 'joint venture', and affected by acronymy and elliptical variation.

[19] `http://zil.ipipan.waw.pl/SEJFEK`

(3.26)

	Variant	Lemma	Features
(PL)	*Bank BPH Spółka Akcyjna*	Bank BPH Spółka Akcyjna	subst:sg:**nom**:m3
	Banku BPH Spółki Akcyjnej	Bank BPH Spółka Akcyjna	subst:sg:**gen**:m3
	Bankowi BPH Spółce Akcyjnej	Bank BPH Spółka Akcyjna	subst:sg:**dat**:m3
	...	...	...
	Bank BPH SA	Bank BPH Spółka Akcyjna	subst:sg:**nom**:m3
	Banku BPH SA	Bank BPH Spółka Akcyjna	subst:sg:**gen**:m3
	...	...	...
	Bank BPH S.A.	Bank BPH Spółka Akcyjna	subst:sg:**nom**:m3
	Banku BPH S.A.	Bank BPH Spółka Akcyjna	subst:sg:**gen**:m3
	...	...	...
	Bank BPH	Bank BPH Spółka Akcyjna"	subst:sg:**nom**:m3
	Banku BPH	Bank BPH Spółka Akcyjna	subst:sg:**gen**:m3
	...	...	...

'BPH Joint-Stock Bank'

The grammatical lexicons such as SAWA, SEJF and SEJFEK are currently oriented towards the generation of lists of unstructured forms and variants, which may later be applied e.g. in the process of a straightforward text search. This does not allow us to transmit the data about the internal, syntactic or semantic, structure of a recognized MWUs to further stages of linguistic processing. Therefore, first experiments were performed with transforming SEJFEK into a fully lexicalized shallow grammar ***SEJFEK4Spejd*** (Savary et al., 2012b). Each MWU grammatically annotated lemma was semi-automatically transformed into one grammar rule in the *Spejd*[20] formalism (Przepiórkowski, 2008; Zaborowski, 2012). The resulting grammar compiles into a cascade of regular rules so that the nesting structure of terms is preserved. An evaluation of SEJFEK both as a lexicon and as a grammar was performed on a manually annotated 220,000-token corpus of Polish economic Wikipedia articles. The two resources were applied by the *Spejd* engine to the unannotated version of the corpus, and the results were compared with its annotated version. Only 0.13 to 0.21% of all MWU terms recognized in the corpus were false positives, which shows a very good quality of both resources. The coverage was estimated in terms of *correctness* (the percentage of fully correctly recognized MWU terms) and *weak-correctness* (the percentage of fully correctly recognized one-token fragments of MWU terms). The values obtained for these measures were equal to 42% and 68%, respectively.

The construction of SEJF and SEJFEK was funded by the ERDF Nekst[21] project. The enhancement of SAWA, SEJFEK and SEJF, as well as making them available within the META-SHARE[22] exchange platform were funded by the European CESAR project[23].

A comparative study of the three Polish e-dictionaries, SAWA, SEJF and SEJFEK reveals interesting specificities, summarized in Table 3.2. MWU **nesting** is:

- virtually inexistent in general-language compounds,
- frequent in urban proper names, mostly due to people and places after whom urban objects are named, as in Figure (3.17),
- particularly prevalent in a terminological sublanguage (new terms are coined by extending the former ones).

The low lemma/graph and the high form/lemma ratios in SAWA indicate a high morphosyntactic **variability** in urban proper names, which notably possess numerous elliptical variants and

[20] `http://zil.ipipan.waw.pl/Spejd`
[21] `http://zil.ipipan.waw.pl/NEKST`
[22] `http://metashare.dfki.de/repository/search/`
[23] `http://www.meta-net.eu/projects/cesar`

Table 3.2: Statistics on Polish MWU dictionaries of different scopes

Dictionary name	Scope	Lemmas	Lemmas with nested MWUs	Lemmas per graph	Forms per lemma	Words per form
SAWA	urban toponyms	9,000	14%	19.8	35 nouns	4.6
SEJF	general language	3,200	0.0003%	22.7	12 nouns 100 adjectives	4.5
SEJFEK	economic terms	11,000	19%	38.7	13 nouns	4.1

acronyms, but which are also annotated with pragmatic labels, as shown in example (3.13), p. 59. Conversely, economic terms more frequently follow common inflection and variation rules and thus require an almost twice smaller amount of graphs (for a comparable amount of lemmas). Nominal MWUs both in general language and in economic sublanguage have less than 14 forms (2 numbers * 7 cases) per lemma on average, mostly due to defective inflection paradigms of type *singulare* or *plurale tantum*. Finally, the adjectival compounds have as many as 100 forms on average due to case (7 values), gender (9 values) and number (2 values) inflection (recall Figure 3.2 p. 48). Note that their inflected forms do not sum up to 126 because some features never combine, e.g. $p1$, $p2$ and $p3$ gender values are restricted to plural forms only. Finally, the average number of words per form is above 4 in each e-dictionary, and it is the highest in urban proper names.

SAWA, SEJF and SEJFEK also have an interesting distribution of different syntactic structures and morphosyntactic variability, as shown in Table 3.3. Binary **agreement structures** (in which two, possibly compound, components agree) constitute about 40% of all entries in each of the dictionaries, while the **government structures** (in which a head noun or a preposition governs the subordinate noun) are about twice more frequent in domain-specific vocabulary (SAWA and SEJFEK) than in general language. Proper names admit no variability in number, while more than one third of both general language and economic MWUs have both singular and plural forms. Variability in order of components is very rare in proper nouns and general language compounds but concerns at least 5% of economic terms.

3.6 Contributions and Perspectives

The contributions of Multiflex, as well as its accompanying research, to the field of NLP include: (i) a better understanding of the behavior of MWUs due to in-depth linguistic studies on their properties (Savary, 2008), (ii) a multilingual view on MWUs allowed by contrastive and complementary studies on different language families, Germanic, Romance and Slavic (Savary, 2000; Savary et al., 2007), (iii) a universal formalism for the high-quality lexical description of MWUs, (iv) an interoperable tool capable of integrating different methods of morphological processing of single words.

Multiflex is close to the international research community using the Unitex system as computational framework for the Paris LADL school's linguistic theory. Thus, the implementation of the Multiflex' formalism relies in particular on two modules adapted from Unitex: a user-friendly graph editor, and a generic finite-state library for binary representation and exploration of graphs. However, the semantics introduced in Multiflex' graphs is novel, although formally close to decorated RTNs (Blanc & Constant, 2005), regular expressions with feature structures (Drożdżyński et al., 2004), and flag diacritics (Beesley & Karttunen, 2003). Its implementation is based on an extensive recursion due to two factors: the depth-first search exploration of a

finite-state transducer behind a graph, and systematic instantiation of unification variables to all possible values from their respective category domains (Savary, 2009).

In Savary (2008) a large contrastive study of 11 lexical approaches to the inflection and variation of MWUs in 7 languages has been performed. It analyzes more than a dozen linguistic properties of MWUs such as exocentricity, orthographic variability, irregular agreement, defective paradigms, abbreviations, syntactic and semantic variants, sense computation, etc. It also considers desirable descriptive and computational facilities such as unification, non-redundancy, inflectional analysis and generation, encoding interface, etc. In the light of this study Multiflex belongs to the most expressive and effective tools along with *lexc* (Karttunen et al., 1992), *FASTR* (Jacquemin, 2001), and *HABIL* (Alegria et al., 2004). Its main drawbacks include the lack of modeling of derivational and semantic variants, and its inability to express dependencies existing between a described MWU and neighboring external elements. For instance, the German example (3.9) in section 3.3.2 fails to reflect the fact that the adjective component agrees in definiteness with the accompanying article:

(3.27) (DE) *die Vereinten Nationen*
Vereinte_ Nationen
**die Vereinte_ Nationen*

Since *Vereinte Nationen* and *Vereinten Nationen* obtain in (3.9) the same morphological tag in nominative and accusative (*nmF* and *amF*), there is presently no means to express such constraints.

Another possible extension concerns the morphological models of highly inflected languages such as Polish. As explained in Section 3.3.1, the Polish *flexemic* tagset behind the Morfeusz-Multiflex suite admits a delimitation of classes according to the criterion of homogeneous morphological behavior. Different *flexemes* denoting the same semantic entry are then grouped into *lexemes*. For instance, the lexeme *student* 'student' divides into two flexemes. The first one is the "neuter" noun (i.e. of class *subst* mentioned in Figure 3.3 p.49) inflecting for number and case and having the gender value *m1*. The other one is its depreciative (*depr*) form *studenty* appearing in plural nominative and vocative case only and having the gender value *m2*. Presently, Multiflex represents the morphological model in terms of flexemes and categories only. Links between flexemic classes and their corresponding classes of lexemes (Woliński, 2003) are not expressed. As a result, the inflection of an entry like in example (3.28) yields the neuter plural form (3.29) but not the stylistically marked form (3.30).

(3.28) wieczny(wieczny:adj:sg:nom:m1:pos) student(student:subst:sg:nom:m1), subst(NC-O_O-1) 'eternal student'

(3.29) wieczni studenci,wieczny student:subst:pl:nom:**m1**

(3.30) wieczn**e** studen**ty**,wieczny student:subst:pl:nom:**m2**

Resolving this issue in a general case is non-trivial. Different flexemes of a given lexeme have very different inflectional behavior, thus switching from one flexeme to another while staying within the same lexeme might be seen as a derivation rather than an inflection process. While formal modeling of nominal and adjectival derivation in Polish has already been addressed (Rabiega-Wiśniewska, 2006), automatic generation of inflected derivatives seems unresolved.

In the long run Multiflex needs to be enlarged to non-contiguous MWUs such as verbal multiword expressions, admitting insertions of free external elements. Sense calculation in MWUs, suggested e.g. by Copestake et al. (2002), might be another ambitious perspective.

As seen in the preceding sections, the encoding process under Multiflex consists in analyzing the MWUs one by one, annotating their possibly inflected components, attributing them already

existing inflection graphs, or creating new graphs. This process can be done either manually or within an automated encoding interface such as LeXimir (Krstev et al., 2006a) or Toposław (Sikora & Woliński, 2009). These graph and dictionary management tools integrating Multiflex highly facilitate the lexicographer's complex work on a new entry by: (i) an automatic lookup of a compound's constituents in the underlying morphological module for simple words, (ii) automatic generation of all resulting forms and variants. Moreover Toposław helps to organize complex and numerous graphs by means of: (iii) graph naming convention, (iv) graph debugging by highlighting paths corresponding to a generated morphological variant, (v) graph filtering based on the morphological characteristics of the entry that is being encoded, (vi) automated creation of new graphs (Woliński et al., 2009). The usability of this environment has been studied and compared to another tool dedicated to encoding Polish MWUs in (Graliński et al., 2010). LeXimir on its turn allows for prediction of inflection graphs for new incoming data due to rules defined on the basis of previously encoded MWUs (Krstev et al., 2010). The integration of a similar facility is planned for Toposław, too. In the long run, a corpus-based solution, possibly including machine learning, might predict inflection paradigms of MWUs on the basis of their occurrences in large texts.

Table 3.3: Distribution of syntactic structures and number/word order variability in Polish MWU e-lexicons. The following codes are used: substantive (S), substantive in genitive (S_{gen}), substantive in a case governed by the preposition (S_{gov}), adjective (Adj), and preposition ($Prep$).

	Syntactic structure	Variability in number	Variability in order	Examples SAWA	Examples SEJF	Examples SEJFEK	Percentage of entries SAWA	Percentage of entries SEJF	Percentage of entries SEJFEK
Agreement	S Adj Adj S	✓			*osoba prawna* *babie lato*	*spółka akcyjna* *agresywna [zmiana cen]*	0%	24%	23%
	S Adj Adj S			*ulica Żyzna* *Aleje Jerozolimskie*	*literatura piękna* *bajońskie sumy*	*[produkt narodowy brutto] realny* *wtórne [ryzyko płynności]*	29%	15%	10%
	S Adj Adj S		✓		*bicz boży* *hiobowa wieść*	*[dług ekonomiczny] użytkowy* *lokalne [dobro publiczne]*	0%	1%	5%
	S S			*Adam Mickiewicz* *generał [Józef Bem]*	*śmichy-chichy*	*kraj imitator, Wawel [Spółka Akcyjna]*	12%	0.2%	0.1%
Government	S S_{gen}			*aleja [Jana Rodowicza „Anody"]* *plac Defilad* *pomnik [generała [Józefa Bema]]*	*kwadratura koła* *gest Kozakiewicza*	*krzywa Beveridge'a, [ryzyko inwestycyjne] obligacji,* *demonetyzacja [zagranicznych [środków płatniczych]]*	24%	3%	13%
	S S_{gen}	✓			*prawo dżungli* *walki byków*	*centrum rozliczeń,* *[[czytnik elektroniczny][kodów kreskowych]],* *podstawa [wymiaru [składek [ubezpieczeń społecznych]]]*	0%	6%	12%
	S S_{gen} S_{gen}			*ulica Zachodu Słońca*		*częstotliwość dokonywania zakupu*	3%	0%	6%
	S Prep S_{gov}			*ulica Ku Wiśle* *Lasek na Kole*	*sztuka dla sztuki* *cuda na kiju*	*egzekucja z [wynagrodzenia za pracę]* *[poziom dobrobytu] w [skali krajowej]*	0.5%	2%	2%
	S Prep S_{gov}	✓			*dowcip z brodą*	*[cena dewizowa] w imporcie*	0%	3.5%	2%
Others				*Zamek Królewski w Warszawie*	*baby sitter* *sodoma i gomora*	*teoria powiązań pionowych i poziomych między firmami*	31.5%	45%	27%

Chapter 4

Compound Named Entities and Beyond

Proper names and, more generally, named entities (NEs), carry a particularly rich semantic load in each natural language text since they refer to persons, places, objects, events and other entities crucial for its understanding. Their central role in natural language processing (NLP) applications is unquestionable. They are good keyword candidates in automatic indexing and categorization of documents. They are subject to specific translation rules. They play key roles in information extraction and in question answering systems.

Similarly to multi-word expressions, named entities are hard to define. Ehrmann (2008) points out that these terminological problems stem from the very nature of NLP resulting from tensions between theoretical studies and strong applicative motivations on the one hand, and from different disciplines composing NLP on the other hand. She cites several NE definition attempts from both the onomasiological (from concepts to words) and the semasiological (from words to concepts) points of view, and she studies the linguistic foundations underlying NE-related studies such as the sense, the reference and the definite descriptions. She then proposes her own NLP-dedicated definition: *Etant donné un modèle applicatif et un corpus, on appelle entitée nommée toute expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le corpus.* 'Given an applicative model, a named entity is any linguistic expression which refers in a corpus in an autonomous manner to a unique entity of the model.'

An efficient modeling and processing of NEs calls for a combination of complementary language resources and tools covering their morphological, syntactic, semantic and discourse aspects. In this section, I am particularly interested in language phenomena related to the multi-word nature of named entities (cf. Section 4.1). I summarize the state of the art in annotation, recognition and lexical semantics of NEs (Section 4.2) and I describe my contributions in these fields in an inflectionally rich language, Polish, and in a multilingual context (Sections 4.3–4.5). I also address the task of coreference annotation (Section 4.6), which offers a richer semantic content to the idea of entity detection and thus paves the way for entity linking.

4.1 Named Entities as Particular Types of MWEs

Even if named entities comprise both single words (e.g. *Europe*) and multi-word units (*European Union*) the proportion of the latter is largely dominating in lexicographic resources. Recall for instance that MWUs account for about 98% of all lemmas in the Grammatical Lexicon of Warsaw Urban Proper Names (SAWA), discussed in Section 3.5. Also in the Polish module of Prolexbase (cf. Section 4.5.1), about 66% of proper names contain at least two tokens.

In corpora the proportion of single-word vs. MWU named entities is much more balanced. As shown in Table 4.1, in the manually annotated part of the Polish National Corpus (NKJP) discussed below, all named entities amount to over 82,000 units, only 20% and 22% of which,

Table 4.1: Distribution of single-word and MWU named entities, with and without nested structures, in the manually annotated part of the National Corpus of Polish

		Single-token NEs			MWU NEs			Total
		With nested NEs	Without nested NEs	All	With nested NEs	Without nested NEs	All	
All NEs	Base forms	**12,313 (15%)**	53,759	66,072	**10,942**	**5,311**	16,253 **(20%)**	82,325
	Occurrences	**12,337 (14%)**	55,517	67,854	**10,919**	**8,620**	19,539 **(22%)**	87,393
Outermost NEs	Base forms	**11,953 (23%)**	23,850	35,803	**10,573**	**4,891**	15,464 **(30%)**	51,267
	Occurrences	**11,977 (21%)**	25,595	37,572	**10,550**	**8,196**	18,746 **(33%)**	56,318

at the level of occurrences and of lemmas[1], respectively, are MWUs. The remaining 80% and 78% are single tokens. This, however, does not mean preference for single-word NEs in textual utterances. NEs in the NKJP corpus, as extensively discussed below (cf. Section 4.3.1), are annotated not only for their maximum-length occurrences but also for all nested NEs included therein (e.g. $[ulica\ [[Mikołaja]_{\text{forename}}\ [Kopernika]_{\text{surname}}]_{\text{persName}}]_{\text{geogName}}$ 'Mikołaj Kopernik Street'). Moreover, many single-word names, such as person and street names, are seen as elliptical variants of larger MWU names, as explained in Section 4.3.3, p. 88 (e.g. $[[Adam]_{\text{forename}}]_{\text{persName}}$). These can, therefore, also contribute to the prevalence of MWUs in NEs. If all MWU NE occurrences, as well as single-word NEs with nested structures, are considered, MWU named entities account for as many as 35% of lemmas and 36% of occurrences. If only outermost, i.e. maximum-length, NEs are taken into account, these proportions rise up to 53%. In other words **named entities concerned by MWU-related phenomena are more frequent in the corpus than single-word NEs**. Let us also note that as many as 35% of all names in the corpus are nested within other NEs, which shows that an appropriate modeling of nested structures is crucial for a high quality NE annotation, recognition and categorization.

Multi-word named entities can clearly be seen as particular types of multi-word expressions. Sag et al. (2002) classify them as semi-fixed expressions and note both their high idiosyncrasy and lexical proliferation, which jeopardize the words-with-spaces approaches to NEs. Our studies on urban proper names collected in the SAWA e-dictionary (cf. Section 3.5) largely confirm this point of view.

4.2 Named Entity Processing – State of the Art

The interest of the international NLP community in processing named entities shifted in time along two main axes: the nature of units in focus and the degree of multilinguality.

Initially, named entities were mainly understood as the *signifiants* (de Saussure, 1916), i.e. proper names and other naming lexemes and phrases, which would more appropriately be called *naming entities*. According to the general survey by Nadeau & Sekine (2007), the need for named entity recognition and classification (NERC), more often called **named entity recognition** (NER), was identified by the Message Understanding Conference (MUC) in 1996[2] as having a crucial importance in Information Extraction (IE). Occurrences of proper names and related

[1]The differences between numbers of occurrences and of lemmas stem from the fact that temporal expressions are assigned normalized ISO forms instead of lemmas.

[2]`http://www-nlpir.nist.gov/related_projects/muc/proceedings/muc_7_toc.html`

naming entities were to be automatically tagged in a text and categorized with a small set of types (e.g. persons, locations, organizations, and miscelaneous). This task was further promoted by evaluation campaigns within MUC-7, CoNLL-2002[3] and CoNLL-2003[4].

Later on, the need for focusing on the *signifiés* (named entities in the literal sense) was stressed with the advent of the Automatic Content Extraction (ACE) program (Doddington et al., 2004). It redefined the research objectives so as to focus on the target objects (entities, relations, events, etc.) rather than on the linguistic units naming them. Thus, all *mentions* of an entity in a text, e.g. definite expressions (*the former Soviet president*) or pronouns (*he*), became of equal interest to proper names (*Mikhail Gorbachev*), which notably implied **co-reference resolution**.

A step forward in this entity-centered view was taken more recently by the Text Analysis Conference[5] (TAC), organized since 2008, a successor of TREC[6]. It added a new connection between named entities in lexical and ontological resources and their occurrences in texts. Notably, in its **entity linking** track (within the Knowledge Base Population task) competitors are given an initial knowledge base (KB) consisting of several hundred thousand entities from English Wikipedia annotated with 4 types. Given a named entity and a source text in which it appears, the task is to provide the identifier of the same entity in the KB. All non-KB (NIL) entities have to be clustered in order to allow for the KB population. In this way, recognizing a NE in a text truly leads to identifying the very object or concept referenced by it, i.e. to **named entity disambiguation** (NED) (Hachey et al., 2013).

The ultimate stage of these extensions in problem definition towards the semantics of NEs stems from the Semantic Web and the **Linked Open Data** (LOD) (Bizer et al., 2009; Mendes et al., 2012; Suchanek et al., 2007; Hoffart et al., 2011). They aim at (automatically or semi-automatically) building an ontological layer over open data such as Wikipedia, GeoNames, etc. They use unique item identifiers (URIs) and provide well-defined formal knowledge representation and querying models and languages (RDF, RDFS, OWL, SPARQL), due to which ontology navigation, lookup and inference may be more consistent, reliable, simple and fast (provided that the inducted ontology is sound). Thus, if a named entity recognition in a text is accompanied by its linking with the LOD, not only does this NE become fully identified (by its URI) but it is also attached to a whole range of extra-linguistic data and relations which may greatly contribute to the text understanding and to reasoning. (Rizzo et al., 2012) provide an overview of 10 named entity disambiguation engines, 7 of which are multilingual, 4 cover French and one covers a language with a rich declension of nouns (Russian). Three of them use DBpedia as classification ontology. Three give unrestricted on-line access for academic use, all others limit the number of calls per day. Results of such systems still leave much room for improvement, e.g. *DBpedia Spotlight* (Daiber et al., 2013) obtains about 49% precision and 55% recall for Dutch NE spotting and around 0.8 accuracy for entity linking in both Dutch and English.

4.2.1 Named Entity Annotation

In order to gather an important part of the NER community around evaluation campaigns such as MUC and CoNLL, annotation efforts were needed to provide both training and evaluation corpora. For instance, the CoNLL-2003 corpus has the format presented in Figure 4.1, where the columns contain text segments, morphological tags, syntactic group tags and NE tags. In the last column, `O` describes a segment outside any NE, while `I-PER`, `I-ORG`, `I-LOC` and `I-MISC` denote components of person, organization, location and miscellaneous names, respectively.

[3] `http://www.clips.ua.ac.be/conll2002/`
[4] `http://www.clips.ua.ac.be/conll2003/`
[5] `http://www.nist.gov/tac/about/index.html`
[6] `http://trec.nist.gov/`

U.N.	NNP	I-NP	I-ORG
official	NN	I-NP	O
Ekeus	NNP	I-NP	I-PER
heads	VBZ	I-VP	O
for	IN	I-PP	O
Baghdad	NNP	I-NP	I-LOC
.	.	O	O

Figure 4.1: CoNLL-2003 corpus structure

Corpora of a similar structure were created for many languages in order to provide a benchmark in differently defined NLP tasks. Their major drawback is to be hardly applicable for other, even related, tasks. For instance, as mentioned in the named entity linking survey by Hachey et al. (2013), a particular corpus annotation strategy in the Knowledge Base Population task of the Text Analysis Conference promoted those systems which properly dealt with ambiguous named entity. Such a corpus clearly cannot be used for other evaluation settings, notably for evaluating overall performances of this complex task.

In this dissertation, I am particularly interested in those approaches to annotation which view corpora not only as training and evaluation material for NLP tools, but – more importantly – as a means of a, possibly application-independent, modeling of language phenomena. Ideally, NE annotation in such reference corpora should be just one of the many aspects of linguistic annotation. *Prague Dependency Treebank* (PDT) (Böhmová et al., 2003) is a good example of such an approach in the Czech language. Newswire texts are annotated at three layers: morphological, syntactic and "tectogrammatical" (i.e. expressing semantic relationships). Named entities are identified within the tectogrammatical layer together with other multi-word expressions (Bejček & Straňák, 2010), assigned one of 9 types and linked with external lexico-semantic resources. Another multi-level annotated corpus (Desmet & Hoste, 2010) for Dutch contains one million words with manually corrected annotations. The layer of NEs relies on a taxonomy of 6 main types and 17 subtypes. Metonymy is treated with a special care: metonymic occurrences are assigned both their "primary" and their target type. Hinrichs et al. (2005a) describe the multi-level German *TüBa-DZ Treebank*, in which NEs are annotated in the same layer as syntactic groups but seem not be assigned to any particular taxonomy. Further in this chapter I describe some aspects of another multi-level reference corpus annotation, the National Corpus of Polish.

Performing automatic pre-annotation of a corpus prior to human correction is a frequent methodology in corpus development. For instance, the Penn Treebank (Marcus et al., 1993) annotation methodology is based on this principle, both for the parts-of-speech and for the syntactic structures. It is interesting to consider a possible bias introduced by this methodological principle. On the one hand, automatic methods may not only accelerate the annotation but also increase its quality. Namely, they may be relatively reliable in systematically proposing consistent annotations for simple but repetitive phenomena. On the other hand, the annotators may tend to rely too much on the automatic pre-annotation results and skip a number of errors therein. Several studies address this bias problem in different NLP tasks. Marcus et al. (1993) show that, while annotating Penn Treebank with an English tagset of 36 POS tags, semi-automatic tagging is substantially faster, more accurate and more consistent than tagging from scratch. Fort & Sagot (2010) present similar considerations for French: automatic pre-annotation allows for a gain in annotation quality, both in terms of accuracy with respect to a reference and in terms of inter-annotator agreement. The same authors mention state-of-the-art results of experiments with and without automatic pre-annotation in the tasks of: (i) part-of-speech annotation in morphologically complex languages (Hindi and Bangla), (ii) human

information extraction in biomedicine, (iii) annotation with semantic frames. These results do not unanimously confirm the benefit of an automatic pre-annotation. They show at least that a pre-annotation almost never hurts, since globally few negative bias was detected, except a negative, briefly mentioned, bias in annotating gene expressions by (Fort et al., 2009).

4.2.2 Named Entity Recognition and Classification

Many systems in the early stage of NER for English were based on hand-crafted rules and gazetteers. Later on, the trend was towards an increasing use of data-driven methods: supervised (requiring a large annotated training corpus), semi-supervised (using seeds of sample names to extract contexts for new names), and unsupervised (using clustering, co-occurrence analysis or external resources like WordNet). For other languages, e.g. Portuguese, rule-based approaches are still dominant, as shown in the evaluation campaign HAREM (Freitas et al., 2010).

Rule-Based NER Systems

Rule-based NER methods rely on an explicit expression of linguistic and word knowledge in resources such as lexicons and grammars. Such lexicons, also called gazetteers, may have very variable sizes and numbers of features associated to an entry. For instance in (Farmakiotou et al., 2000), the Greek gazetteer contains about 3,000 entries representing only lemmas although the Greek language has a relatively rich inflection. The reason is that the items in the source text are stemmed before they are subject to rule matching. In (Gaizauskas et al., 1995), a flat gazetteer of about 6,000 English names and trigger words is used. In (Wolinski et al., 1995), 8,000 French names are represented in a knowledge base which relates them to attributes such as type, domain and location, as well as to their aliases (orthographic variants, acronyms, etc.). In (Wacholder et al., 1997), the gazetteer contains 3,000 English trigger words and 20,000 first names. In (Mikheev et al., 1999) the English gazetteer consists of 45,000 entries. In (Schäfer, 2006), gazetteers with rich sets of features (here including 20,000 English entries) are automatically extracted from OWL/RDF-encoded ontologies. Some authors note that the quality of NER does not necessarily improve with the growing size of gazetteers. In Mikheev et al. (1999), arguments are given towards using application-tuned gazetteers of limited size rather than far-fetched examples of little known places and organizations.

The size of different named-entity grammars is not always indicated in the reference papers. Note that the number of rules does not always give a good idea of the coverage and precision of the grammar, as it is related to the particular grammar formalism. For instance, (Gaizauskas et al., 1995) rely on about 200 rules (almost 50% thereof address organization names), while Appelt et al. (1995) use only 12 so called macro rules and 15 domain-dependent rules. The latter authors claim that due to compile-time transformations these rules cover approximately as many phenomena as would be described by a hundred explicit patterns.

As far as formal tools for the representation of rules are concerned, NER rule-based engines frequently use finite-state methods. For instance, Krstev et al. (2011) develop e-dictionaries and local grammars for Serbian NEs and get results close to 0.9 F-measure. Finite-state methods are often supported by cascading mechanisms, where the transformed text output from lower-level rules becomes the input for higher-level rules, which gives more expressive power to the resulting systems. This mechanism is used for instance by Hobbs et al. (1997) for English and by Friburger & Maurel (2004) for French.

As far as the set of types and subtypes covered by NER systems is concerned, early systems, e.g. those evaluated within MUC and CoNLL campaigns, such as those by Appelt et al. (1995), Gaizauskas et al. (1995), and Mikheev et al. (1999) use 3 main types: ENAMEX (proper names), TIMEX (temporal expressions), and NUMEX (expressions of quantities and measures), later

completed by artefacts. The first type is subdivided into 3 categories: names of persons, locations and organizations. The recent French evaluation campaign ESTER-2 (Galliano et al., 2009) for NER in spoken corpora uses a much richer typology with 7 main types and 78 subtypes. All competing systems – e.g. Nouvel et al. 2010 which is the rule-based transducer-cascade system having evolved from Friburger & Maurel (2004) – take this hierarchy into account. In HAREM, the NER evaluation for Portuguese (Freitas et al., 2010), the admitted typology is also rather large: it consists of 10 main types and 47 subtypes. An interesting methodological innovation in this last framework is to account for possible vagueness of NE interpretation by allowing more than one tag per annotated NE. According to our experience with manual corpus post-editing (Savary et al., 2012a), this feature proves useful in some cases of actual ambiguity of types and/or attributes.

As far as temporal expressions are concerned, since the beginning of the 21st century an international community has been proposing an elaborate annotation and normalization standard *TimeML*[7]. Consequently, the normalization of temporal expressions has been addressed in some approaches, e.g. for Italian and English in (Tommaso Caselli & Bartolini, 2008) and in (Krstev et al., 2012) for Serbian.

The NE type attached to a recognized entry may be accompanied by a series of application-dependent attributes, which is however rarely discussed in the literature. In Wolinski et al. (1995) these attributes include city, sector of activity, market, financial index, etc. In the rule-based approach addressed in Section 4.4.1, the attributes are of a more linguistic nature, notably base forms, normalized time expressions, as well as derivational bases and their types.

Nested Named Entity Recognition Based on Machine Learning

While a full-fledged state-of-the-art survey in machine learning-based NER is not precisely within the scope of my dissertation, let us recall that many existing approaches admit a flat and contiguous nature of NEs and represent NER as a sequential tagging problem (Jurafsky & Martin, 2009, Chapter 22.1). According to the reference method originally proposed for noun phrase chunking by Ramshaw & Marcus (1995), corpus tokens are annotated with the so-called IOB tags. The `O` tag represents a word outside any NE, `B-T` – an initial component in a NE of type T and `I-T` – a non initial component of a NE of type T. This extends the CoNLL-style tagset presented in Section 4.2.1 with `B-T` tags, which helps distinguish consecutive NEs of the same type. Observation templates are then defined containing different features, which are instantiated for each corpus token: its orthographic form, shape, base form, affixes, occurrences in external lexicons, etc. Each sentence in the annotated training corpus is transformed into a sequence of feature vectors (one vector per token) and a sequential classifier is trained on this corpus. The resulting model can be applied to a new untagged corpus in that each sentence is assigned the most probable sequence of tags given the feature vectors of its component tokens. Tagging decision are based on local contexts of the current token and its several (usually one to four) neighboring tokens.

Taking nested NEs into account dramatically changes the point of view on NE modeling since NEs can no longer be seen as flat contiguous sequences of tokens but are most accurately represented as trees. Until recently, NER community has proposed relatively few contributions to this redefined problem. Early efforts on this task have been made in the biomedical domain due to existence of corpora annotated with nested structures. Named entities in this domain are not related to proper names but are names of proteins, cell types, viruses, lipids, drugs etc. Alex et al. (2007) address the problems of nesting, possible discontinuities and overlapping in such NEs, notably in coordinations. Three simple methods for recognition of such nested NEs

[7] `http://www.timeml.org`

are described: layering, cascading, and joined label tagging. All reduce the problem to layered sequence tagging. In layering, each level of nesting is modeled as a separate IOB problem. The output of models trained on individual layers is combined subsequent to tagging. Cascading reduces the nested NER task to several IOB problems by grouping one or more entity types and training a separate model for each group. Joined label tagging, presented in more details in Section 4.4.2, relies on concatenating the IOB tags of all levels of nesting. Results for all three techniques are very similar and included between 62% and 70% F_1 measure. The same paper reports on previous efforts by other authors of tagging biomedical nested NEs with hybrid models, where a probabilistic sequential tagger for innermost NEs is combined with rule-based methods for outermost NEs.

Ramírez-Cruz & Pons-Porrata (2008) propose another sequence-based approach, which takes advantage of deep parsing in Spanish. Candidate NEs are detected in deep constituency trees of a sentence in that each definite noun phrase is considered a potential NE. Candidate NE trees are then represented as sequences of nodes spelled out during the postorder traversal of the trees. The classification of such candidate NE trees is considered a sequence classification problem. The results show a 70.45 and 57.65 F_1 measure for in-domain and out-of-domain evaluation, respectively.

Finkel & Manning (2009c) put forward a more refined solution in which tagging nested NEs is understood as a particular instance of the probabilistic parsing problem. Sentences are represented as constituency parse trees with constituents for each named entity and its embedded NEs. Each tree node is annotated by both its parent and grandparent labels, and each tree is transformed into an equivalent binary tree. A corpus thus transformed is used to train a discriminative CRF-based constituency parser, similar to a probabilistic context-free grammar (PCFG) parser (Jurafsky & Martin, 2009, Chapter 14). The set of features includes notably embedded NE features which represent dependencies between parents of adjacent nodes. The evaluation results on English biomedical data, as well as Spanish and Catalan newswire data, span from 64.55% to 70.33 F_1 measure.

Nested NE recognition in French has been boosted by a recent ETAPE evaluation campaign[8] using partly noisy transcribed speech data from TV and radio broadcasts. Its NE detection task was based on a 1.2-million word corpus (Gravier et al., 2012) manually annotated with a NE taxonomy of 7 types and 32 sub-types, in which embedded NEs are explicitly marked. Among several participating systems, *mXS* (Nouvel et al., 2013) implements a novel idea of detecting left and right NE boundaries independently via pattern extraction techniques, and correcting non-consistent NE tag sequences by dynamic programming. Dinarelli & Rosset (2012) propose another approach in which sequential labeling via CRFs is combined with PCFG parsing.

Named Entity Annotation and Recognition in Polish

The National Corpus of Polish (Przepiórkowski et al., 2012), discussed extensively below in this chapter, is probably the largest and the most comprehensive attempt towards creating a manually annotated reference NE corpus in Polish. Previous efforts of manual annotation include two domain-specific corpora: (i) a multi-level annotated corpus of dialogs concerning the Warsaw transportation system (Mykowiecka et al., 2008) containing 81,000 words and about 6,200 annotated named entities, (ii) a corpus of texts in economy and stock exchange (Marcińczuk & Piasecki, 2011) with an annotation schema and a format similar to CoNLL corpora, containing 330,000 words and about 9,000 annotated person, location and institution names. (Broda et al., 2012) report on a work in progress on *KPWr*, a multi-genre multi-level annotated corpus dedicated to training and evaluation of machine learning-based tools. Its NE annotation

[8] `http://www.afcp-parole.org/etape.html`

schema assumes 57 NE categories and annotating nested structures. By 2012 the number of annotated NEs exceeded 16,000 instances. The corpus is distributed under the Creative Commons Attribution 3.0 Unported Licence.

Nothman et al. (2013) report on an effort of automatically creating huge *silver-standard* NER-annotated corpora extracted from Wikipedia for 9 languages including Polish. Wikipedia articles are classified with a fine-grained (19 types) and a coarse-grained (6 types) taxonomy using a supervised approach with category propagation among languages. Contents of Wikipedia articles are then tagged for each outgoing link with the type of its target article. Additional links are inferred from redirects, disambiguation pages, and anchor texts. Multiple (unlinked) occurrences of the same entities are recognized by a naive prefix-based approach. The resulting, non evaluated, corpus contains almost 53 million Polish tokens, and its training subcorpus of roughly 3.5 million words is publicly available[9].

As far as Polish NER is concerned, the work reported in Piskorski (2005) describes, to our best knowledge, the first systematic attempt towards creation of a fully automated rule-based NER system for Polish, built on top of *SProUT*. This very system was the starting point for our automatic pre-annotation tools described below. It covers the classical named-entity types, i.e., persons, locations, organizations, as well as numeral and temporal expressions. The NER resources created in this first study were adapted and further extended by Abramowicz et al. (2006) in order to create information extraction tools used in cadastral information systems.

Marcińczuk & Piasecki (2007) report on a memory-based learning approach to automatically extract information on events in the reports of Polish Stockholders. In particular, resources for extracting locations and temporal expressions for Polish were created. In a follow-up work, (Marcińczuk & Piasecki, 2010), which focused on the same domain, some accuracy results of NER algorithm based on the Hidden Markov Model are presented. Also in (Lubaszewski, 2007) and (Lubaszewski, 2009) some general-purpose information extraction tools for Polish are addressed.

Graliński et al. (2009b) present *NERT*, another rule-based NER system for Polish which covers similar types of NEs as Piskorski (2005), but the underlying grammar formalism is simpler. *NERT* has been mainly implemented for deployment in machine anonymisation and translation (Graliński et al., 2009a).

More recently, Marcińczuk et al. (2013) describe *Liner2*, an open source NE recognizer based on Conditional Random Fields (CRF), using 5 categories (first names, surnames, city names, road names and country names) and a set of orthographic, morphological, lexical and semantic features. It refers notably to the Polish WordNet for synonyms and hypernyms, and to a gazetteer of 1.37 million Polish proper names. The authors report a 95.57% and 79.63% F_1 measure in an in-domain and cross-domain evaluation, respectively.

In Section 4.4 I report on *Nerf*, another CRF-based system, developed within the National Corpus of Polish Project. Both Liner2[10] and Nerf[11] are available as web services.

4.2.3 Lexical and Semantic Resources for Named Entities

As mentioned above, recent advanced in NER introduce close links between named entities in lexical and ontological resources and their occurrences in corpora. Creation and enrichment of such resources has a rich bibliography most of which was initially dedicated to English, and has been more recently applied to other languages. Several approaches are based on aligning WordNet with Wikipedia (Toral et al., 2008, 2012; Fernando & Stevenson, 2012; Nguyen & Cao, 2010), *YAGO* (Suchanek et al., 2007) and *YAGO2* (Hoffart et al., 2011). Others build new

[9] `http://sydney.edu.au/engineering/it/~joel/wikiner/aij-wikiner-pl-wp3.bz2`
[10] `http://nlp.pwr.wroc.pl/inforex/index.php?page=ner`
[11] `http://glass.ipipan.waw.pl/multiservice/`

semantic layers over Wikipedia alone: *Freebase* (Bollacker et al., 2007), *MENTA* (de Melo & Weikum, 2010), *DBpedia*[12] (Bizer et al., 2009; Mendes et al., 2012). Given the quantities of data to be processed, relatively few efforts are made towards manual data validation.

Many other efforts have been made towards the construction of particular application- or language-oriented proper name thesauri and their exhaustive study is out of the scope of this dissertation. *JRC-NAMES* (Steinberger et al., 2011) is a notable example in which a lightly structured thesaurus of several hundred thousand named entities, mainly person names, is being continuously developed for 20 languages. New names and their variants are extracted by a rule-based named-entity recognizer from 100,000 news articles per day and partly manually validated.

In (Savary et al., 2013b) we present a contrastive state-of-the-art survey in the domain of lexical and ontological resources including NEs, which shows a large variability of eight approaches in terms of the languages covered, the sizes of the resulting knowledge bases, the methods of ontology mapping and population, and the coverage of linguistic features. In the light of this study, large multilingual ontologies open exciting perspectives in many NLP domains but they still have insufficient explicit links with morphological and syntactic data necessary for morphologically rich languages, in particular those with a complex declension system in nouns and adjectives.

The following sections describe my contributions in creating high quality language resources and tools for one of such highly inflected languages, i.e. Polish. I show how NEs are annotated in a large reference corpus according to high annotation standards. I describe a pioneering work on nested NE recognition in this language. Finally, I report on efforts towards a manually validated fine-grained multilingual (Polish-English-French) NE ontology containing both semantic and morphological data.

4.3 Annotating Named Entities in the National Corpus of Polish

The National Corpus of Polish[13] (Pol. *Narodowy Korpus Języka Polskiego*; **NKJP**) is a 1.5-billion ($1.5 * 10^9$) word corpus of Polish annotated at various levels, with a 300-million balanced subcorpus (Przepiórkowski et al., 2012). The following linguistic annotation layers are distinguished (Bański & Przepiórkowski, 2009): segmentation (word-level and sentence-level), morphosyntax, word sense disambiguation (limited to around 100 lexemes), syntactic words, syntactic groups and named entities.

A **1-million word balanced subcorpus** contains randomly chosen paragraphs[14] of the whole corpus. It underwent manual annotation at all the abovementioned layers. Each time (except at the word sense layer) texts were automatically pre-annotated and then manually corrected and completed by two independent annotators. Finally, discrepancies were reviewed by an adjudicator. The 1-million word subcorpus served as a training corpus for various annotation tools. It is represented in a stand-off annotation format and distributed under the GNU GPL v. 3 license.

Different aspects of named entity annotation in the National Corpus of Polish have been described in several publications. In (Savary et al., 2010) we outline the annotation scope, the TEI-P5-inspired hierarchy of named entities and the multi-level stand-off annotation format, as well as some methodological strategies. In (Savary & Piskorski, 2010) and (Savary & Piskorski, 2011) we show how existing lexical resources and grammars for Polish named entity recognition

[12] `http://dbpedia.org`

[13] `http://nkjp.pl/`

[14] This helps overcome copyright problems related to some types of source texts.

have been adapted in order to be used in the process of automatic pre-annotation of the corpus. Evaluation and error analysis is performed for the resulting rule-based NER system. In (Waszczuk et al., 2010) we describe methods and tools used during the simultaneous annotation of both named entities and syntactic words and groups, from automatic pre-annotation, through file management, manual annotation and adjudication, as well as various format conversions. In (Waszczuk et al., 2013) we further discuss the inter-annotator agreement and we introduce a baseline probabilistic NER tool trained on the manually-annotated corpus. In (Savary et al., 2012a) we document the annotation guidelines and discuss a large range of interesting linguistic phenomena encountered during the NE annotation, such as metonymy, ellipsis, type and nesting ambiguity and geopolitical issues. In (Savary & Waszczuk, 2012) we revisit the NE annotation tools used in pre-annotation and manual annotation of the one-million word subcorpus, as well as in the automatic annotation of the whole 1,5-billion word corpus.

In this chapter I resume the general characteristics of the NKJP corpus, and I summarize the annotation schema and tools for NEs. I then elaborate on specific challenges posed by Polish **multi-word NEs** such as nesting, coordination, discontinuity, ellipsis, metonymy and derivation.

4.3.1 Named Entity Annotation Schema

The rules admitted for named entity annotation in the NKJP project result from a compromise between the precision of linguistic data and the richness of naming phenomena in Polish texts. The NE type taxonomy was inspired by TEI P5 (Burnard & Bauman, 2008), as shown in Fig. 4.2. We take into account most NE types common for different NE projects, such as names of persons, locations, organizations, and numerical expressions. Note that some differences exist in our list of basic NE categories with respect to other state-of-the-art approaches such as (Sekine et al., 2002). Notably, locations are distributed within two types called `placeName` and `geogName`. According to TEI P5, the former is meant for hierarchically-organized geo-political or administrative units (districts, regions etc.), while the latter refers simply to objects having geographical features such as mountains or rivers. This distinction may be useful because names of administrative units frequently appear as metonyms (designating the inhabitants of the unit), in which case they should be seen as organizations rather than locations (cf Chinchor 1997).

Note also that the hierarchy in Figure 4.2 is non homogeneous. Personal subtypes correspond to parts of a personal name, while geographical subtypes refer to types of objects they name and their mutual relations. Such heterogeneity is common for many existing taxonomies. A solution to this problems was more recently proposed in the Quaero annotation guidelines (Rosset et al., 2011) for French, where annotation is divided into two dimensions: NE components (surname, given name, zip code, month, etc.) and NEs (persons, locations, time, etc.).

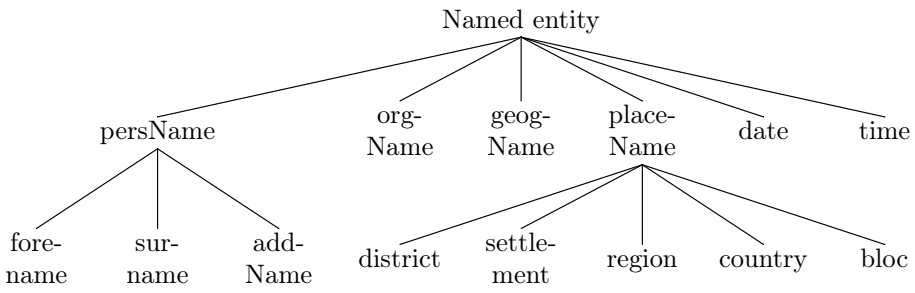


Figure 4.2: Type hierarchy of Polish NEs

We did not annotate other NEs, such as events, quantities and measures, product and vessel names, titles of works and texts. Within temporal expressions we did not treat expressions of

duration (*przez dwa dni* 'for two days'), sets (*co drugi dzień* 'every other day') and relative time (*wczoraj* 'yesterday'). We were, however, interested in some units that are less frequently covered by other projects, such as **relative adjectives** stemming from person, location and organization names (e.g. *warszawski* 'Warsaw-related'), as well as what we call **personal derivations**, i.e. names of inhabitants (e.g. *warszawiak* 'inhabitant of Warsaw') and organization members. Their annotation in the corpus includes indicating the semantic *derivational bases*, e.g. *warszawiak* → *Warszawa* 'Warsaw', *amerykański* → *Stany Zjednoczone* 'American' → 'United States'. Note that this attachment is context-dependent and cannot always be unambiguously done by an external lexicon. For instance *ostrowski* is an adjective related to several Polish towns: *Ostrów Wielkopolski*, *Ostròw Mazowiecka*, etc., while *europejski* 'European' can refer to *Europa* 'Europe' or to *Unia Europejska* 'European Union'. Derived names are, thus, annotated with a two-dimensional typology. The first dimension describes the type or subtype of their derivational bases, according to the type hierarchy in Fig. 4.2. The second dimension concerns the type of derivation. For instance, *warszawiak* 'inhabitant of Warsaw' receives the type `city` and the derivation type `persDeriv` (personal derivation).

Apart from the main type, and possibly the subtype of the NE, other annotated attributes important for the creation of resources and grammars include:

- Lemma (attribute `@base`, e.g. *Stany Zjednoczone* for *Stanów Zjednoczonych* 'United States')
- TEI-P5-inspired normalization of date and time (attribute `@when`, e.g. *2009-10-30*, *09:45:00*)

Note that determining the **lemma of a NE**, is a non trivial task in a highly inflected language such as Polish, in particular for **compound and personal names**, as discussed in Piskorski et al. (2009). That is why we put a special impact on the creation of NE resources containing such lemmas, as well as their automatic deduction in grammar rules.

Traditional NER, MUC and CoNLL campaigns, have focused on identifying and classifying flat maximum-length NEs. More recent research shows the importance of representing the internal structure in **recursively embedded NEs** (Alex et al., 2007; Galicia-Haro & Gelbukh, 2009; Ramírez-Cruz & Pons-Porrata, 2008; Finkel & Manning, 2009c; Kravalová & Žabokrtský, 2009; Dinarelli & Rosset, 2012; Nouvel et al., 2013) and their overlapping with nominal phrases (Finkel & Manning, 2009b; Osenova & Kolkovska, 2002) in multi-level annotation. Thus, in NKJP we annotated each NE together with other NEs possibly included in it. For instance:

(4.1) $[[Maria]_{\text{forename}}\ [Skłodowska]_{\text{surname}}\text{-}[Curie]_{\text{surname}}]_{\text{persName}}$

(4.2) $[ulica\ [[Mikołaja]_{\text{forename}}\ [Kopernika]_{\text{surname}}]_{\text{persName}}]_{\text{geogName}}$
STREET MIKOŁAJ$_{gen}$ KOPERNIK$_{gen}$
'Mikołaj Kopernik Street'

(4.3) $[[Wydział\ Prawa]_{\text{orgName}}\ [Uniwersytetu\ [Warszawskiego]_{\text{relAdj:settlement(Warszawa)}}]_{\text{orgName}}]_{\text{orgName}}$
FACULTY$_{nom}$ LAW$_{gen}$ UNIVERSITY$_{gen}$ VARSOVIAN$_{gen}$
'Law Faculty of the University of Warsaw'

We believe that such representation has three advantages: (i) it enlarges the density of annotated NEs in the corpus, (ii) it facilitates further treatment of coreferences, as well as relations occurring between different NEs, (iii) it may help in NE type disambiguation.

The richness of the whole annotation schema is illustrated in Figure 4.5, described in the next section, showing a sample NE with an embedded relational adjective, encoded in the TEI-P5 format.

4.3.2 Annotation Data Flow

The data flow in the 1-million-word subcorpus is shown in Fig. 4.3 (see Section 4.4.2 for the dataflow in the automatic annotation of the entire 1.5-billion word main corpus). The left-hand side presents different annotation levels in NKJP.

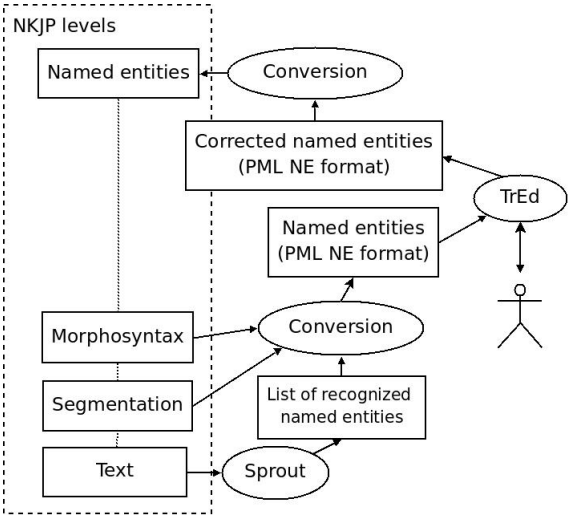


Figure 4.3: Data flow in the NE manual annotation task of the 1-million word NKJP subcorpus

Raw texts taken from the corpus repository were processed by lexical resources and grammar rules within the ***SProUT*** platform (Becker et al., 2002; Drożdżyński et al., 2004) (cf. Section 4.4.1). This tool offers several convenient features such as: (i) a rather rich grammar formalism with finite-state operators, unification and cascading, (ii) a very fast gazetteer lookup, (iii) an XML-based output, called Sproutput, in the form of typed feature structures whose type hierarchy can be defined by the user. SProUT was adapted to processing Polish texts by Piskorski et al. (2004) and Polish-specific NE lexical resources and grammars were addressed in (Piskorski, 2005). These resources and grammars, meant for an information retrieval (IR) task, had to be adapted to the NKJP annotation task (Savary & Piskorski, 2010). In particular, we had to redesign the rules so that the output structures contain the **features of all NEs embedded in the outermost sequences**, as discussed in Section 4.4.1.

As shown in Fig. 4.3, SProUT's output was further converted into another XML format, called PML-NE, defined for the tree editor ***TrEd*** (Pajas & Štěpánek, 2008)[15]. TrEd was selected, after evaluation of several annotation platforms including Synpathy[16], MMAX[17], and GATE Wilcock (2009), for the following reasons: (i) admitting pre-annotated input and multi-level annotation, (ii) customizable open XML-based abstract data format (PML), (iii) easy manipulation of tree representations decorated with user-defined feature structures, (v) ergonomic customizable graphical user's interface, (vi) parallel editing of concurrent annotations, (vii) rich documentation, (viii) technical reliability. Each corpus fragment was edited in TrEd by two human annotators. Then, an adjudicator (called *super-annotator*) reviewed the cases of disagreement and chose the correct annotation. Each annotator and super-annotator worked off-line with TrEd installed locally. She consulted remote project repositories in order to get new versions of NKJP extensions for TrEd. She also had an access to a remote subversion repository, where files to be annotated were stored.

[15] `http://ufal.mff.cuni.cz/~{}pajas/tred/`
[16] `http://www.lat-mpi.eu/tools/synpathy`
[17] `http://mmax2.sourceforge.net`

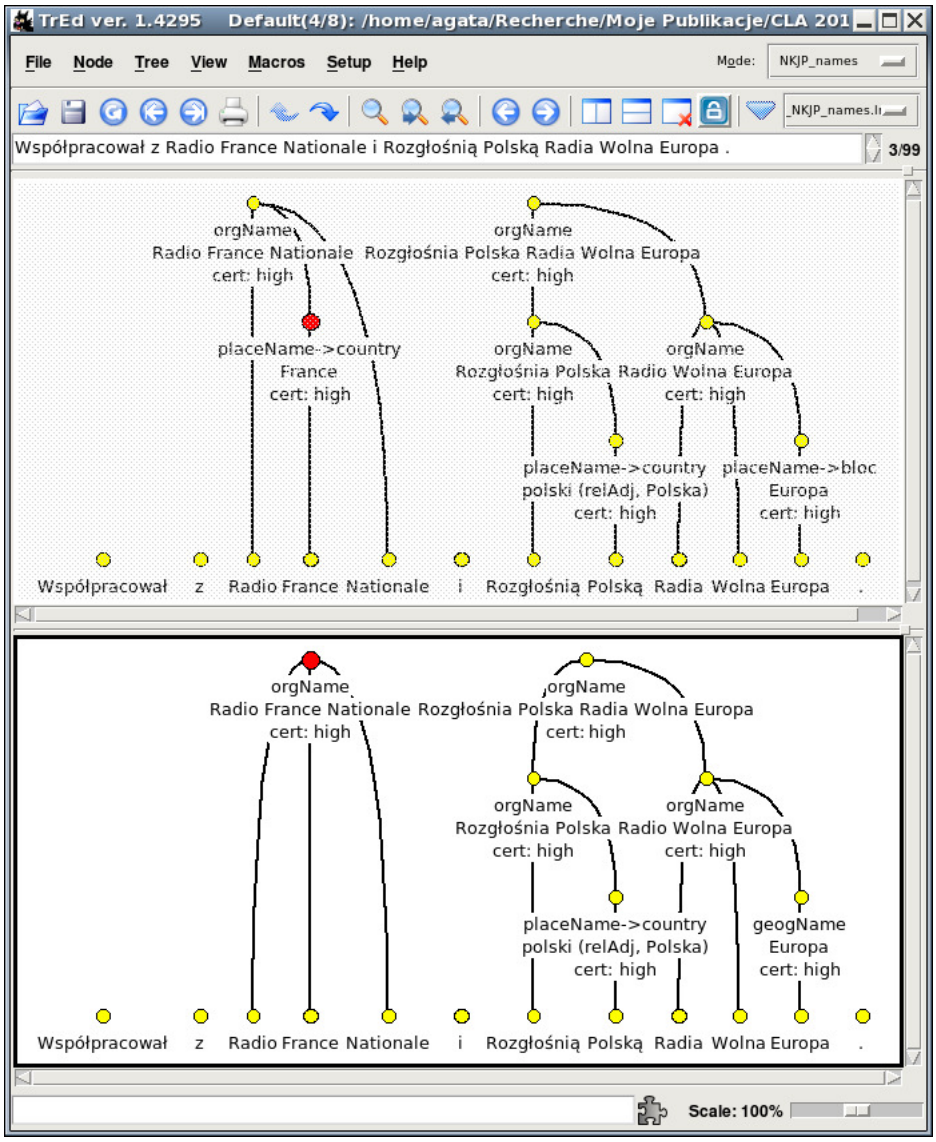


Figure 4.4: Super-annotation in TrEd of a sentence with multiply embedded NEs: 'He collaborated with Radio France Nationale and the Polish Station of the Free Europe Radio.'

TrEd had been customized by various macros, layout stylesheets and keyboard shortcuts in order to best fit the needs of NE annotation and super-annotation in NKJP. Figure 4.4 shows the NKJP-customized TrEd super-annotation interface with a sample sentence edited differently by two annotators. In each of the two windows, the lowest level contains sentence tokens (here: *Współpracował*, *z*, etc.). The highest level shows the outermost NEs (here *Radio France Nationale* of type *orgName*, etc.). All other intermediate levels represent embedded NEs. The essential node attributes — main type, subtype (if any), base form, and certainty level[18] — are visible under the node. Discrepancies between two annotations are detected automatically and single keyboard shortcuts allow to shift parts of annotations from one annotator's version to the other. Here, we can transfer the highlighted node $[France]_{country}$ from the upper to the lower window, over the node *France* and under the node *Radio France Nationale*.

[18]The certainty level represented by the attribute *cert* represents the degree of annotator's confidence with respect to her own annotation.

The last stage consisted in converting the PML-NE format of the validated annotations into the final **stand-off multilevel NKJP format** (Przepiórkowski & Bański, 2009). A stand-off annotation consists in keeping the source text intact and expressing annotations in layer L_n in an external file containing pointers to the underlying layers $L_1, \ldots, L_{n-1}$. Thus, the level of named entities L_{named} is built upon the level of morphosyntax[19] $L_{morphosyntax}$. In parallel, the level of syntactic groups L_{groups} builds upon the level of syntactic words[20] L_{words}, which in their turn build upon $L_{morphosyntax}$. As discussed extensively in Section 3.4, the morphosyntax of a compound NE is not always a straightforward function of the morphosyntax of its constituents. However, within NKJP we did not annotate the morphosyntax of NEs manually. We expect instead that it can be deduced later, largely automatically, from the underlying level of syntactic words, from the lemma of each NE, and from the annotated syntactic groups (Głowińska & Przepiórkowski, 2010).

Figure 4.5 shows a sample NE with an embedded relational adjective, encoded in the TEI-P5 format. The organization name *Irlandzka Armia Republikańska* 'Irish Republican Army' points to ⟨`seg`⟩ments `morph_1.2-seg` (*Armia* 'Army') and `morph_1.3-seg` (*Republikańska* 'Republican') at the $L_{morphosyntax}$ level (in file `ann_morphosyntax.xml`), and to ⟨`seg`⟩ment `named_1.34-s_n3` (*Irlandzka* 'Irish') defined just below at the L_{named} level. Both named entities have a set of attributes defining their types (`ne_type`), subtypes, if any (`ne_subtype`), corpus occurrence forms (`orth`), lemmas (`base`), and the degree of annotator's certainty with respect to this annotation (`certainty`). Additionally, the derivational adjective *Irlandzka* 'Irish' is assigned its type of derivation (`derivType`) and its derivational base *Irlandia* 'Ireland' (`derivedFrom`).

4.3.3 Annotation Challenges from Multi-Word Named Entities

Subtasks of the NE annotation according to the rules summarized in Section 4.3.1, are – formally speaking – classification problems. Each single- or multi-word unit considered a NE had to be assigned exactly one of the pre-defined types and/or subtypes, and the correct values were to be determined for each of its attributes. However, named entities, like many other linguistic objects, have a controversial status, fuzzy boundaries between categories and fuzzy lexical and semantic relations. In (Savary et al., 2012a), I show some linguistic properties of NEs and interesting problems which challenged the NE annotation process in NKJP. In this chapter I summarize those properties and problems which specifically concern multi-word named entities and NE nesting.

Henceforth, I admit two equivalent notations for annotated NEs and their attributes. In the first one, shown in examples (4.4)–(4.5), the annotated NE is bracketed. The subscript index contains: (i) a type of derivation, if any (`relAdj` for a relative adjective and `persDeriv` for an inhabitant or organization member), (ii) the main type and an optional subtype. The superscript index shows: (i) the lemma or the normalized form (in case of temporal expressions), (ii) the derivation base (for derivations). The second notation, illustrated in examples (4.6)–(4.7), consists in an annotation tree where each NE is represented as a node labeled with its base or normalized form, a possible derivation type, the main type and subtype, and a possible derivation base. The NE components are children of this node.

(4.4) *w* $[Paryżu]^{\text{Paryż}}_{\text{placeName.settlement}}$ $[dnia\ 21\ września\ 1960\ r.]^{\text{1960-09-21}}_{\text{date}}$
IN PARIS_{loc} DAY_{gen} 21 SEPTEMBER_{gen} 1960 Y.
'in Paris on the 21st of September 1960'

[19] Initially, the level of named entities was supposed to be built upon the level of syntactic words. Since the annotation of these two levels was performed in parallel, the already available level of morphosyntax was chosen instead.

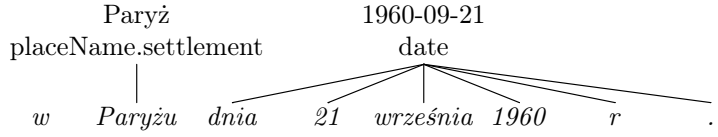
[20] minimal single tokens or groups of tokens corresponding to traditional parts of speech

```
<teiCorpus xmlns:xi="http://www.w3.org/2001/XInclude" xmlns="http://www.tei-c.org/ns/1.0">
<xi:include href="NKJP_1M_header.xml"/>
<TEI>
  <xi:include href="header.xml"/>
  <text><body>
  <p xml:id="named_1-p" corresp="ann_words.xml#words_1-p">
    <s xml:id="named_1.34-s" corresp="ann_words.xml#words_1.34-s">
      <seg xml:id="named_1.34-s_n2">
        <fs type="named">
          <f name="ne_type"><symbol value="orgName"/></f>
          <f name="orth"><string>Irlandzka Armia Republikańska</string></f>
          <f name="base"><string>Irlandzka Armia Republikańska</string></f>
          <f name="certainty"><symbol value="high"/></f>
        </fs>
        <ptr target="named_1.34-s_n3"/>                        <!-- Irlandzka -->
        <ptr target="ann_morphosyntax.xml#morph_1.2-seg"/>   <!-- Armia -->
        <ptr target="ann_morphosyntax.xml#morph_1.3-seg"/>   <!-- Republikańska -->
      </seg>
      <seg xml:id="named_1.34-s_n3">
        <fs type="named">
          <f name="derived">
            <fs type="derivation">
              <f name="derivType"><symbol value="relAdj"/></f>
              <f name="derivedFrom"><string>Irlandia</string></f>
            </fs>
          </f>
          <f name="ne_type"><symbol value="placeName"/></f>
          <f name="ne_subtype"><symbol value="country"/></f>
          <f name="orth"><string>Irlandzka</string></f>
          <f name="base"><string>irlandzki</string></f>
          <f name="certainty"><symbol value="high"/></f>
        </fs>
        <ptr target="ann_morphosyntax.xml#morph_1.1-seg"/>
      </seg>
    </s>
  </p>
  </body></text></TEI>
</teiCorpus>
```

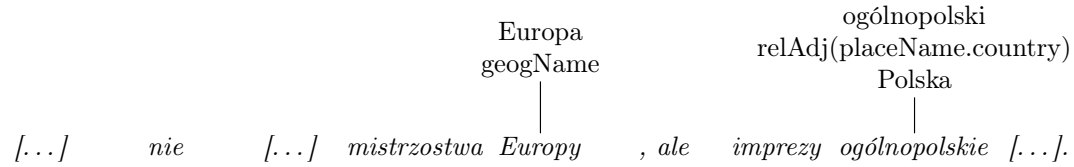
Figure 4.5: TEI-P5-conformant encoding for the named entity *Irlandzka Armia Republikańska* 'Irish Republican Army'.

(4.5) *może nie od razu mistrzostwa* [*Europy*]$^{\text{Europa}}_{\text{geogName}}$*, ale [. . .] imprezy* [*ogólnopolskie*]$^{\text{ogólnopolski; Polska}}_{\text{relAdj(placeName.country)}}$ *rangi juniorskiej*
MAYBE NOT AT ONCE CHAMPIONSHIPS EUROPE$_{gen}$ BUT EVENTS GENERALLY-POLISH RANK$_{gen}$ JUNIOR-ADJ$_{gen}$
'maybe not Europe championships but Polish-national junior-level events'

(4.6)



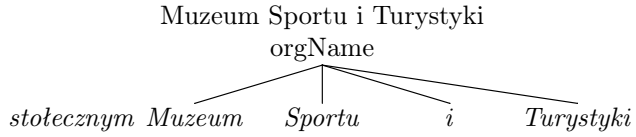
(4.7)



Coordinated Names

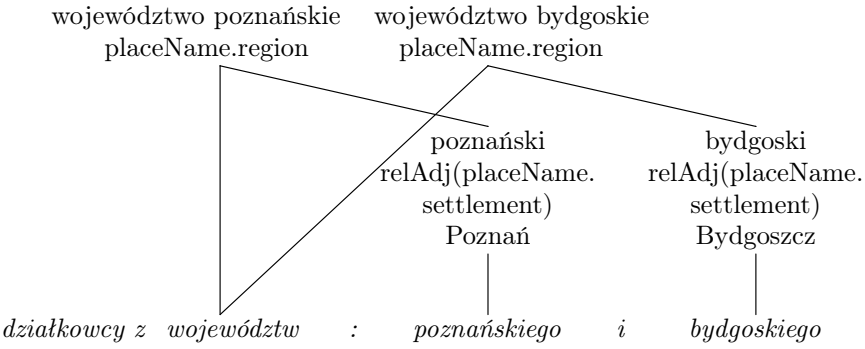
One of novel annotation principles in NKJP concerns coordinated NEs. If a conjunction is an inherent component of the NE it is attached to the NE annotation tree, as in example (4.8). If, however, the coordination spans over two different NEs, each of them is annotated separately, as in example (4.9). If such coordinated NEs share a common component two problems appear: partial overlapping and discontinuity. A stand-off annotation (cf Section 4.3.2) is crucial for these phenomena since they are easily representable by trees but not by bracketing.

(4.8)



CAPITAL-ADJ$_{loc}$ MUZEUM$_{loc}$ SPORT$_{gen}$ AND TOURISM$_{gen}$
'Museum of Sports and Tourism in the capital'

(4.9)



ALLOTMENT-OWNERS FROM VOIVODESHIPS$_{gen}$: POZNAŃ-ADJ$_{gen}$ AND BYDGOSZCZ-ADJ$_{gen}$
'allotment owners from the voivodeships of Poznań$_{relAdj}$ and Bydgoszcz$_{relAdj}$'

Some related works have stressed similar problems in NEs, multi-word expressions and other syntactic groups. Bejček & Straňák (2010), while annotating MWEs in Prague Dependency Treebank, duplicate the nodes shared in coordinated structures (at the tectogrammatical layer). In BulTreeBank, a syntactically annotated corpus of Bulgarian (Osenova & Simov, 2004), head-words missing due to phrase coordination are represented as 'zero elements' whenever they belong to coreference chains. Mazur & Dale (2007) analyze the ambiguity of conjunctions in candidate named entity strings and propose a supervised machine learning approach to conjunction disambiguation.

Person Names

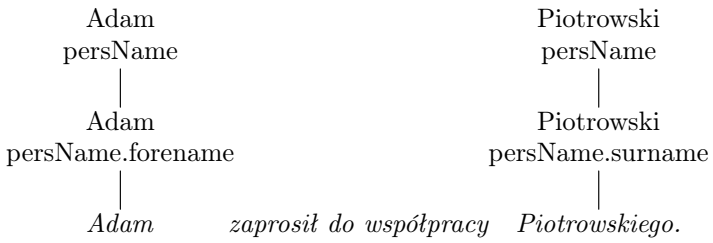
Among all Polish NEs, person names show specific behavior:

- as previously mentioned, their subtypes denote their components rather than types of the named objects,
- the morphology of these components is richer than in other names; e.g. they inflect for number unlike NEs of other types,

- they appear more frequently in coordinated structures.

For these reasons, some specific annotation rules have been defined for person names. Firstly, at least two nodes are created for each person name (as long as it contains a given name, a surname or an additional name), even if this name is composed of one unit only, as in example (4.10). In other words, we consider that single-token names like *Adam* or *Piotrowski* are in fact occurrences of multi-word NEs with elided components (here: the surname and the given name, respectively). Similar proposals are contained in the more recent Quaero annotation guidelines (Rosset et al., 2011) for French. The modeling proposed by these authors is more elegant though due to the two-dimensional (components vs. NEs) type hierarchy.

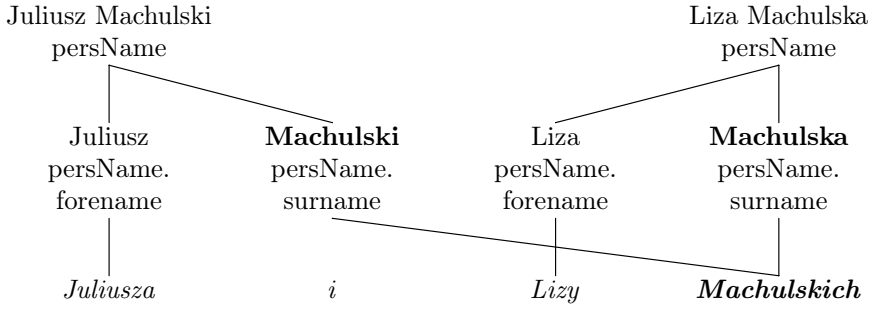
(4.10)



ADAM INVITED TO COLLABORATION$_{gen}$ PIOTROWSKI$_{gen}$
'Adam invited Piotrowski to collaborate'

A surname is always considered a noun. Recall (Section 3.3.1) that Polish nouns have gender but they do not inflect for gender. Thus, an inflected surname – even stemming from an adjective – keeps its gender (masculine or feminine) when lemmatized. For instance, the lemma of *Machulskiej* 'Machulski$_{sg:gen:f}$' is *Machulska* 'Machulski$_{sg:nom:f}$' rather than *Machulski* 'Machulski$_{sg:nom:m1}$'. Moreover, plural masculine human surnames shared in coordinated structures by a male and a female name obtain two different base forms, as shown in example (4.11). At least one of them is obviously different from the lemma assigned to the surname on a per-token basis during the process of morphosyntactic annotation (here: *Machulski*).

(4.11)



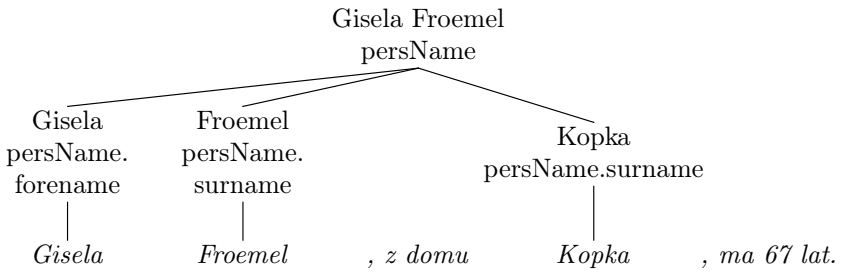
JULIUSZ$_{sg:gen:m1}$ AND LIZA$_{sg:gen:f}$ MACHULSKIS$_{pl:gen:m1}$
'of Juliusz and Liza Machulski'

When a surname appears alone in plural, with no indication of the natural gender of the group of persons meant, the base form is given in nominative plural, as in example (4.12).

(4.12) *Najmłodsi* [[*Kowalscy*]$^{\textbf{Kowalscy}}_{\text{persName.surname}}$]$^{\textbf{Kowalscy}}_{\text{persName}}$ *dostali jeszcze od rodziców po małej ciupadze.*
YOUNGEST$_{pl:nom:m1}$ KOWALSKIS$_{pl:nom:m1}$ RECEIVED YET FROM PARENTS EACH SMALL SHEPHERD'S-AXE
'The youngest Kowalskis also received small Shepherd's axes from their parents'

Some special cases related to maiden names, nicknames and coats of arms may introduce discontinuity, as in example (4.13).

(4.13)



GISELA FROEMEL, FROM HOME KOPKA, HAS 67 YEARS
'Gisela Froemel, born Kopka, is 67 years old'

Elliptical Variants

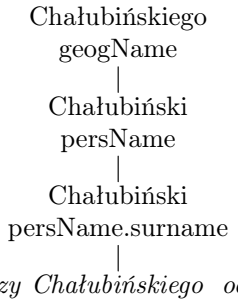
Ellipsis, i.e. omission of one or more components of a phrase, is a frequent phenomenon in NKJP. As mentioned in the previous section, given names or surnames occurring alone can be seen as elliptical variants of full person names. Also organization names, whose official forms usually consist of several words, are subject in texts to substantial reduction, sometimes to a single word, especially in anaphoric occurrences, as in example (4.14).

(4.14) *w* [*Sojuszu Lewicy Demokratycznej*]$^{\text{Sojusz Lewicy Demokratycznej}}_{\text{orgName}}$ *[. . .] zaczęły się procesy naprawcze [. . .]* [***Sojusz***]$^{\textbf{Sojusz}}_{\text{orgName}}$ *składał te obietnice*
'reforms started within the [Democratic Left Alliance]$^{\text{Democratic Left Alliance}}_{\text{orgName}}$ [. . .]
the [Alliance]$^{\text{Alliance}}_{\text{orgName}}$ made those promises'

According to the general NE annotation rules in NKJP, determining the base form of a NE was only performed on the inflectional level, i.e. boiled down to establishing the nominative (usually singular) form, while all other variation aspects were maintained. In particular, we did not try to restore full forms for elliptical variants, as for *Sojusz* in example (4.14).

A notable exception concerned street names appearing in "extreme" elliptical variants, cf. (Savary et al., 2009). While head shifting discussed in example (3.9), p. 55, is a frequent behavior in Polish NEs, street names with a genitive complement do not follow this rule. Despite the omission of their headword *ulica* 'street' or *aleja* 'avenue', their complements remain in genitive instead of adapting their case to the context. Thus, in example (4.15) the street name *Chałubińskiego* does not take the instrumental form *Chałubińskim* despite the case requirement imposed by the preceding preposition *przy* 'at'. Consequently, the base form of this elliptical name remains in genitive, as shown in the topmost tree node. The nominative base form *Chałubiński* appears at the two lower levels, where an embedded person name is represented by a usual hierarchy of nodes, as discussed in the preceding section. As a conclusion, the token *Chałubińskiego* takes – here again, like in example (4.11) – a different base form at the morphosyntactic annotation level than it does at the NE level.

(4.15)



[. . .] w willi „Palace" przy Chałubińskiego od kilku lat [. . .].

IN WILLA "PALACE" AT CHAŁUBIŃSKI$_{sg:gen:m1}$ FROM SEVERAL YEARS

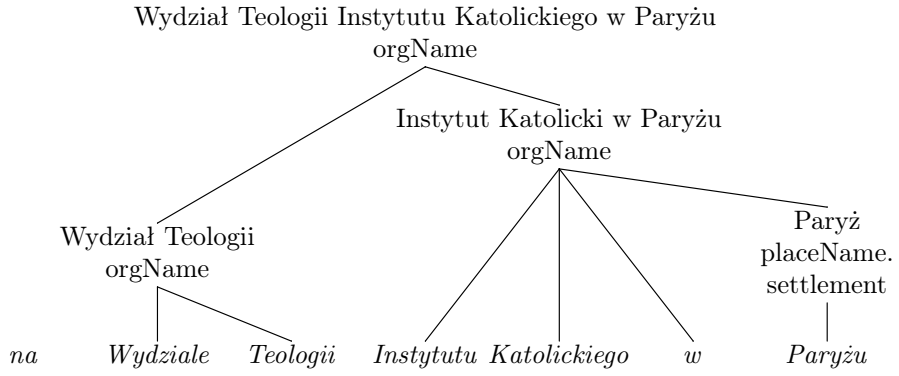
'in the "Palace" villa at the Chałubiński Street'

Ellipsis and Nesting Ambiguities

Recall that the NKJP NE annotation concerned both the maximum-length and the nested NEs. Nesting was understood semantically rather than syntactically, thus distinguishing nested structures may have been problematic.

As a general rule, if a name fragment corresponded to a clearly delimited administrative unit of a bigger organization it was annotated as a nested structure, as in example (4.16). Otherwise it received no individual tree node. In some cases, the application of this rule required a rich world knowledge. In example (4.17), despite syntactic similarity to (4.16), the parliamentary club is not a separate organization within a political party (here: *Platforma Obywatelska* 'Civic Platform') but denotes those members of parliament who belong to the party. It should not, thus, be annotated as a nested NE.

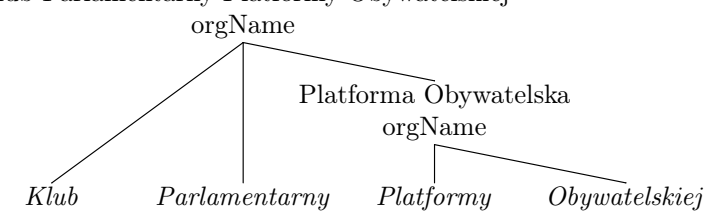
(4.16)



ON FACULTY$_{loc}$ THEOLOGY$_{gen}$ INSTITUTE$_{gen}$ CATHOLIC$_{gen}$ IN PARIS$_{loc}$

'in the Faculty of Theology of the Catholic Institute in Paris'

(4.17)



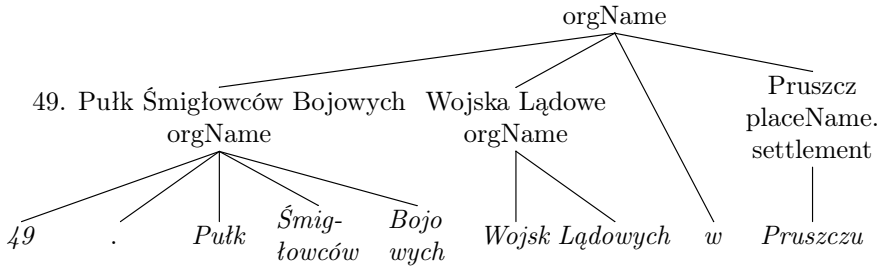
CLUB$_{nom}$ PARLIAMENTARY$_{nom}$ PLATFORM$_{gen}$ CIVIC$_{gen}$

'Parliamentary Club of the Civic Platform'

Identifying nested NEs was particularly difficult if nesting co-occurred with ellipsis. In examples (4.18)–(4.20) the same sequence *49. Pułk Śmigłowców Bojowych* 'the 49th Regiment of Attack Helicopters' was annotated in three different ways. In (4.18) it was nested in a larger

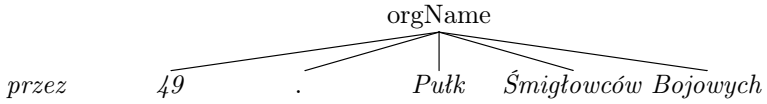
organization name, in (4.19) it denoted a separate organization, and in (4.20) it did not receive a separate tree node. These discrepancies resulted from ellipsis. Only case (4.18) consists in the full name[21] of a military unit. Example (4.19) shows an elliptical variant of this full form, with no mention of a superior organization. Finally, case (4.20) contains another elliptical variant, with an embedded place name. This last example is, thus, similar to (4.17) in that the regiment is not a part of the Pruszcz city (in the sense of an organization), like the parliamentary club is not a part of a political party.

(4.18) 49. Pułk Śmigłowców Bojowych Wojsk Lądowych w Pruszczu



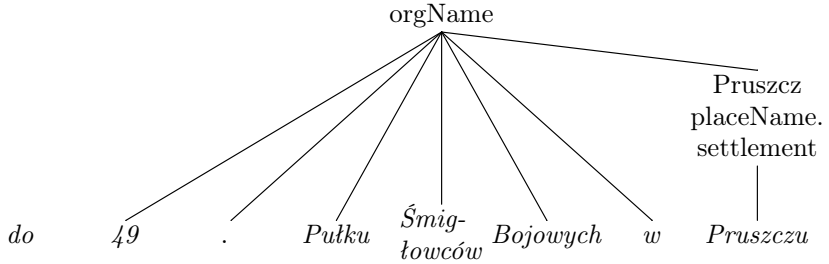
49-TH. REGIMENT$_{nom}$ HELICOPTERS$_{gen}$ ATTACKING$_{gen}$ ARMIES$_{gen}$ LAND-ADJ$_{gen}$ IN PRUSZCZ$_{loc}$
'the 49th Regiment of Attack Helicopters of the Land Forces in Pruszcz'

(4.19) 49. Pułk Śmigłowców Bojowych



BY 49-TH. REGIMENT$_{acc}$ HELICOPTERS$_{gen}$ ATTACKING$_{gen}$
'by the 49th Regiment of Attack Helicopters'

(4.20) 49. Pułk Śmigłowców Bojowych w Pruszczu



TO 49-TH. REGIMENT$_{gen}$ HELICOPTERS$_{gen}$ ATTACKING$_{gen}$ IN PRUSZCZ$_{loc}$
'to the 49th Regiment of Attack Helicopters in Pruszcz'

From the perspective of a more recent experience in coreference annotation (cf. Section 4.6), we think that a wrong decision was probably taken concerning delimiting names of administrative units of organizations as embedded names, as in example (4.16). Namely, the *Faculty of Theology* is just an ellipsis of the full name *Faculty of Theology of the Catholic Institute in Paris* in the sense that they describe the same referent. Thus, the annotation trees in examples (4.16) and (4.18) should be analogous to example(4.17) (i.e. with the leftmost intermediate node deleted). Correcting these cases is envisaged.

[21]The official name is even more complex: *49. Pułk Śmigłowców Bojowych Wojsk Lądowych Rzeczpospolitej Polskiej w Pruszczu Gdańskim.* 'the 49th Regiment of Attack Helicopters of the Land Forces of the Polish Republic in Gdańsk Pruszcz'

Metonymy, Ellipsis and Nesting

According to (Polański, 1993), metonymy is a syntagmatic deviation based on reduction. It appears when a syntactic position p is not occupied by an appropriate expression $e1$ but by another expression $e2$ such that: (i) $e2$'s semantics is incompatible with p, (ii) $e2$ remains in a syntactic relation with $e1$.[22] This definition shows intimate connections of metonymy with ellipsis (reduction) on the one hand, and with nesting (if $e1$ and $e2$ are NEs) on the other hand.

For instance, in sentence (4.21), the country name *Niemcy* 'Germany' ($e2$) is a metonymy since: (i) it cannot, in its original meaning, be linked with a verb requiring a human subject, (ii) it stands for the 'Germany National Football Team' ($e1$). By NKJP NE annotation rules, NEs were always to receive types and subtypes corresponding to the particular, possibly metonymic, contexts. Therefore, *Niemcy* is assigned here the `orgName` type instead of `placeName`→`country`.

(4.21) $[Niemcy]^{\text{Niemcy}}_{\textbf{orgName}}$ *pokonały* $[Kazachstan]^{\text{Kazachstan}}_{\textbf{orgName}}$ 3–0.
'Germany defeated Kazakhstan 3–0'

Other sports-related NEs are good examples of complex inter-dependencies between metonymy, ellipsis and nesting. For instance football club names are frequently formed by apposition of proper names and/or acronyms, as in *Wisła Kraków* 'literally: Vistula Cracow', *Bayern Monachium* 'Bayern Munich', *FC Porto*, etc. It is not quite clear if the geographical and place names are to be annotated as nested NEs, like in examples (4.22), (4.24) and (4.25), or as units concerned by metonymy, like in (4.23), (4.25) and (4.26).

(4.22) $[FC\ [Porto]^{\text{Porto}}_{\text{placeName.settlement}}]^{\text{FC Porto}}_{\text{orgName}}$

(4.23) $[FC\ Porto]]^{\text{FC Porto}}_{\text{orgName}}$

(4.24) $[[Wisła]^{\text{Wisła}}_{\text{geogName}}\ [Kraków]^{\text{Kraków}}_{\text{placeName.settlement}}]^{\text{Wisła Kraków}}_{\text{orgName}}$

(4.25) $[Wisła\ [Kraków]^{\text{Kraków}}_{\text{placeName.settlement}}]^{\text{Wisła Kraków}}_{\text{orgName}}$

(4.26) $[Wisła\ Kraków]^{\text{Wisła Kraków}}_{\text{orgName}}$

If nesting-based annotations (4.22) and (4.24) or (4.25) are chosen, then *Porto* alone could be seen as an ellipsis of *FC Porto* by analogy to example (4.15). Consequently, annotation (4.27) might be preferred over (4.28), which, however, contradicts the analogy to example (4.21).

(4.27) *W niedzielnym meczu* $[[Porto]^{\text{Porto}}_{\text{placeName.settlement}}]^{\text{Porto}}_{\text{orgName}}$ *wygrało na wyjeździe z Beira Mar 1:0.*
'In Sunday's away game Porto defeated Beira Mar 1:0'

(4.28) *W niedzielnym meczu* $[Porto]^{\text{Porto}}_{\text{orgName}}$ *wygrało na wyjeździe z Beira Mar 1:0.*

Within the NKJP project we did not manage to solve these contradictions. Annotations of types (4.22)–(4.26) may currently co-occur in the corpus.

[22] Author's translation and paraphrase.

Ambiguity of Derivational Bases

As mentioned in Section 4.3.1, one of novel aspects in NKJP NE annotation was the fact of taking relative adjectives, as well as inhabitant and organization member names into account. These items were assigned *derivational bases*, i.e. the names of persons, organizations, places or geographical objects from which the annotated items were derived. The derivational bases were semantically rather than morphologically motivated, thus e.g. the adjective *amerykański* 'American' was frequently attributed the derivational base *Stany Zjednoczone* 'United States' rather than *Ameryka* 'America'. Some cases were clearly of a fuzzy nature, e.g. the derivational base in example 4.29) might be replaced by *Stany Zjednoczone* 'United States' with no harm to the overall meaning.

(4.29) *Magazyn podał w środę wieczorem czasu* $[amerykańskiego]^{\text{amerykański, }\mathbf{Ameryka}}_{\text{geogName}}$ [...].
'The magazine announced on Wednesday evening American time [...].'

Determining the proper entity whom a derivative was related to was not always straightforward, notably due to multi-word geographical names with an identical or a similar headword. For instance, the adjective *ostrowski* 'Ostrów-related' refers, according to Kubiak-Sokół & Łaziński (2007), to at least 10 different Polish settlements: *Ostrów Wielkopolski*, *Ostrów Lubelski*, *Ostrów Mazowiecka*, *Ostrów*, *Ostrowo*, *Ostrowo Kościelne*, *Ostrowo Mogileńskie*, *Ostrowsko*, *Ostrowy nad Okszą* and *Ostrowy Tuszowskie*. Some contextual clues, as in example (4.30) helped filter dictrict headquater settlements only (here: *Ostrów Mazowiecka* or *Ostrów Wielkopolski*). In many cases the actual referent remained unknown due to the fact that the 1-million NKJP subcorpus contains randomly selected text samples of one-paragraph length.

(4.30) *Sąsiedni powiat* $[ostrowski]^{\text{ostrowski, }\mathbf{Ostrów}?}_{\text{relAdj(placeName.settlement)}}$ *w tym temacie jest przodującym w całym kraju, a więc pieniądze można zdobyć, trzeba tylko chcieć.*
NEIGHBORING DISTRICT OSTRÓW-ADJ IN THIS SUBJECT ...
'The neighboring Ostrów (?) is a leading district is this matter, so money can be found; where there is a will, there is a way.'

Even if the proper referent was easy to determine, the choice of the appropriate lemma for the derivational base could be an issue. For instance, in case of countries, a derivational base could correspond to the full official name – e.g. *Zjednoczone Królestwo Wielkiej Brytanii i Irlandii Północnej* 'The United Kingdom of Great Britain and Northern Ireland', *Republika Czeska* 'Czech Republic', *Stany Zjednoczone Ameryki Północnej* 'the United States of America' – or to its more commonly used abbreviated variant: *Wielka Brytania* 'Great Britain', *Czechy* 'Czechia', *Stany Zjednoczone* 'the United States'. For the sake of annotation coherence, we established a list of normalized (usually abbreviated) derivational bases to be systematically used in NKJP (provided that their subtypes had been previously fixed to `country`) – cf. examples 4.31–4.34.

(4.31) *amerykański* ← *Stany Zjednoczone* 'American ← United States'

(4.32) *angielski*, *brytyjski* ← *Wielka Brytania* 'English, British ← Great Britain'

(4.33) *południowoafrykański* ← *Republika Południowej Afryki* 'South African ← Republic of South Africa'

(4.34) *sowiecki*, *radziecki* ← *Związek Radziecki* 'Soviet ← Soviet Union'

Note that normalization problems of this kinds disappear as soon as we perform not only the named entity annotation/recognition but the **entity linking** as well. The latter task (cf

Section 4.2) consists in attaching NE occurrences in text to nodes of an external ontology, whose central items are semantic objects rather than linguistic labels naming these objects. In this context, the term "named entity" gains its literal meaning since what is really looked for are the semantic referents (entities in the sense of the Semantic Web) rather than just particular surface realizations of their names.

Ambiguous NE Span

Determining the left and right boundaries of an NE in texts is a well known issue in automatic named entity recognition. Even in the process of manual annotation, this problem may be hard to solve.

For instance, common nouns *dzień, godzina, rok, stulecie, era* 'day, hour, year, century, era' were considered integral parts of temporal NEs, as in example (4.35). The preceding prepositions, as a rule, were excluded from the NE span – cf (4.36)–(4.37). Note, however, that this rule could sometimes be counter-intuitive since adverbial expressions (4.35) and (4.37) are equivalent in many contexts.

(4.35) $[\textbf{\textit{dnia}}\ \textit{15 maja 2002}\ \textbf{\textit{r.}}]^{\text{2002-05-15}}_{\text{date}}$
[DAY$_{gen}$ 15 MAY$_{gen}$ 2002 Y.]
'the 15th of May, 2002'

(4.36) $\textbf{\textit{do}}\ [\textit{dnia 15 maja 2002 r.}]^{\text{2002-05-15}}_{\text{date}}$
UNTIL [DAY$_{gen}$ 15 MAY$_{gen}$ 2002 Y.]
'**until** [the 15th of May, 2002]'

(4.37) $\textbf{\textit{w}}\ [\textit{dniu 15 maja 2002 r.}]^{\text{2002-05-15}}_{\text{date}}$
IN [DAY$_{loc}$ 15 MAY$_{gen}$ 2002 Y.]
'**on** [the 15th of May 2002]'

In organization names, the entity boundaries should agree as far as possible with official names. In particular, place names may or may not participate in names of geographical objects, buildings, institutions, etc. In example (4.38), the city name is clearly a part of the building name but in more fuzzy cases the inclusion of the settlement name might be questionable, as illustrated in alternative annotations (4.39)–(4.40).

(4.38) $w\ [\textit{Muzeum Narodowym}\ w\ [\textit{Warszawie}]^{\text{Warszawa}}_{\text{placeName.settlement}}]^{\text{Muzeum Narodowe w Warszawie}}_{\text{geogName}}$
'in the [National Museum in Warsaw]'

(4.39) $w\ [\textit{Centrum Handlowym}]^{\text{Centrum Handlowe}}_{\text{geogName}}\ w\ [\textit{Czeladzi}]^{\text{Czeladź}}_{\text{placeName.settlement}}$
'in the [Shopping Mall] in [Czeladź]'

(4.40) $w\ [\textit{Centrum Handlowym}\ w\ [\textit{Czeladzi}]^{\text{Czeladź}}_{\text{placeName.settlement}}]^{\text{Centrum Handlowe w Czeladzi}}_{\text{geogName}}$
'in the [Shopping Mall in [Czeladź]]'

4.3.4 Inter-Annotator Agreement in Tree Structures

The inter-annotator agreement is a classical quality indicator for the results of an annotation task: (i) the clearer and the more detailed the annotation guidelines are, the fewer ambiguities, underspecifications and contradictions need to be resolved by the annotators, (ii) the better the project methodology is, the clearer the annotation procedures and requirements are, and the better the chance of coherent actions among independent annotators. This indicator also allows

to estimate the cost of the super-annotation: the higher it is, the less discrepancies need to be revised by an adjudicator.

The inter-annotator agreement, despite its intuitive simplicity, is a rather complex notion in an annotation task like ours, mainly due to multi-word and nested names. The weighted kappa measure (Cohen, 1960) is not easily applicable here because it assumes that the units to be annotated are known beforehand. In our task, annotators first have to identify the boundaries of existing names before they categorize them, thus there is no a priori list of units for which different annotations are to be compared. (Bejček & Straňák, 2010) describe similar considerations on annotating multi-word expressions. Therefore, we use simpler classical information retrieval measures. If annotators *a1* and *a2* have annotated the same corpus text, we admit that annotations produced by *a1* constitute the reference corpus and calculate the precision and the recall of *a2* with respect to this reference. Note that if we invert the roles of both annotators, we obtain complementary results: the precision (recall) of *a2* with respect to *a1* is equal to the recall (precision) of *a1* with respect to *a2*. Thus, the F_1-measure of *a1* w.r.t. *a2* is equal to the F_1-measure of *a2* with respect to *a1*.

Estimating the precision/recall, however, is not quite straightforward in our task. As mentioned before, each NE is assigned an annotation tree whose leaves are segments from the morphosyntactic annotation level, the tree's height can exceed 1 and its every node obtains a set of attributes (syntactic and semantic head, lemma, derivational base etc.). We assume that an annotation tree node is correct with respect to the reference corpus if the latter contains a node which:

- has the same attributes
- covers the same, possibly non-adjacent, segments at the morphosyntactic level

Thus, a parent node may be correct even if its subnodes are incorrect or incomplete. Consider, for instance, the named entity in example (4.41):

(4.41) Instytutu Podstaw Informatyki Polskiej Akademii Nauk
INSTITUTE$_{sg:gen:m3}$ FOUNDATIONS$_{pl:gen:f}$ INFORMATICS$_{sg:gen:f}$ POLISH$_{sg:gen:f}$ ACADEMY$_{sg:gen:f}$ SCIENCES$_{pl:gen:f}$
'Institute of Computer Science of the Polish Academy of Sciences'

and suppose that:

- annotator *a1* has created a node of type *orgName* covering all 6 words with the lemma *Instytut Podstaw Informatyki Polskiej Akademii Nauk*, and an embedded node of type *orgName* covering the last 3 words with the lemma *Polska Akademia Nauk* 'Polish Academy of Sciences':
[*Instytutu Podstaw Informatyki* [*Polskiej Akademii Nauk*]$_{\text{orgName}}$]$_{\text{orgName}}$
- annotator *a2* has created the same nodes as *a1* but, additionally, he also created an embedded *orgName* node covering the first three words with the lemma *Instytut Podstaw Informatyki* 'Institute of Computer Science'
[[*Instytutu Podstaw Informatyki*]$_{\text{orgName}}$ [*Polskiej Akademii Nauk*]$_{\text{orgName}}$]$_{\text{orgName}}$

We assume then that two out of three named entities have been correctly annotated by *a1* w.r.t. *a2* ($P_1 = 1$, $R_1 = 2/3$). If, however, *a1* made a mistake in the lemma of a name (e.g. **Polska Akademia Nauka*) then only one name will be considered correct ($P_1 = 1/2$, $R_1 = 1/3$).

Results of the inter-annotator agreement based on the above assumptions are given in Table 4.2. *Persons*, which are the most numerous NEs, correspond to all NEs of type `persName`,

and possibly any of its 3 subtypes. *Locations* represent all NEs of types `geogName` or `placeName` (and any of its 5 subtypes). *Organizations* relate to type `orgName`. *Temporal expressions* designate types `date` and `time`. Finally, *derivations* embrace relative adjectives, inhabitant and organization member names.

Named entities					
Persons	Locations	Organizations	Temporal expressions	Derivations	Overall
0.89	0.78	0.69	0.88	0.71	0.83

Table 4.2: Inter-annotator agreement results

These results are reasonably high, given the fact that NE annotation is largely of a semantic nature, and that the admitted agreement criteria are rather severe (partial agreement counts as non agreement).

4.4 Named Entity Recognition with Multi-Word and Nested Structures

The NE annotation level of NKJP was supported by two named entity recognition (NER) tools. A unification grammar and gazetteers developed for the ***SProUT*** platform were used in automatic pre-annotation prior to the manual correction and adjudication of the 1-million part of the corpus, as mentioned in Section 4.3.2. A novel machine learning-based tool ***Nerf*** was trained on the 1-million subcorpus and further used for automatic annotation of the whole 1.5-billion word main corpus. In this section, we summarize the construction of these two tools and stress the development challenges resulting from **multi-word** and **nested NEs**.

4.4.1 Rule-Based Named Entity Recognition with Multi-Word and Nested Structures

SProUT (Becker et al., 2002; Drożdżyński et al., 2004) is a general purpose multi-lingual NLP platform. It is equipped with a set of reusable Unicode-capable processing components for basic linguistic operations (a tokenizer, a sentence splitter, a morphological analyzer, a gazetteer look-up component, etc.) and a cascaded unification-based finite-state grammar parser and interpreter.

SProUT has been adapted to Polish (Piskorski et al., 2004), and grammars for extracting 'classical' named-entities (e.g., names of persons, organizations, locations, etc.) from Polish texts have been developed (Piskorski, 2005). If this Polish-oriented SProUT version, originally meant for information extraction tasks, were to be applied to NKJP pre-annotation, it had to be modified so as to fit the annotation guidelines (Savary & Piskorski, 2010, 2011).

Firstly, the SProUT morphological and semantic lexicons (called gazetteers) were extended with new (partly multi-word) entries, and with their inflected forms, whenever they were known to the morphological generator of the Morfeusz system (cf. Section 3.3.5). Example (4.42) shows a sample (slightly simplified) gazetteer entry describing the instrumental form of a masculine surname *Kowalski*. The whole gazetteer contains over 289,000 inflected forms for 54,000 lemmas. Its size is comparable to the 45,000 entry English gazetteer used by Mikheev et al. (1999) but our entries are assigned relatively large lists of grammatical and semantic features. Our gazetteer (with the exception of some proprietary data concerning derivation forms) was exported into a custom LMF format (Savary, 2012) and made available[23] under the 2-clause BSD license[24].

[23]The Polish Named Entity Gazetteer (PNEG) is downloadable from `http://clip.ipipan.waw.pl/Gazetteer`

[24]`http://en.wikipedia.org/wiki/BSD_licenses`

(4.42) **Kowalskim** | GTYPE:gaz_surname | G_LEMMA:Kowalski | G_GNUMBER:singular | G_CASE:ins | G_GENDER:masc1

Secondly, the hierarchy of concepts was redesigned so as to: (i) fit the taxonomy shown in Figure 4.2, (ii) introduce new gazetteer attributes such as G_DERIV_TYPE or G_DERIVED_FROM (for derivation type and base), (iii) conflate all NE-types into one main `ne-nkjp` type, (iv) complete the `ne-nkjp` type with a special attribute TREE meant to accumulate data about nested NE structures.

Finally, SProUT grammars had to be thoroughly reviewed and modified in order to serve the annotation task. Handling nested NEs (which was missing in the original grammar) belonged to the main challenges of this process. Figure 4.6 shows a (slightly simplified) grammar rule, named `surname_gaz_based`. A SProUT rule has a left-hand side (LHS) and a right-hand side (RHS), separated by `->`. The LHS is a regular expression over typed feature structures (TFS), representing the recognition pattern, and RHS is a TFS specification of the output structure. Additionally, functional operators may be used on both sides of the rules. They provide a gateway to the outside world, and are primarily utilized for forming the output of a rule and for introducing complex constraints. The symbol `&` denotes unification, and variables are strings preceded by the symbol `#`. Here, the LHS allows to recognize any gazetteer entry provided that it is a surname. The RHS triggers creation of an `ne-nkjp` structure. Seven slots are assigned values. In particular, `SURFACE`, `BASE` and `MORPH` are directly instantiated, via variables, with the corresponding attributes from the gazetteer entry. Attributes `CSTART` and `CEND` denote the starting and the ending character numbers of the recognized sequence in the text, and are straightforwardly instantiated by the analyzer. The `TREE` attribute is used for storing **nested annotations** that are transformed into trees in the annotated corpus. Its value is created by the functional operator `ConcWithBlanks`, which concatenates (with separating blanks and bars) the surface form, the lemma, the beginning and ending character number. For instance, the value of this attribute for the occurrence of the inflected form from entry (4.42) at position 127 in the input text would be `[ Kowalskim | Kowalski | surname | 127 | 135 ]`. A similar rule exists for the recognition of names of all other types appearing in the gazetteer (forenames, cities, organizations, etc.).

```
surname_gaz_based :/ gazetteer & [SURFACE #surface, G_LEMMA #lemma,
                                  GTYPE gaz_surname,G_NUMBER #number,
                                  G_CASE #case, G_GENDER #gender,
                                  CSTART #s, CEND #e]
->
ne-nkjp & [SURFACE #surface, BASE #lemma, NE_TYPE surname,
           MORPH  agr-nkjp & [NE_NUMBER #number,
                              NE_CASE #case,NE_GENDER #gender],
           TREE #tree, CSTART #s, CEND #e],
where #tree=ConcWithBlanks("[", #surface, "|", #lemma, "| surname |", #s, "|", #e]").
```

Figure 4.6: Grammar rule for the recognition of a surname, e.g. *Kowalskim*, belonging to the gazetteer

Grammar rules can be recursively embedded. Fig. 4.7 shows the `person_1` rule for recognition of person names. First, an optional ('?' denotes optionality) position and title are matched, via a call to adequate rules: `@seek(full_position)` and `@seek(title)`. Next, one or two forenames are sought: `@seek(forename)`. Finally, a surname is consumed by an embedded rule roughly equivalent to Fig. 4.6. In the resulting `ne-nkjp` structure the `SURFACE` slot is created via concatenation of the forenames and the surname (by a call to `ConcWithBlanks`), whereas the `BASE` collects base forms on the LHS. The attribute `TREE` is a list of the `TREE` values of the

embedded names, followed by the description of the whole structure. For instance, matching the text fragment *Prezydentem Janem Kowalskim* would result in producing the structure depicted in Figure 4.8. All output structures of this type obtained for a given text were transferred to the Sproutput-to-PML converter, as explained in Section 4.3.2. As a result, the TREE attributes were transformed to TrEd annotation trees, as in Figure 4.4, ready to be edited by human annotators and super-annotators.

```
person_1 :> ((@seek(full_position) & #position])(token & [TYPE comma])?)?
            (@seek(title) & #title) ?
            (@seek(forename) & [SURFACE #surf1, BASE #lemma1, MORPH #morph,
                                TREE #tree1, CSTART #s1, CEND #e1])
            (@seek(forename) & [SURFACE #surf2, BASE #lemma2, MORPH #morph,
                                TREE #tree2, CSTART #s2, CEND #e2]]) ?
            (@seek(surname) & [SURFACE #surf3, BASE #lemma3, MORPH #morph,
                               TREE #tree3, CSTART #s3, CEND #e3] & #surname)
            (@seek(name_suffix) & #suffix)?
->
ne-nkjp & [SURFACE #surface, BASE #lemma, TYPE persName,
           TREE #tree, CSTART #s1, CEND #e3],
where #surface = ConcWithBlanks(#surf1, #surf2, #surf3),
      #lemma = ConcWithBlanks(#lemma1, #lemma2, #lemma3),
      #tree = ConcWithBlanks(#tree1,#tree2,#tree3,
              "[",#surface,"|",#lemma,"| persName |",#s1,"|",#e3," ]").
```

Figure 4.7: Grammar rule for the recognition of a person name, with embedded rule calls

In (Savary & Piskorski, 2011) we also show how competing TFSs for the same sequence were handled, how the so-called internal and external evidences of NEs were included in the type hierarchy and in grammar rules, and how special functional operators allow us to synthesize multi-word NE lemmas whose components are not lemmas themselves (e.g. *Najwyższa Izba Kontroli* 'Supreme Chamber of Control').

The final adapted SProUT grammar consist of 120 rules, most of them specific to a particular NE type or subtype, and only 10% generic. For instance a generic rule `capitalized_adj_with_special_stem` allows to create the lemma of organization or place names containing common names. The common names in this lemma must be capitalized and in the correct gender, e.g. *Morze Martwe* 'Dead$_{neut}$ Sea$_{neut}$' instead of *Morze martwy* 'dead$_{masc}$ Sea$_{neut}$'.

The detailed evaluation of the grammar was performed on an NKJP subcorpus containing about 56,000 NEs. Those texts were chosen for evaluation whose manual annotation was performed after the stabilization of the grammar and the gazetteer. Two different criteria guided

SURFACE	Janem Kowalskim
BASE	Jan Kowalski
TYPE	persName
	[Janem | Jan | forename | 121 | 125]
TREE	[Kowalskim | Kowalski | surname | 127 | 135]
	[Janem Kowalskim | Jan Kowalski | persName | 121 | 135]
CSTART	121
CEND	135

Figure 4.8: Structure resulting from processing the text *Prezydentem Janem Kowalskim* by the rule `person_1`.

the evaluation design: (i) taking all existing, notably nested, NE occurrences into account, or the maximum-length occurrences only, (ii) considering all attributes (text span, type, subtype, lemma, normalized date form, derivation type and base) or only the first three of them. Consequently, four sets of results were obtained whose general conclusions are the following:

- the overall precision varies from 68% to 78%, and the overall recall from 35% to 39%,
- the results are, obviously, higher if only tokens, types and subtypes are considered than if all attributes are taken into account; differences between these two scenarios range from 2% to 13% of precision and from 2% to 5% of recall,
- the best results are obtained for temporal expressions and the worst for organizations.

Interestingly enough, the results for persons and locations are significantly better when all NEs, including the nested ones, are considered than if only the longest-match occurrences are taken into account. This might be due to at least two factors. Firstly, the specific annotation guidelines for person names (with at least two nodes for single-word NEs) artificially increase the number of easy-to-spot nested entities (cf. Section 4.3.3). Secondly, longer names offer disambiguating contexts for shorter (nested) ones. For instance, the person name *Jurek* appears in the gazetteer both as a forename and as a surname but the context of the sentence (4.43), on which the rule $\texttt{person}_1$ in Figure 4.7 can be triggered, allows us to reject the former and retain the latter analysis.

(4.43) *W sprawie formalnej zgłasza się pan poseł* $[[Marek]^{\text{Marek}}_{\text{persName.forename}}\ [Jurek]^{\text{Jurek}}_{\text{persName.surname}}]^{\text{Marek Jurek}}_{\text{persName}}$.
'The deputy Marek Jurek is submitting a formal issue.'

In (Savary & Piskorski, 2011) we also perform a detailed qualitative analysis of errors produced by our adapted SProUT grammar. Interesting problems are revealed concerning, notably, the lemmatization of compound person names, street names and plurale tantum geographic names.

4.4.2 Machine Learning and Named Entity Recognition with Multi-Word and Nested Structures

Automatic annotation of the whole 1.5-billion word NKJP corpus was performed with a tool based on machine learning (ML), developed by Jakub Waszczuk and trained on the manually annotated 1-million word subcorpus (Savary & Waszczuk, 2012; Waszczuk et al., 2013). This baseline version[25] later evolved into ***Nerf***, which contains modules of external lexicon lookup compatible with a list of Polish NE resources, including the SProUT gazetteer described in the previous version, the Polish Named Entity Triggers (PNET)[26] and the Polish Prolexbase module (see Section 4.5.1). This section summarizes some aspects of the baseline Nerf version related to the recursive embedding of NE structures.

Nerf is based on the method called *Joined Label Tagging* introduced by Alex et al. (2007). It reduces the problem of NER with nested structures into the sequential tagging problem using the classical IOB tags (cf. Section 4.2.2). In this method, a NE-tagged sentence is represented by a sequence of tags, where *B-t* represents a token at the beginning of an NE of type *t*, *I-t* – a token inside an NE of type *t* and *O* – a token outside any NE. For instance, the sentences (4.44) and (4.46) from examples (4.4) and (4.5) are represented by the sequences of labels in (4.45) and (4.47), respectively.

[25] Downloadable at `http://zil.ipipan.waw.pl/Nerf?action=AttachFile&do=view&target=nerf.dist.0.2.tgz` and usable under the terms of the GPL v3 license.

[26] Downloadable from `http://zil.ipipan.waw.pl/PNET` and usable under the terms of the 2-clause BSD license.

(4.44) *w Paryżu dnia 21 września 1960 r.*
IN PARIS$_{loc}$ DAY$_{gen}$ 21 SEPTEMBER$_{gen}$ 1960 Y.
'in Paris on the 21st of September 1960'

(4.45) [O, B-settlement, B-date, I-date, I-date, I-date, I-date]

(4.46) *imprezy ogólnopolskie rangi juniorskiej*
EVENTS GENERALLY-POLISH RANK$_{gen}$ JUNIOR-ADJ$_{gen}$
'Polish-national junior-level events'

(4.47) [O, B-country@reladj, O, O]

As far as **multi-word discontinuous NEs** are concerned, the same encoding can be used, whereas it is possible to have an *I-t* directly after an *O* label, which should never happen in continuous NEs. For instance, if subtypes and nesting are disregarded, the sentence (4.48) from example (4.13) is represented by the sequence of labels in (4.49).

(4.48) *Gisela Froemel, z domu Kopka*
GISELA FROEMEL, FROM HOME KOPKA
'Gisela Froemel, born Kopka'

(4.49) [B-persName, I-persName, O, O, O, I-persName]

In the IOB encoding of **nested structures**, each nesting level is encoded separately and then labels from all levels are merged, whereas *O* labels merged with non-*O* ones are omitted. For instance, the sentence (4.50) from example (4.16) is represented by the sequence of labels in (4.51), while the full encoding of sentence (4.48) is given in (4.52).

(4.50) *na Wydziale Teologii Instytutu Katolickiego w Paryżu*
ON FACULTY$_{loc}$ THEOLOGY$_{gen}$ INSTITUTE$_{gen}$ CATHOLIC$_{gen}$ IN PARIS$_{loc}$
'in the Faculty of Theology of the Catholic Institute in Paris'

(4.51) [O, B-orgName#B-orgName, I-orgName#I-orgName, B-orgName#I-orgName, I-orgName#I-orgName, I-orgName#I-orgName, B-placeName#I-orgName#I-orgName]

(4.52) [B-forename#B-persName, B-surname#I-persName, O, O, O, B-surname#I-persName]

The only case where the IOB encoding cannot be used in accordance with the NKJP annotation rules is the one of coordinated NEs with overlapping, as in examples (4.9) and 4.11). These cases account for about 1% of all NEs in the 1-million word NKJP subcorpus. Since in Nerf they could not be represented as overlapping structures, they are not annotated as such in the whole 1-billion word corpus.

Nerf is based on the CRF (Conditional Random Fields) probabilistic model. It is trained on four annotation levels of the 1-million word NKJP subcorpus: text level, segmentation, morphosyntax and the NE level (converted into the joint-label representation as explained above). Each word of the training corpus is represented as a list of observations containing its:

1. orthographic form (`orth`) in lowercase,
2. prefixes and suffixes of length k, $k-1$, $k-2$, and $k-3$ (if non-empty), where k is the word's length (they are substitutes of lemmas),
3. suffixes of length 3, 4 and 5 (they help in modeling derivations),

4. shape, where each lowercase and uppercase letter is replaced by an *l* or an *u*, respectively, each digit is replaced by a *d*, and each non-letter and non-digit character by an *x*,

5. compressed shape, in which sequences of identical shape characters are squeezed into a single character, e.g. *ullxx* is compressed into *ulx*,

6. concatenated shapes of two neighboring tokens,

7. concatenated compressed shapes of two neighboring tokens.

The observation schema is based on unary features $(observation, label_i)$ concerning the observations 1–7 on the current word i only, and on binary features $(observation, label_{i-1}, label_i)$ taking observations 1–5 into account for both the current and the previous word.

Nerf, in its baseline version, was evaluated on the 1-million word NKJP subcorpus with a 5-cross validation. The results show an overall precision of 0.83, an overall recall of 0.76, and an overall F_1-measure of 0.79. Like for SProUT (cf Section 4.4.1), the best scores were obtained for temporal expressions and person names, and the worst for organization names. Nerf significantly outperforms SProUT in all NE categories. The differences between both tools vary from 3% (for temporal expressions) to 8% (for organization names) of precision, and from 11% (for derivations) to 46% (for person names) of recall.

These results show the complementary nature of rule-based and ML-based tools. The spectacular increase in NER quality in Nerf with respect to SProUT is, notably, due to the high quality of the 1-million word NKJP subcorpus, whose labor-intensive manual annotation and adjudication could be reduced to 10 person-months thanks to automatic pre-annotation with SProUT. Note also that SProUT grammars take all attributes of the matched NE into account, while Nerf can only assign them their types and subtypes, and ignores those attributes whose values are not limited to a closed set of labels (base forms, derivational bases, and normalized dates). We think, therefore, that a hybrid – data-driven and rule-based annotation method – would be the most useful for a future high quality automatic NER system. In such a model, a CRF-based tool would be mainly responsible for the identification and categorization of NEs, while the grammar rules could provide lemmas, normal forms and derivation bases of recognized NEs.

4.5 Named Entities as Concepts in a Multilingual Ontology

As previously discussed, the description of the internal structure of a compound NE, and of a MWU in general, paves the way for understanding its (non-)compositionality not only at morphosyntactic but also at semantic level. In particular, nested structures, which are NEs on their own, might straightforwardly contribute to compositional meaning calculation of their nesting entities in most examples in Section 4.3.3, in particular in (4.16), (4.17), (4.18), (4.38), etc.

Another way of representing the semantics of MWUs and NEs is their attachment to a general or domain-specific ontology. In this section I summarize my efforts towards a semi-automatic population of a multilingual ontology of proper names, ***Prolexbase***. A more detailed presentation of this work can be found in (Savary et al., 2013a) and (Savary et al., 2013b).

4.5.1 Prolexbase

Prolexbase (Krstev et al., 2005; Tran & Maurel, 2006; Maurel, 2008) offers a fine-grained multilingual model of proper names whose specificity is to be both concept-oriented and lexeme-oriented. Namely, it comprises a language-independent ontology of concepts referred to by proper names,

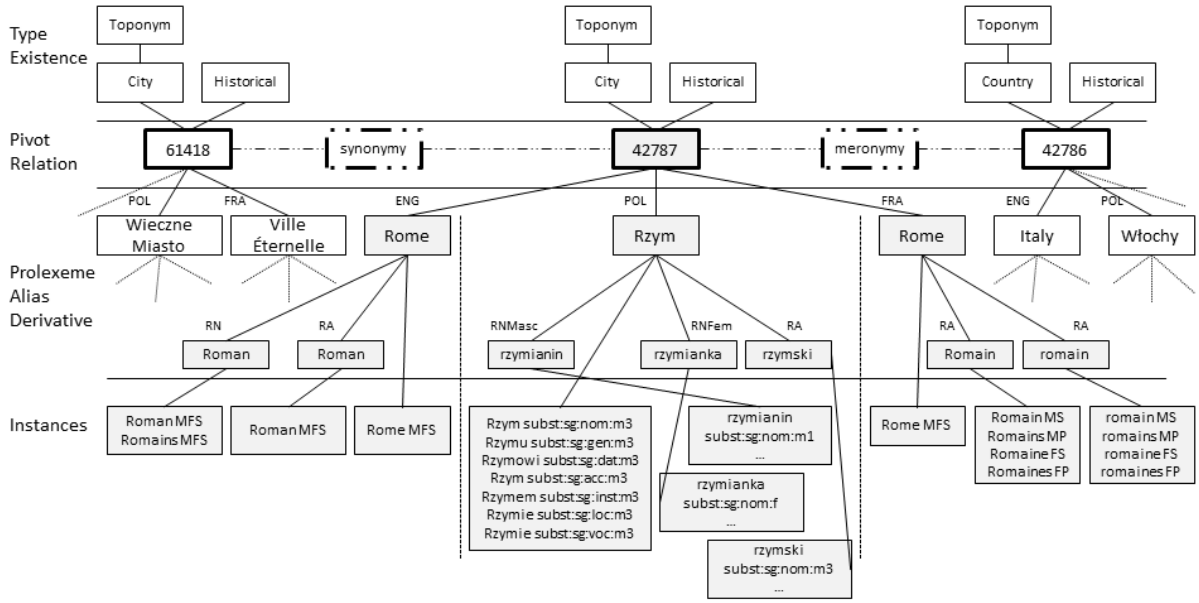


Figure 4.9: Extract of Prolexbase with four levels and three prolexemes in Polish, English and French.

as well as detailed lexical modules for proper names in several languages (French, English, Polish and Serbian being the best covered ones). Prolexbase is structured in four levels for which a set of relations is defined.

The **metaconceptual level** defines a two-level typology of four **supertypes** (anthroponym, toponym, ergonym and pragmonym) and 34 **types** (celebrity, association, country, product, disaster, etc.).

Some types have secondary supertypes, e.g. a city is not only a toponym but also an anthroponym and a pragmonym. The metaconceptual level contains also the **existence** feature which allows to state if a proper name referent has really existed (*historical*), has been invented (*fictitious*) or whether its existence depends on religious convictions (*religious*).

The originality of the **conceptual level** is twofold. Firstly, proper names designate concepts (called **conceptual proper names**), instead of being just instances of concepts, as e.g. in WordNet population by (Toral et al., 2008, 2012). Secondly, these concepts, called **pivots**, embrace not only objects referred to by proper names, but also points of view on these objects: *diachronic* (depending on time), *diaphasic* (depending on the usage purpose) and *diastratic* (depending on sociocultural stratification). For instance, although *Alexander VI* and *Rodrigo Borgia* refer to the same person, they get two different pivots since they represent two different points of view on this person. Each pivot is represented by a unique interlingual identification number allowing to connect proper names that represent the same concepts in different languages. Pivots are linked by three language-independent relations. **Synonymy** holds between two pivots designating the same referent from different points of view (*Alexander VI* and *Rodrigo Borgia*). **Meronymy** is the classical relation of inclusion between the meronym (*Samuel Beckett*) and the holonym (*Ireland*). **Accessibility** means that one referent is accessible through another one, generally better known (Tran & Maurel, 2006). The accessibility **subject file** with 12 values (*relative*, *capital*, *leader*, *founder*, *follower*, *creator*, *manager*, *tenant*, *heir*, *headquarters*, *rival*, and *companion*) informs us about how/why the two pivots are linked (*The Magic Flute* is accessible from *Mozart* as *creator*).

The **linguistic level** contains **prolexemes**, i.e. the lexical representations of pivots in a given language. For instance, pivot 42786 is linked to the prolexeme *Italy* in English, *Italie* in French and *Włochy* in Polish. There is a 1:1 relation between pivots and prolexemes within a language, thus homonyms (*Washington* as a celebrity, a city and a region) are represented by different prolexeme instances. A prolexeme can have language-dependent variations: **aliases** (abbreviations, acronyms, spelling variants, transcription variants, etc.) and **derivatives** (relational nouns, relational adjectives, prefixes, inhabitant names, etc.). The language-dependent relations defined at this level include, in particular: **classifying context** (the *Vistula river*), **accessibility context** (*Paris* — the *capital* of *France*), **frequency** (*commonly used*, *infrequently used* or *rarely used*), and **language** (association of each prolexeme to one language).

The **level of instances** contains inflected forms of prolexemes, aliases and derivatives, together with their morphological or morphosyntactic tags. These forms can either be materialized within Prolexbase itself or be represented by links to external morphological models and resources.

Figure 4.9, inspired by Krstev et al. (2005), shows an extract of the intended contents of Prolexbase containing the vicinity of the prolexeme *Rzym* 'Rome', in particular its pivot, stylistic synonym, meronym, derivatives, and instances.

The motivation behind Prolexbase is not to represent as many available names as possible, like in the case of other large automatically constructed ontologies such as YAGO (Suchanek et al., 2007) or DBpedia (Mendes et al., 2012). We aim instead at a restricted scope but a high quality, i.e. manually validated, incremental resource dedicated to NLP. This implies: (i) appropriate selection criteria for selecting only the most relevant, popular and stable names, (ii) data integration avoiding duplication of data, (iii) NLP-targeted features, particularly with respect to highly inflected languages. Prolexbase might, thus, correspond to the *kernel NE lexicon*, i.e. the common shared NE vocabulary appearing in texts of different dates, types and subjects, as opposed to the *peripheral NEs* used infrequently and in domain-specific jargons. As suggested by Saravanan et al. (2012), handling peripheral NEs might then rely on their co-occurrence with the kernel NEs.

Note that the rich Prolexbase model accounts for different aspects of variability in proper names. Firstly, aliases, inflected forms, relative adjectives and inhabitant names, when encountered in texts, can be recognized as different surface occurrences of the same underlying concept (represented by the pivot). Secondly, variability in time and aspect is represented by synonymy. Finally, variability in language can be resolved by reference to a common interlingual pivot.

4.5.2 Prolexbase Population from Open Sources

Prolexbase initially contained mainly French proper names, even if its model supports multilingualism. In order to extend its coverage of other languages we created *ProlexFeeder* (Savary et al., 2013a), a tool meant for a semi-automatic population of Prolexbase from Wikipedia and, to a lesser extent, from GeoNames.

Figure 4.10 shows the dataflow in our Prolexbase population process. Three main data sources were: (i) Polish, English and French Wikipedia, (ii) Polish names in GeoNames[27], (iii) Polish inflection resources in *Translatica* (Jassem, 2004), a Polish machine translation software.

We first automatically selected 1016 Wikipedia classes (including 340 relevant infobox templates and 676 person-related categories). Those were then manually mapped on the Prolexbase types and relations. For instance, the Polish Wikipedia category *Władcy Blois* 'counts of Blois' was assigned the Prolexbase type *celebrity*, *historical* existence, and *accessibility* relation with the pivot representing the town of *Blois* and the subject file *leader*.

[27] `http://www.geonames.org/`

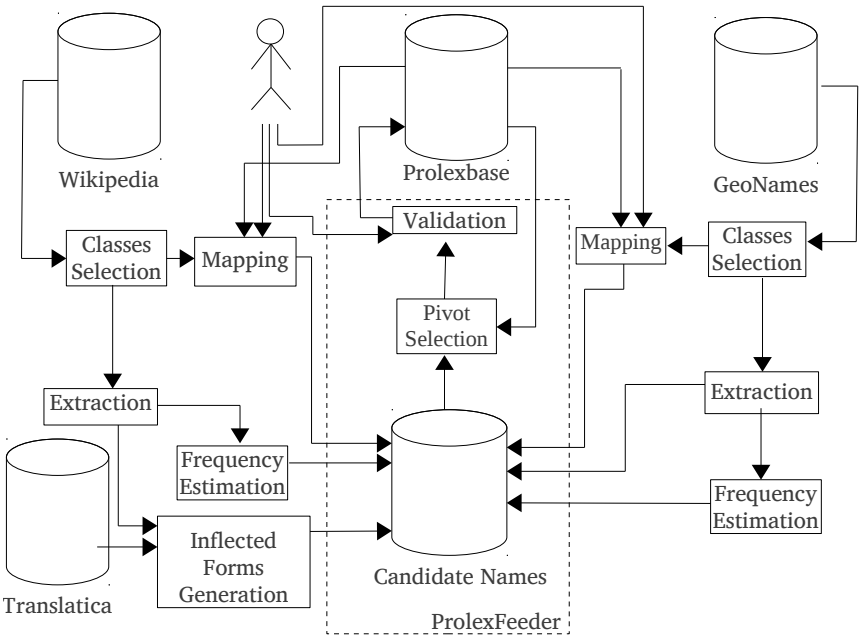


Figure 4.10: Data flow in Prolexbase population via ProlexFeeder.

The titles of articles belonging to the mapped classes (including Wikipedia redirects) were automatically extracted and their popularity (or frequency) was estimated from statistical data[28] on Wikipedia hits in 2010. For instance, the football club *Wisła Kraków*, obtained frequency code 1 (*commonly used*) in Polish (175,785 hits), and code 2 (*infrequently used*) in English (123,360 hits) and French (7,516 hits). Additional country names, Polish names and alternate names were selected from GeoNames.

Inflection modules from Translatica were used to automatically predict inflected forms of both simple and multi-word Polish entries.

The resulting set of candidate names was fed to ProlexFeeder pivot selection module, which automatically checked if the entity represented by a candidate entry was already present in Prolexbase. This process was based on a similarity function taking prolexemes, types, aliases and Wikipedia URLs into account. Table 4.3 shows a sample set of Wikipedia data resulting from the preprocessing described above.

Each entry, together with its translations, variants, relations and inflected forms was manually validated by an expert lexicographer. Figure 4.11 shows fragments of the validation interface for the data from Table 4.3. The pivot selection procedure has found the proper existing pivot. Special care had to be taken in a proper treatment of Wikipedia redirects. For instance, the redirect *ONZ* 'ONU' towards *Organizacja Narodów Zjednoczonych* 'United Nations Organization' should become an alias. Here, however, the redirect *Wieczne miasto* 'eternal city' had to be transformed into a new pivot related by diastratic synonymy to *Rzym* 'Rome'.

An evaluation performed on a sample of 150 entries of different types showed that ProlexFeeder predicts the correct (existing or new) pivot for an incoming entry with a 97.2% accuracy. On average, the manual correction and validation of an entry takes about 2 minutes. Most of this time is taken by completing and/or correcting the inflected forms of Polish prolexemes. Inflecting celebrity names proves the most labor-intensive since Translatica's automatic inflection tool makes some errors concerning person names. This confirms the hardness of their

[28] Available at `http://stats.grok.se/`

Table 4.3: Sample preprocessed Wikipedia data. The attributes represent: Wikipedia lexemes (*PL.lex*, *EN.lex*, *FR.lex*), the number of Wikipedia hits in 2010 (*PL.hits*, *EN.hits*, *FR.hits*), frequency (*PL.freq*, *EN.freq*, *FR.freq*), the Wikipedia page URL (*PL.url*, *EN.url*, *FR.url*), Wikipedia redirects proposed as aliases (*PL.aliases*, *EN.aliases*, *FR.aliases*), the predicted Polish inflected forms (*PL.infl*), predicted Prolexbase type, meronymy-related pivot (*meroPivot*), existence and pivot.

Attribute	**Value**	**Attribute**	**Value**	**Attribute**	**Value**
PL.lex	*Rzym*	EN.lex	*Rome*	FR.lex	*Rome*
PL.hits	315,996	EN.hits	3,160,315	FR.hits	450,547
PL.freq	1	EN.freq	1	FR.freq	1
PL.url	pl.wikipedia.org/ wiki/Rzym	EN.url	en.wikipedia.org/ wiki/Rome	FR.url	fr.wikipedia.org/ wiki/Rome
PL.aliases	*Wieczne miasto*	FR.aliases	*Ville Éternelle*, *Ville éternelle*	EN.aliases	*Capital of Italy*, *Castel Fusano*, *Città Eterna*, ...
PL. infl	*Rzymu:sg:gen:m3*, *Rzym:sg:acc:m3*, ...	type	city	existence	historical
		meroPivot	none	pivot	42787

automatic processing in Polish, addressed by Piskorski et al. (2007).

A first evaluation of Prolexbase application has been performed with Nerf (cf. Section 4.4.2). We used the named entity level of the manually annotated 1-million word NKJP corpus (cf Section 4.3) divided into 10 parts of a roughly equal number of sentences. In each fold of the 10-fold cross validation Nerf was trained once with no external resources (setting A), and once with the list of Polish Prolexbase instances and their types (setting B). Each setting admitted 20 training iterations. We considered an NE as correctly recognized by Nerf if its span and type matched the reference corpus. In setting A the model obtained the mean F_1 measure of 0.76819 (with mean $P = 0.79325$ and $R = 0.74477$), while in setting B the mean F_1 measure was equal to 0.77409 (with mean $P = 0.79890$ and $R = 0.75092$). The paired Student's t-test yielded the p-value equal to 0.0001145 which indicates that the results are statistically significant with respect to the the commonly used significance levels (0.05 or 0.01). It should be noted that the majority of names appearing in the NKJP corpus correspond to person names, while Prolexbase contains a relatively small number of such names. Conversely, settlement names (cities, towns, villages, etc.) constitute a relatively high percentage of Prolexbase entries. In this subcategory the enhancement of Nerf's scores is the most significant: the mean F-measure increased by 0.03894 (from $F_1 = 0.79202$ to $F_1 = 0.83096$) and the Student's t-test p-value was equal to $8.011e - 08$.

Table 4.4 shows the state of Prolexbase at the end of March 2013. The dominating role of toponyms is due to the initial contents of Prolexbase, which essentially focused on French geographical names. The most numerous types are city (48,340 pivots), celebrity (7,979 pivots), hydronym (4,580 pivots) and region (4,190 pivots), the number of pivots of the remaining types is between 1 and 1,374. Recall that one of original aspects of Prolexbase is the synonymy relation between pivots referring to the same object from different points of view. Currently, 3.35% of all pivots, mainly celebrities and countries, are in synonymy relation to other pivots. Moreover, about 89% and 8% of pivots are concerned with meronymy and accessibility relations, respectively.

The Prolexbase data are referenced in the META-SHARE infrastructure[29] and available[30]

[29] `http://www.meta-net.eu/meta-share`

[30] Downloadable from `http://zil.ipipan.waw.pl/Prolexbase`

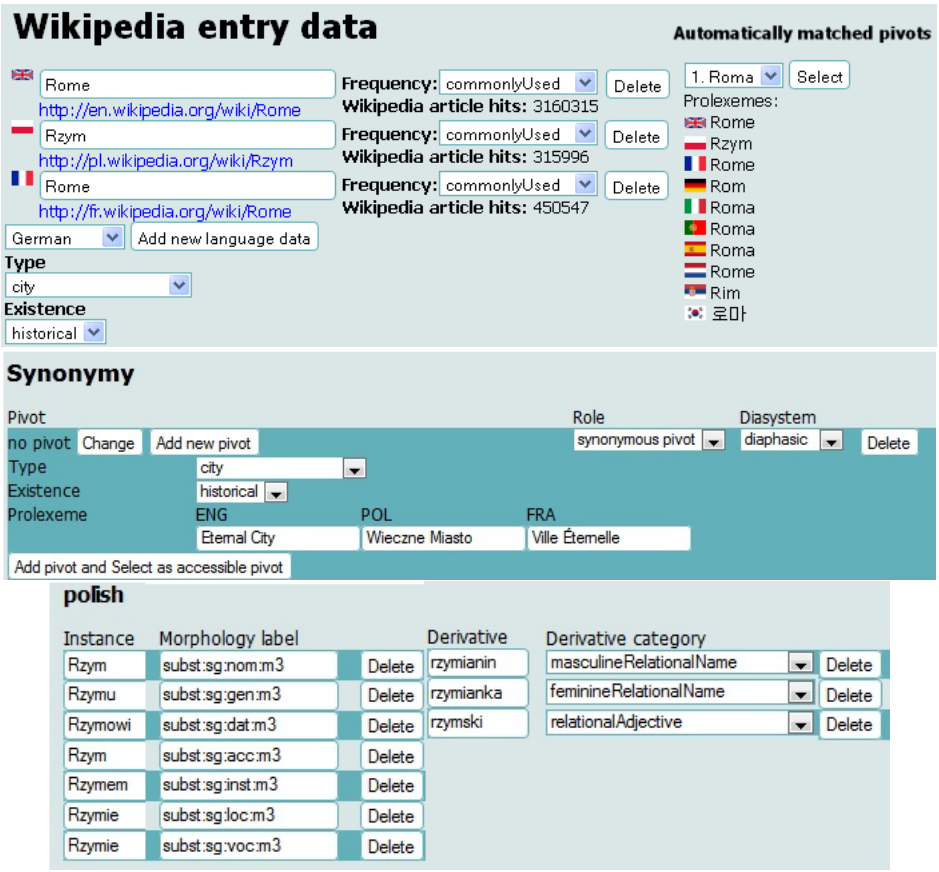


Figure 4.11: Fragments of the ProlexFeeder GUI for correcting and validating pivots, prolexemes, Wikipedia links, type, existence, frequency, synonyms, inflected forms and derivatives.

under the CC BY-SA license[31], i.e. the same as for Wikipedia and GeoNames. We are currently working on their LMF exchange format according to Bouchou & Maurel (2008).

4.6 Coreference Annotation with Nested Structures

As mentioned in Section 4.2, coreference annotation is an NLP task whose strong links with shallow parsing and named entity recognition have been stressed notably by the Automatic Content Extraction (ACE) program (Doddington et al., 2004). Since information extraction tasks in ACE focus on the target objects (entities, relations, events, etc.) rather than on the linguistic units naming them, it is crucial to be able to identify different surface *mentions* (or *markables*) referring to the same object. Hachey et al. (2013) also stress the strong, though still underestimated, interdependence of coreference resolution and named entity linking/disambiguation via Linked Open Data (LOD).

In this chapter I describe some aspects of my contribution towards the creation of the Polish Coreference Corpus and I stress those issues which relate to coreference in nested and coordinated mentions, as well as in multi-word expressions.

[31] `http://creativecommons.org/licenses/by-sa/3.0/`

Table 4.4: Current state of Prolexbase. Polish instances include inflected forms of prolexemes only.

Pivots				
All	Toponyms	Anthroponyms	Ergonyms	Pragmonyms
73,405	81.3%	16.8%	1.4%	0.4%

Relations			
All	Meronymy	Accessibility	Synonymy
72,672	92.9%	5.3%	1.8%

	Pivots in synonymy relation		Pivots in meronymy relation		Pivots in accessibility relation	
All	2,457	(3%)	65,768	(90%)	6,312	(9%)
Most	celebrity 1,325	(17%)	city 48,110	(100%)	city 2,214	(5%)
frequent	country 390	(45%)	celebrity 7,053	(88%)	region 1,696	(40%)
types	city 157	(0.3%)	region 4,052	(97%)	celebrity 1,129	(14%)

Language	Prolexemes	Aliases	Derivatives	Instances
PL	27,408	8,724	3,083	166,479
EN	19,492	14,039	94	18,575
FR	70,869	8,488	20,919	142,506

4.6.1 Polish Coreference Corpus

The ***Polish Coreference Corpus***[32] (PCC) is the first large corpus of general Polish coreference. It has a comparable size to the anaphora annotation layer of the Polish KPWr corpus (Broda et al., 2012) but its scope is significantly broader (e.g. coreference links are not restricted to named entities and markables are not limited to heads) and its development methodology includes revision of annotations.

The PCC adds a new annotation level to the National Corpus of Polish (cf. Section 4.3). It is manually annotated but it builds over a different subset of texts than the 1-million manually annotated NKJP subcorpus. Recall that the latter is composed of randomly selected paragraphs of the whole 1.5-billion corpus. We judged a one-paragraph length insufficient for a reliable coreference annotation, since coreference chains may easily span over multiple paragraphs. Thus, PCC contains 1,773 text extracts of at least 250 tokens each, selected randomly (respecting the genre balance) for the total number of about 504,000 tokens. Additionally, in order to be able to study coreference properties in full documents, 21 non-reduced texts of 1000 through 4000 tokens were also selected. With its total number of 540,000 tokens, PCC belongs to the largest coreference corpora in the international community, together with Tüba/DZ (Hinrichs et al., 2005a) for German, NAIST Text (Iida et al., 2007) for Japanese, OntoNotes 2.0 (Pradhan et al., 2007) for English, Arabic and Chinese, the Prague Dependency Treebank (Nedoluzhko et al., 2009) for Czech and ANCOR (Muzerelle et al., 2013) for French.

The automatic pre-annotation of these texts consisted in segmentation and morphosyntactic tagging (Acedański, 2010), as well as mention and coreference chain detection (Ogrodniczuk & Kopeć, 2011). Mentions and coreference chains were then manually corrected by a human annotator, within a customized version of the MMAX2 tool (Müller & Strube, 2006), and further reviewed by an expert (super-annotator). A part of the texts was double-annotated by two annotators and adjudicated by the super-annotator in order to estimate the inter-annotator agreement. In the final TEI-P5-conformant export format the level of coreference chains $L_{coreference}$

[32] `http://zil.ipipan.waw.pl/PolishCoreferenceCorpus`

builds upon the level of mentions $L_{mentions}$, which in its turn builds upon the morphosyntactic level $L_{morphosyntax}$ (see Section 4.3.2 for details of the multi-level organization of the NKJP corpus).

The resulting annotated corpus is available under the Creative Commons CC BY 3.0 license[33] and is also browsable on-line[34].

In (Ogrodniczuk et al., 2013a), we present the major aspects of the annotation scope, schema and strategies. The annotation scope covers **all nominal groups** (NGs), including pronouns since we consider the difference between an NG and a mention too controversial to be reliably decided in a general case.

As far as introducing coreference links is considered, we limit ourselves to those semantic relations which cannot be deduced directly from syntax. Firstly, unlike e.g. Haghighi & Klein (2009), nominal predicates (*Helena jest dyrektorką.* 'Helena is the director.')[35] are never included in coreference chains (although, as all other NGs, they are considered mentions). Secondly, unlike in (Linguistic-Data-Consortium, 2006) and (Nedoluzhko et al., 2009), an apposition is not seen as a sequence of coreferent mentions but as one mention only (*Dyrektorka, młoda kobieta [...]. Ona [...]* 'The director, a young woman [...]. She [...]'). Thirdly, like (Hinrichs et al., 2005b), (Nedoluzhko et al., 2009) and (Recasens & Martí, 2010), we mark split NGs as unitary mentions (*naszym yyy to znaczy nauczycieli akademickich obowiązkiem* 'our hmm it means academic teachers' duty'). Finally, like (Osenova & Simov, 2004), (Pradhan et al., 2007), (Iida et al., 2007), and (Recasens & Martí, 2010), we take special care in annotating zero subjects, pervasive in Polish.

We take two coreferential relations into account: the **identity** (leading to splitting the set of mentions into clusters, i.e. equivalence classes) and – experimentally – the **near-identity** proposed by Recasens et al. (2011). The latter happens, for instance, when two mentions refer to the same entity but the text suggests the opposite (refocusing), as in example (4.53)[36] or, conversely, when two mentions refer to different entities but the text suggests the opposite (neutralization), as in (4.54), where first a bottle and then its contents are meant. The definition of the near-identity is interesting in that it allows us to see coreference in terms of a degree of identity rather than as a binary relation. Nevertheless, as discussed in (Ogrodniczuk et al., 2013b), the frequency of near-identity links introduced by our annotators and the inter-annotator agreement are too low in our corpus to consider this relation as reliably annotated. Note also that synonymy between conceptual proper names in Prolexbase, discussed in Section 4.5.1, is very close to the idea of near-identity since it takes different points of view on the same referent into account (e.g. its previous and current name, its official name and a nickname, etc.).

(4.53) *Warszawa przedwojenna i ta z początku XXI wieku*
'Pre-war Warsaw and the one at the beginning of the 21st century'

(4.54) *Wziął wino z lodówki i wypił je.* 'He took the wine from the fridge and drank it.'

Due to the pioneering (with respect to Polish) nature of our project, all relations different from identity and near-identity are outside the annotation scope: indirect (bridging or associative) anaphora and discourse deixis (Hinrichs et al., 2005b; Poesio & Artstein, 2008; Nedoluzhko et al., 2009; Korzen & Buch-Kromann, 2011), ellipses (with the exception of zero anaphora), predicative and bound relations (Hendrickx et al., 2008), split antecedent (Hinrichs et al., 2005b), identity of sense (Iida et al., 2007), etc.

[33] `http://creativecommons.org/licenses/by/3.0/deed.en_US`

[34] `http://glass.ipipan.waw.pl:11111/index.xhtml#/core/`

[35] Henceforth, I will mark coreferent NGs with (possibly multiple) underlining, and non-coreferent NGs with dashed underlining.

[36] Mentions linked by near-identity are marked by dotted underlining.

Another element of our annotation schema, is to indicate, in every coreference chain, the **dominant expression**, i.e. the expression that carries the richest semantics or describes the referent the most precisely. The best candidates for dominant expressions are named entities, and phrases that denote a particular object in the discourse world, as in example (4.55). In many cases, pointing at the dominant expression helps the annotators sort out a large set of pronouns denoting various persons (e.g. in fragments of plays or novels). We think that it might also facilitate cross-document annotation or the creation of a semantics frame containing different descriptions of the same object. In some cases it might later help link clusters to URIs of their referents in the linked data (cf. Section 4.2) or to identifiers in another external ontology (Osenova & Simov, 2004; Pradhan et al., 2007; Iida et al., 2007).

(4.55) Cluster*: {David Beckham, rozgrywający Realu Madryt}* 'David Beckham, Real Madryt playmaker'
Dominant expression*: David Beckham*

(4.56) Cluster*: {stwierdzili, powiedzieli}* 'stated, said'
Dominant expression*: lekarze w Polsce* 'doctors in Poland'

In 62% of all cases the dominant expression was selected from among NGs contained in the cluster. 23% of them were transformed into their base forms, while 77% were taken without any changes (they already appeared in their base forms). For 38% of the clusters, the dominant expression was not present in the text but given by the annotator instead, e.g., when the cluster consisted of verb forms only (zero subjects are represented by marking their corresponding verbs as mentions), as in example (4.56).

Another novel feature of our annotation schema is the fact of pointing at the **semantic heads** of each mention. The semantic head in most nominal groups is the same element as the syntactic head but in Polish numeral groups (*pięć kobiet* '$\lfloor$five women$_{pl:gen:f}\rfloor_{pl:nom:n}$') the numeral is the syntactic head, while the noun is the semantic head. ACE annotation guidelines (Linguistic-Data-Consortium, 2006) and the associated competing systems (Hinrichs et al., 2005b) have already stressed the interest of pointing at syntactic heads of mentions. Our intuition behind annotating semantic heads is that they should help establish discourse links, notably in future automatic coreference resolvers. In particular, it seems promising to examine agreement in gender, number, synset, etc. between semantic rather than syntactic heads in potentially co-referring mentions.

4.6.2 Annotation Challenges from Nested and Coordinated Expressions

Both the ACE program (Linguistic-Data-Consortium, 2006) and some previous coreference annotation works (Osenova & Simov, 2004; Pradhan et al., 2007; Recasens & Martí, 2010) have addressed the necessity of delimiting, as mentions, not only the maximum-length NGs but **nested NGs** as well. This idea of a "semantic nesting", adopted in our coreference annotation schema, as shown in example (4.57), is very close to the one of nested named entities extensively addressed in Section 4.3. The main difference concerns person names, where given names and surnames are considered embedded NEs at the NE annotation level but not at the level of mentions, as in example (4.58), since they do not denote a different referent than the one referred to by the whole name. Rephrasing Section 4.3.3, p. 91, we can say that a forename or a surname within a full person name is (from the point of view of the coreference) an ellipsis rather than en embedded mention.

(4.57) *W 1977 roku postanowiono opracować Fiata 126p z przednim napędem [...]. Przednie koła pojazdu zawieszone były [...].*

Table 4.5: Frequencies of nested and outermost mentions in the Polish Coreference Corpus

	Nested	**Outermost**	**All**
Singleton	47,500	63,513	111,013
Non-singleton	20,155	49,039	69,194
All	67,655	112,552	180,207

'In 1977 it was decided to design Fiat 126p with a front wheel drive [...]. The front wheels of the car were suspended [...].'

(4.58) *Prof. Władysław Bartoszewski ukończył w lutym br. 80 lat. Jest [...]*
'Prof. Władysław Bartoszewski had his 80th birthday in February. [He] is [...]'

Nesting is a quantitatively and qualitatively important phenomenon in coreference annotation. As shown in Table 4.5, nested NGs account for 38% of all mentions and 29% of all non-singleton mentions, i.e. of those included in reference chains of length 2 or more. In some cases, clusters contain only nested expressions, as in the double-underlined example in (4.59).

(4.59) *W tamtych czasach często karykaturowano przedplebiscytowe poczynania rządu niemieckiego [...] rysunek wydrwił działania rządu berlińskiego [...]*
'Those days the pre-plebiscite initiatives of the German government were often caricatured [...] the drawing mocked the activity of the Berlin government [...]'

Paradoxically, a nested mention may be sometimes coreferent with the nesting one, as in example (4.60).

(4.60) *Azja bierze do niewoli m.in. Ewę Nowowiejską - siostrę Adama Nowowiejskiego, syna człowieka, który go wychował i skatował za amory do **córki**.*
'Azja emprisons notably Ewa Nowowiejska – a sister of Adam Nowowiejski, a son of the man who brought him up and tortured for romance with **his daughter**'

Finally, nesting is particularly challenging in coordinated structures, since coreference may take place with respect to elementary components of the coordination, as in example (4.61). For this reason, the annotation of mentions should probably follow the examples of coordinated names (4.9), p. 88 and (4.11), p. 89.

(4.61) *Jednoosobowe i kolegialne organy uczelni państwowej [...]. Kadencja kolegialnych organów uczelni publicznej [. . .]. Osoba pełniąca funkcję organu jednoosobowego uczelni publicznej [. . .].*
'Individual or collective bodies of a state academy [. . .]. Term of office of collective bodies of a public academy [...]. A person acting as an individual body of a public academy [. . .]'

In legal texts, coordination is pervasive, as noticed by Mazur & Dale (2007). Accumulation of coordinated structures, as in example (4.62) leads sometimes to a proliferation of potential mentions – cf. (4.63)–(4.66) – whose precise annotation is not easy to perform.

(4.62) ***Nabycie lub objęcie** przez cudzoziemca **udziałów lub akcji** w spółce handlowej z siedzibą na terytorium Rzeczypospolitej Polskiej, będącej **właścicielem lub wieczystym użytkownikiem** nieruchomości na terytorium Rzeczypospolitej Polskiej [...].*

'**Purchase or accession** by a foreigner of **an interest or a share** in a commercial company located on the territory of the Polish Republic, being **the owner or a perpetual usufructuary** of a real estate on the territory of the Polish Republic [...]'

(4.63) ***Nabycie*** *przez cudzoziemca* ***udziałów*** *w spółce handlowej z siedzibą na terytorium Rzeczypospolitej Polskiej, będącej* ***właścicielem*** *nieruchomości na terytorium Rzeczypospolitej Polskiej*
'**Purchase** by a foreigner of **an interest** in a commercial company located on the territory of the Polish Republic, being **the owner** of a real estate on the territory of the Polish Republic [...]

(4.64) ***Objęcie*** *przez cudzoziemca* ***udziałów*** *w spółce handlowej z siedzibą na terytorium Rzeczypospolitej Polskiej, będącej* ***właścicielem*** *nieruchomości na terytorium Rzeczypospolitej Polskiej*
'**Accession** by a foreigner of **an interest** in a commercial company located on the territory of the Polish Republic, being **the owner** of a real estate on the territory of the Polish Republic [...]

(4.65) ***Nabycie*** *przez cudzoziemca* ***akcji*** *w spółce handlowej z siedzibą na terytorium Rzeczypospolitej Polskiej, będącej* ***wieczystym użytkownikiem*** *nieruchomości na terytorium Rzeczypospolitej Polskiej*
'**Purchase** by a foreigner of **a share** in a commercial company located on the territory of the Polish Republic, being **a perpetual usufructuary** of a real estate on the territory of the Polish Republic [...]

(4.66) *etc.*

4.6.3 Mentions Embedded in Multi-Word Expressions

Multi-word expressions show opaque semantics, thus the NGs they include might be seen as non-referential. However, most MWEs do inherit some part of the semantics of their components, i.e. are partly semantically compositional, and might be coreferential in some stylistically marked cases, as in (4.67). Defining a clear-cut frontier between non-referential and referential NGs in these cases seems very hard. This is another reason why we consider all NGs as mentions, also those included in MWEs.

(4.67) *Nie wahał się włożyć kij w mrowisko.*
Mrowisko to, czyli cały senat uniwersytecki, pozostawało zwykle niewzruszone.
'He didn't hesitate to put a stick into an anthill (i.e. to provoke a disturbance).
This anthill, i.e. the whole university senate, usually didn't care.'

Real corpus examples of this type include many cases of compound prepositions or conjunction (4.68)–(4.72) with a variable degree of non-compositionality and fixedness. In examples (4.68)–(4.70) the units *ze strony* 'from-the-side-of = from', *na przykład* 'for example' and *w trakcie* 'in the course of' are syntactically or semantically opaque and delimiting their component nouns as mentions or parts of mentions seems incorrect. Examples (4.71)– (4.72) are more controversial. It would be interesting to perform quantitative studies showing how often an anaphoric reference to nouns in such expressions appears in the corpus. We think that such a test might be a measure of semantic or syntactic opaqueness in multi-word expressions.

(4.68) Present annotation: *ze strony społeczeństwa*
'lit. from-the-side-of the society = from the society'
Alternative annotation: *ze strony społeczeństwa*
'lit. from-the-side-of the society'

(4.69) Present annotation: *na przykład to przepytywanie za pomocą wiarografu*
'for example this questioning with a polygraph'
Alternative annotation: *na przykład to przepytywanie za pomocą wiarografu*
'for example this questioning with a polygraph'

(4.70) Present annotation: *w trakcie prac legislacyjnych*
'in the course of legislative work
Alternative annotation: *w trakcie prac legislacyjnych*
'in the course of legislative work'

(4.71) Present annotation: *W przypadku wygaśnięcia mandatu*
'In case of mandate expiry'
Alternative annotation: *W przypadku wygaśnięcia mandatu*
'In case of mandate expiry'

(4.72) Present annotation: *brak sukcesów w zakresie pozyskiwania środków Unii Europejskiej*
'failure in the scope of obtaining means from the European Union = in obtaining [...]'
Alternative annotation: *brak sukcesów w zakresie pozyskiwania środków Unii Europejskiej*
'failure in the scope of obtaining means from the European Union'

4.7 Contributions

Our contributions to the fields of (named) entity modeling and processing are manifold. The NKJP corpus is one of the relatively few application-independent large reference corpora aiming at a rigorous modeling of language phenomena. It fits the high annotation standards based on the principle of a multi-level annotation, double-annotation and adjudication. Due to its TEI P5 stand-off representation format it respects the modern data representation and interoperability principles.

With its NE annotation level, whose contents is summarized in Table 4.6, it is probably the largest and the most comprehensive attempt towards annotating NEs in Polish, and one of the largest in Slavic languages. It was developed according to a rich annotation schema, which accounts for nested, coordinated, overlapping and discontinuous NEs, rarely annotated as such in other projects. To the best of our knowledge, no other NE-annotated corpora represent relational adjectives and inhabitant names as instances of their related NEs.[37] With regard to this aspect our approach might be seen as novel. We also pay special attention to correct lemmatization of NEs. We have found no explicit references to this problem except in the original SProUT grammar by Piskorski (2005) and in the dedicated study by Piskorski et al. (2007).

The annotation tools described in this section belong to important achievements in NE annotation and recognition. The modified SProUT grammar now explicitly takes NE nesting into account and has received an extended gazetteer used in other Polish NER tools, e.g. (Marcińczuk & Kocoń, 2013). Our adaptation of TrEd, initially developed for a dependency treebank, proved its utility for constituency treebanks, its usability and easy customization. Our NE annotation

[37]The Quaero annotation guidelines (Rosset et al., 2011) do recommend to annotate inhabitant names with a dedicated *demonym* type, these are however not associated with their derivational bases. The corpus itself is not publicly available.

Table 4.6: Named entities annotated in the National Corpus of Polish. *SC* and *WC* stand for the 1-million word manually annotated subcorpus, and for the whole 1.5-billion word corpus, respectively.

	persName	**orgName**	**geogName**	**placeName**	**date**	**time**	**relAdj**	**persDeriv**	**Overall**
SC	47,286	11,380	3,893	10,733	4,514	562	7,147	1,785	87,300
WC	94,991,096	22,593,467	7,736,442	28,666,309	9,778,340	2,252,954	9,375,486	1,834,876	166,018,608

file management tools were effectively reused in coreference annotation in NKJP. Finally, Nerf is one of the first tools in the international community dedicated to recursively embedded NEs, and probably the first comprehensive machine-learning method for Polish NER.

Our work on Prolexbase contributes to the domain of ontological and lexical resources of named entities. We have adopted a rich model in which: (i) semantic aspects are represented by conceptual proper names interconnected by a rich set of relations, (ii) lexical aspects are covered by a fine-grained morphological model including prolexemes, aliases, derivatives and instances. Most previous data in Prolexbase were French. We have complemented them with about 18,000 new pivots and 19,000 relations, as well as 23,000 Polish, 19,000 English and 15,000 French prolexemes and many aliases. This over 127% increase in the amount of data allowed us to perform the first large-scale validation of the Prolexbase model in a multilingual context.

Before ProlexFeeder was created, Prolexbase population had been performed mostly manually (Tran et al., 2005). Uniqueness of pivots was based on a prolexeme match alone. Lists of entries and attributes were crafted in spreadsheet files. Data were manually looked up in traditional dictionaries, lists and Internet sources. Inflected forms were generated via external tools. The complexity of the model hardly allowed the users to work in this way on more than one language or more than one type at a time. ProlexFeeder largely facilitates the lexicographer's work in that most data are automatically fetched, pivot uniqueness relies on more elaborate multilingual checks, entry validation is supported by automatic Prolexbase lookup, and inflected forms are automatically generated.

Due to the insertion of Wikipedia URLs in most Prolexbase entries, we have also paved the way for connecting Prolexbase with the Linked Data (Mendes et al., 2012; Hoffart et al., 2013). In this way future NLP applications will be able to benefit from the huge amounts of multilingual interlinked data on the one hand, and from NE-specific relations and morphological data necessary for NE processing in texts on the other hand.

The annotation levels of mentions and coreference in the NKJP extend the one of NEs to more largely understood entities. They offer the first large corpus of general coreference in Polish and one of the first in Slavic languages. It is probably the first effort in the international community to annotate near-identity relations (Recasens et al., 2011) on a large scale. Other novel aspects of the annotation schema include dominant expressions and semantic heads that may prove useful in future automatic in-document and cross-document coreference resolution, as well as in entity linking.

In resources and tools described in this section, particular impact has been made on the proper treatment of multi-word units. We show the quantitative and qualitative importance of these phenomena. By an extensive coverage of nesting in NEs and mentions, we offer an advanced approach to entity modeling and processing, more fine-grained than both the bag-of-words and the words-with-spaces viewpoint.

4.8 Perspectives

The NKJP NE annotation deserves completions and extensions. New types and subtypes, such as product names, quantities and measures should be taken into account. TimeML-inspired annotation of temporal expressions might complete the current annotation schema. Difficult cases of nested, elliptical and metonymic NEs (cf. Section 4.3.3) should be analyzed more deeply and benefit from unified annotation rules. Metonymy could also be annotated explicitly, as in (Desmet & Hoste, 2010), in order to help its automatic resolution. Allowing for competing annotations in case of unsolvable ambiguities (cf. Sections 4.3.3–4.3.3) could enhance the precision of future NE-dedicated applications.

The current corpus can already now be used in linguistic studies. One of interesting aspects is to check how well the segments identified at the level of NEs correspond to those at the level of syntactic groups (Waszczuk et al., 2013) on the one hand, and to corefence mentions on the other hand. This correspondence has a big influence on morphological and syntactic analysis of NEs, since the level of syntactic groups provides data such as syntactic and semantic headwords, as well as inflectional features of each group.

As far as NER tools are concerned, a new model for Nerf is currently being developed, in which nesting of NEs is no longer reduced to sequential labeling but annotation trees are straightforwardly modeled. This allows us, in particular, to take relationships between distant sentence tokens into account. External NE lexicons, such as the SProUT gazetteer and Prolexbase, can now also be used as sources of observations. The first results show that these enhancements contribute to a substantial increase in NER results.

Prolexbase is an open-ended project. We should be able to regularly update the frequency codes both with Wikipedia hits on larger time spans, and with corpus frequency methods. We need to complete the existing entries with missing data (e.g. classifying contexts), and to design an (easier to maintain) intentional description of the morphological data. Pivot matching could take approximate string matching techniques and text mining from Wikipedia articles into account.

New development is also needed for the Prolexbase model itself as far as multi-word and nested structures are concerned. Prolexemes are currently represented as sets of unstructured strings. In order for Prolexbase data to be applicable to tasks where nesting is an issue, prolexemes should ideally be modeled (intentionally or extensionally) as trees of two kinds. Fully lexicalized, possibly theory-independent, syntax trees would represent their grammatical structure and could be reproduced over their occurrences in the corpus in the process of deep parsing. "Semantic" trees would be built similarly to the NE annotation trees in the NKJP corpus (cf. Section 4.3) and could be straightforwardly reused in NE annotation and recognition.

Chapter 5

Finite-State Methods for Word and Tree Languages

Formal languages of words (strings) and trees are basic objects of interest in computer science. They have, notably, often been treated as approximations of natural languages. They also offer representation formalisms for encoding semi-structured data, such as XML, heavily used in linguistic data encoding and interchange standards. This chapter is dedicated to my contributions to finite-state algorithmics, with respect to both word and tree languages.

5.1 Formal Methods for the Representation and Approximation of Words and Trees – State of the Art

5.1.1 Finite-State Techniques for NLP in a Nutshell

Finite-state methods of a varying degree of expressiveness have been heavily applied to the representation and approximation of linguistic data. In the simplest case, subsets of a natural language are seen as regular languages, thus describable by regular expressions. For instance, the lexicon of a natural language can be represented as a finite set of words (i.e. a regular expression). Also infinite sets of language structures can be represented by part-of-speech patterns, e.g. Justeson & Katz (1995) extract well formed noun phrases in English with the following regular expression: ((A | N)$^+$| (A | N)* (N P) (A | N)*) N[1]. The possible patterns matching this regular expression include AN, NN, AAN, ANN, NAN, NNN, NPN, ... Regular expressions are equivalent to **finite-state automata**, i.e. for every regular expression there is a unique minimum deterministic finite-state automaton defining the same language, and vice versa.

Finite-state transducers are more complex tools than automata because of their two-way functioning based on an input alphabet and an output alphabet. They are applied to many areas of natural language processing: phonology (Kaplan & Kay, 1994; Laporte, 1997), morphology (Koskenniemi, 1983; Beesley & Karttunen, 2003), part-of-speech tagging (Roche & Schabes, 1997), and parsing (Roche, 1997).

In the field of information extraction, a **transducer cascade** is an efficient technique. A cascade is a set of transducers that are applied to a text one after another. Each transducer parses the text and performs some transformations on it. The resulting transformed text becomes the input for the following transducer. Three of the systems using this technique are Cass (Abney, 1996), FASTUS (Hobbs et al., 1997) and CasSys (Friburger & Maurel, 2001).

[1]The symbols A, N and P stand for adjective, noun and preposition respectively, alignment of symbols stands for concatenation, "|" stands for union, "+" for one or more occurrences of a symbol, and "*" for zero or more occurrences.

One of the reasons why finite-state automata and transducers are widely used in NLP in their classical and extended (Kornai, 1999) versions is their time and space efficiency obtained by determinisation (sequentialisation) and minimisation (Watson, 1995; Daciuk et al., 2000; Mohri, 1994; Gaál, 2001). These two properties can be characterized as follows. For each non-deterministic finite-state automaton there exists a minimal deterministic finite-state automaton recognizing the same language (Hopcroft, 1971; Hopcroft & Ullman, 1979). In the general case, due to the determinization process, the number of states of the resulting automaton may theoretically increase exponentially, but for some subclasses of finite-state automata the worst-case space complexity of determinization is far lower (Melishar & Skryja, 2001). The problem of minimisation and determinization of finite-state transducers is more complex than that of finite-state automata. A transducer may be interpreted as a simple automaton whose alphabet contains couples of input and output symbols. Then, the minimisation algorithms designed for automata may also be applied to transducers. However, a word lookup in such a transducer may not be deterministic. A transducer which is deterministic with respect to its input alphabet is called a sequential transducer. Not all transducers can be sequentialized, but their sequentiability is decidable (Gaál, 2001).

The time and space complexity of finite-state tools can further be enhanced by compression techniques if internal implementation details are taken into account. For instance, compression techniques of large multilingual lexicons proposed by Daciuk & Weiss (2011) reduce the space requirements to only 1.3 up to 3.9 bits per entry.

5.1.2 String-to-String and String-to-Language Correction

Several NLP applications such as spelling correction, information retrieval with noisy data, morphological analysis of old language, etc., can be modeled as instances of the theoretical problem of *approximate string matching* (Hall & Dowling, 1980). Namely, typing or recognition errors or variants can be interpreted as resulting from one or more elementary *editing operations* on letters: insertions, deletions, replacements and inversions of adjacent letters (Damerau, 1964)[2] The distance between two strings is the minimum cost of all sequences of editing operations that transform one string into another. Different sequences of editing operations may be allowed and different cost functions may be assigned to these editing operations. With the distance measure called *edit distance* proposed in Wagner & Fisher (1974) and Lowrance & Wagner (1975), editing operations may be assigned arbitrary non-negative costs, and they may act on arbitrary positions in the string in arbitrary order (e.g. *ca* can be obtained from *abc* by two operations: deletion of *b*, inversion of *a* and *c*). However, an efficient algorithm for edit distance calculation exists only if $W_I + W_D \leq 2W_S$, where W_S, W_I, W_D are costs assigned to inversion, insertion and deletion operations, respectively.

In (Du & Chang, 1992) this distance measure is modified and renamed to *error distance* by assigning cost 1 to each editing operation and by admitting that errors occur in linear order from left to right so that a later operation may not cancel the effect of an earlier operation. Thus, inversions occur only between letters that are adjacent in the original word and remain adjacent in the erroneous word (e.g. the error distance between *abc* and *ca* is 3). Due to the equal cost of each editing operation, the error distance becomes a *metric*, i.e. a function satisfying four properties: non-negative values, reflexivity, symmetry, and triangular inequality.

The computational solution for the (editing or error) string-to-string distance calculation, belonging to the class of *dynamic programming* algorithms, is based on a matrix *H[0:n,0:m]*, where n and m are the lengths of the two strings to be compared, and *H[i,j]* contains the distance

[2] A reduced set of operations, containing insertions, deletions and replacements, was proposed independently by Levenshtein (1966) in the context of correcting binary words. The word-to-word distance based on these three operations with cost 1 each is often called the Levenshtein distance.

between the prefixes of lengths i and j of the two strings. The calculation is particularly efficient for the error distance matrix, since the value of the element *H[i+1,j+1]* depends only on the values of the elements *H [i-1,j-1]*, *H[i,j]*, *H[i+1,j]*, and *H[i,j+1]*.

An extension of the string-to-string correction problem, motivated notably by spelling correction, is the string-to-language correction. For a given input word w and a given language L, this problem consists in finding those words which belong to L and that are similar to w, the similarity being a reverse function to edit or error distance. Since there is no theoretical distance limit between an erroneous word and its corrections, a trade-off is necessary between three factors: the search time efficiency, the length of the resulting correction candidate list (the user may be unwilling to consult a long list), and the chance that the intended word be on that list. Thus, two of the possible string-to-language correction problem definitions are:

- Finding all valid words which are no more distant from the input word than a given threshold.
- Finding the nearest-neighbors, i.e. the valid words with the minimal distance from the input word (the minimal distance possibly being no bigger than a given threshold).

It is further interesting to retrieve the minimal-cost edit sequences which allow to transform the incorrect word into any of its corrections (or vice-versa).

Boytsov (2011) shows that the string-to-language correction (that he calls approximate dictionary searching) has received attention from various scientific communities, which frequently reduced the scope of this problem and built comparable algorithms, sometimes in parallel, for their specific application contexts. He presents an extensive state-of-the-art survey of about 30 different algorithms. He classifies them into a taxonomy of several dozens of classes, two major ones being: (i) direct methods, that subdivide into methods based on: prefix trees, neighborhood generation or metric-space pivoting, and (ii) sequence-based filtering methods including: pattern partitioning and vector-space frequency-distance methods. He also performs comparative experimental tests, within a common implementation framework, on natural language and DNA data, with the distance threshold of 1, 2, and 3. He shows that due to hardware and software advances an approximate dictionary searching query can be answered in 2 milliseconds on average and is up to four orders of magnitude faster than sequential searching. This search time grows, however, exponentially with the distance threshold.

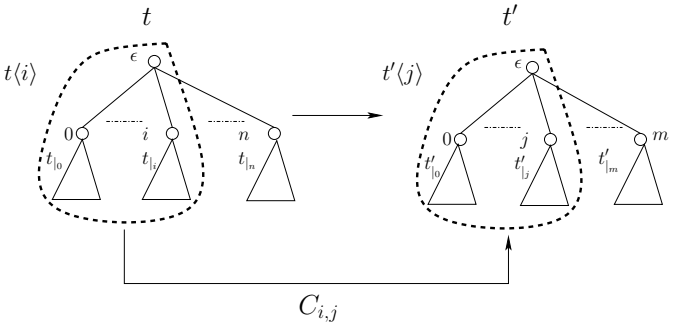
In the following sections I refer to two other string-to-language correction algorithms which could be classified, according to Boytsov's taxonomy, as direct methods based on a prefix tree implemented as a string trie. One of them, by Oflazer (1996) performs the calculation of the error distance matrix during a depth-first search traversal of the finite-state automaton representation of the lexicon. Following a new transition triggers the calculation of a column in the edit distance matrix, labeled with the same character as the transition. Backtracking from a transition leads to deleting the column calculated for this transition. In this way, when a word is searched for in the lexicon, a part of the matrix is calculated only once for all lexicon words that have the same common prefix. The other algorithm, proposed by me (Savary, 2001b), admits a similar approach but retains only the most similar corrections (nearest neighbors), reducing dynamically the search space in the lexicon, and follows the longest correct prefix first, thus allowing to often reach the first correction as soon as possible.

5.1.3 Tree-to-Tree and Tree-to-Language Correction

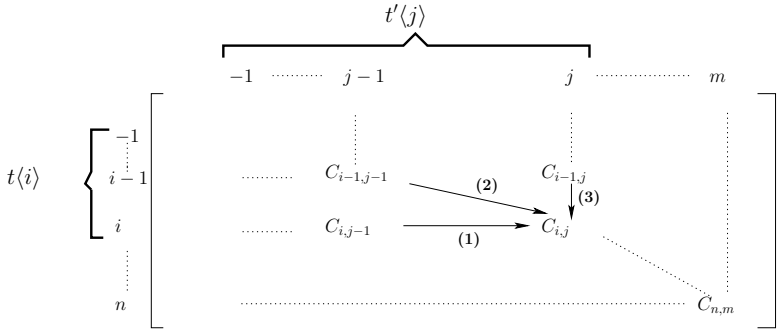
A string of symbols may be viewed as a trivial case of a tree whose depth is 1 and whose leaves are the elements of the string. Thus, the formalization of the string-to-string correction problem naturally inspired research on the tree-to-tree correction problem (David Barnard and

Gwen Clarke and Nicholas Duncan, 1995). Note that the diversity of the possible choices of elementary editing operations is bigger in case of a tree than of a word since one can consider changes not only on the siblings' level but also on some ancestor's level. The most appropriate choice depends on the intuitive notion of tree proximity for the particular application.

Among the tree-to-tree correction algorithms, the one by Selkow (1977) straightforwardly extends the string edit distance definition by Wagner & Fisher (1974) to unranked labeled trees. Three elementary editing operations are considered: (i) changing a node label, (ii) deleting a subtree, (iii) inserting a subtree (the two latter operations can be decomposed into sequences of node deletions and insertions, respectively). A cost is assigned to each of these operations and the problem is to find the minimal cost of all operation sequences that transform a tree t into a tree t'. The edit distance between t and t' is equal to this minimal cost.



(a)



(b)

Figure 5.1: (a) Two partial trees $t\langle i\rangle$ and $t'\langle j\rangle$. (b) Tree edit distance matrix: computation of $H[i,j] = C_{i,j}$.

The computation of the edit distance is based on a matrix H where each cell $H[i,j]$ contains the edit distance between two partial trees $t\langle i\rangle$ and $t'\langle j\rangle$. A partial tree $t\langle i\rangle$ of a tree t consists of the root of t and its subtrees $t_{|_0}, \ldots, t_{|_{i-1}}$ – see Figure 5.1(a). The matrix H is computed column by column, from left to right and top down. Each element $H[i,j]$ is deduced from its three neighbors $H[i-1,j-1]$, $H[i-1,j]$ and $H[i,j-1]$, as shown in Figure 5.1(b). It contains the minimum value among *(1)* its left-hand neighbor's value plus the minimum cost of inserting the subtree $t'_{|_j}$ (Figure 5.1(b), edge (1)), *(2)* its upper-left-hand neighbor's value plus the minimum cost of transforming the subtree $t_{|_i}$ into $t'_{|_j}$ (Figure 5.1 (b), edge (2)), and *(3)* its upper neighbor's value plus the minimum cost of deleting the subtree $t_{|_i}$ (Figure 5.1 (b), edge (3)). Note that computing the edit distance between t and t' implies computing edit distances between subtrees of t and subtrees of t'. The time complexity of Selkow's algorithm is $O(\Sigma_{i=0}^{min(d_t,d_{t'})} h_i h'_i)$, where

d_t and $d_{t'}$ are the depths of t and t', and h_i and h'_i are the numbers of nodes at height i in t and t', respectively.

In (Tai, 1979) a different set of elementary editing operations on trees is considered: (i) relabeling a node, (ii) deleting and (iii) inserting a possibly internal (non leaf) node. Node mappings defined on compared trees t and t' show how a sequence of edit operations transforms t into t' regardless of their order. The edit distance between t and t' is equal to the minimum cost of such mappings. In (Zhang & Shasha, 1989) the same basic operations hold but the tree-to-tree distance problem is transformed into finding the distance between two ordered subforests of the initial trees.

By analogy to strings, a tree may be compared not only to another single tree but to a tree language as well. This problem has been extensively studied by the XML community and has different definitions, as shown by Tekli et al. (2011) and Amavi et al. (2013). Some authors are mainly interested in measuring the *distance between an XML document and a schema* (represented by a DTD, an XML schema or another related formalism) (Bertino et al., 2004; Xing et al., 2006; Staworko & Chomicki, 2006; Tekli et al., 2007; Bertino et al., 2008; Staworko et al., 2008; Thomo et al., 2008). Others also search for *one or all minimal cost corrections*, i.e. trees that are valid with respect to the schema and whose distance from the initial tree is minimal (Boobna & de Rougemont, 2004). Finally, the most complete approach is not only to find the possible corrections but also to *restore the edit sequences* that lead from these corrections to the input tree (Suzuki, 2007; Svoboda, 2010; Svoboda & Mlýnková, 2011). A particular instance of the tree-to-language correction problem appears in the context of XML document and/or schema evolution. When updated documents become invalid, two possible solutions are to: (i) update the schema in such a way that both the previously valid and the newly invalidated documents become valid (Bouchou et al., 2004; Shoaran & Thomo, 2011), (ii) correct the invalidated document but preserve the most recent updates. One of our contributions (Bouchou et al., 2006b,a) described in Section 5.3 belongs to this class of algorithms.

5.2 Correcting Words and Trees

Our early contributions to finite-state algorithmics for NLP concern the problem of string-to-language correction and its application to spelling correction (cf. Section 5.1.2). In (Savary, 2001b) we proposed a method for an error-tolerant lookup in a finite-state lexicon admitting four elementary edit operations introduced by Damerau (1964) (insertions, deletions, replacements and inversions). The main idea of the algorithm is that approximate FSA dictionary search is a fourfold modification of the basic exact FSA search. Namely, given the current string position i, occupied by the character w_i and the current FSA state s, exact search consists in advancing both to a new state s' and to a new position $i+1$ by following a transition labeled with w_i. In approximate search this basic step can be transformed into four: (i) we quit the state s by a transition but we stay at position i (insertion), (ii) we stay in the state s but we pass to the next position $i+1$ (deletion), (iii) we quit the state s and we pass to position $i+1$ but we follow a transition labeled by a different character than w_i (replacement), (iv) we follow two transitions labeled with w_{i+1} and w_i, respectively, and we pass to position $i+2$ (inversion).[3]. Our algorithm is conceptually very close to the one by Oflazer (1996). One of its advantages is to match the longest correct prefix first, which may allow to quickly find the first (and often the most appropriate) solution. Namely, in FSAs encoding natural language lexicons the fan-out is very big for the states close to the initial state and it decreases for those close to the final states. Thus, the depth-first search exploration can take time before finding the first solutions if many

[3]Similar observations underly the idea of a restoration graph introduced by Staworko & Chomicki (2006) in a recursive XML tree correction with respect to a DTD.

backtracking steps to the initial state have to be performed. Moreover, statistical studies have shown that spelling errors are rarely committed at the very beginning of a word. Thus, following the longest correct prefix first often leads to directly achieving the position to be corrected in order to obtain the intended word.

The FSA-based string-to-language correction can be extended to XML document correction with respect to a DTD because structural constraints imposed on document nodes are expressed in a DTD as regular expressions. In other words, the valid words formed by all possible sequences of children of a node form a regular language. This is no longer a natural language, thus the benefits of our algorithm in (Savary, 2001b) with respect to natural language vocabularies do not apply. We rely therefore on Oflazer's algorithm whose advantage is to be more concise and elegant. We restrict, however, the set of elementary editing operations by eliminating inversions, which are hardly justified within the context of XML documents. We combine Oflazer's approach with the tree-to-tree correction algorithm by Selkow (1977), cf. Section 5.1.3. The resulting tree-to-language correction approach was first proposed in an incremental framework, discussed in more details in Section 5.3. Later on, a fundamental, application-independent version of the algorithm was developed, implemented, tested and compared with the state of the art in (Amavi et al., 2013). We illustrate the principles of our algorithm by an example drawn from this paper.

5.2.1 An Example

Let $\Sigma = \{root, a, b, c, d\}$ be a set of tags, and let t be the XML tree in Fig. 5.2. The positions of nodes in t are represented by sequences of integers such that: (i) the children of a node are numbered from left to right by consecutive non-negative integers 0, 1, etc., (ii) the tree's root is at position ϵ, (iii) if node n is at position p, the position of the $(i+1)$-th child of n is given by the concatenation of p and i. For instance, in Fig. 5.2, the node at position 1.0 (labeled with c) is the first child of the node at 1 (labeled b), which on its turn is the second child of the root at ϵ. A tree is seen, formally, as a mapping from positions to labels. Thus, the tree in Fig. 5.2 can be described as the set $\{(\epsilon, root), (0, a), (0.0, c), (0.1, d), (1, b), \ldots\}$.

Let S be the structure description in Fig. 5.3 representing a DTD. Note in particular the finite-state automaton associated with the root element and corresponding to the regular expression $b^*|ab^*c$. The tree t is not valid with respect to S because the word which is formed by the tags of the children of the root node, i.e. abb, does not belong to $L(b^*|ab^*c)$[4].

We wish to correct t with respect to S, i.e. to compute the set of valid trees $\{t'_1, \cdots, t'_n\}$ whose distance from t is no higher than a given threshold th. Let $th = 2$. We construct a tree-to-language edit distance matrix M, inspired from (Selkow, 1977), which contains the sets of operation sequences (of cost no higher than th each) needed to transform partial trees of t into partial trees of t'_i, as shown in Fig. 5.4. Let $M[i][j]$ or (i, j) designate the cell of the matrix M at line i and at column j. Cell $(0, 0)$ contains the operation sequence needed to transform the root node of t to the root node of the trees in $L(S)$. Here, t has the same root as the one specified by the schema S. To keep this root intact cell $(0, 0)$ contains an empty operation sequence denoted by $nos_\emptyset$. Then for computing the other cells of M we use the cells which are already computed. For instance, going from cell $(0, 0)$ to cell $(1, 0)$ we consider deleting the subtree of t rooted at position 0, which has cost 3. Thus, the threshold is exceeded and cell $(1, 0)$ becomes empty as well as all other cells below.

The computation of the matrix M is done column by column. According to (Oflazer, 1996), a new column is added after following a transition in the FSA_{root} automaton associated with the root element of S. For instance for the column $j = 1$ we may use the transition (q_0, b, q_1) and this column will be referred to by the tag b. This means that the subtrees at position 0

[4] $L(E)$ stands for the language described by the regular expression E.

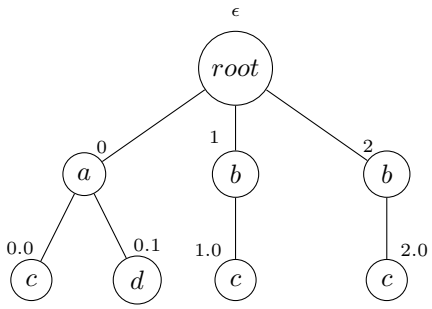


Figure 5.2: An XML tree.

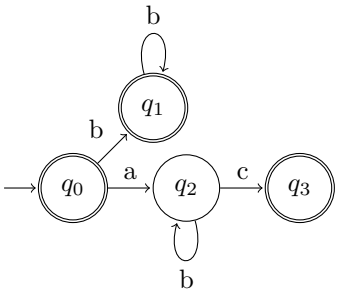
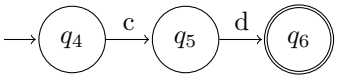
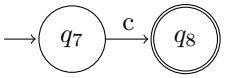
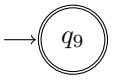
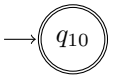
Tag	Regular Expression	Finite State Automaton(FSA)
root	$b^*\|ab^*c$	
a	cd	
b	c	
c	ϵ	
d	ϵ	

Figure 5.3: An example of a structure description.

	M	0 root	1 b	2 b	3 b	**4** ***b***
0	root	$\{nos_\emptyset\}$	$\{\langle(add, 0, b), (add, 0.0, c)\rangle\}$	$\emptyset$	$\emptyset$	$\emptyset$
1	a	$\emptyset$	$\{os_1{=}\langle(relabel, 0, b), (delete, 0.1, /)\rangle\}$	$\emptyset$	$\emptyset$	$\emptyset$
2	b	$\emptyset$	$\emptyset$	$\{os_1\}$	$\emptyset$	$\emptyset$
3	b	$\emptyset$	$\emptyset$	$\emptyset$	$\{os_1\}$	$\emptyset$

Figure 5.4: Content of the matrix M for the word prefix *bbbb*

in the correct tree that we are trying to construct will have a root labeled b. The tags for all columns ($j > 0$) in M form a word u. Fig. 5.4 shows the contents of the matrix M for the word $u = bbbb$.

In order to calculate a new cell (i, j), we concatenate each sequence taken from the left and top neighboring cells $(i, j-1)$, $(i-1, j-1)$ and $(i-1, j)$ with one of the three following, possibly complex, operations (provided that the threshold th is not exceeded):

(i) **Inserting subtrees** (denoted by $\rightarrow$): coming from the left-hand cell we concatenate its operation sequences with an insertion of a subtree in a result tree t'_i. For instance, in order to calculate cell $(1, 1)$ we consider cell $(1, 0)$, which is empty and cannot yield any operation sequence.

(ii) **Correcting a subtree** (denoted by $\searrow$): coming from the upper-left-hand cell we concatenate its operation sequences with a correction of a subtree in t into a valid subtree of t'_i. This correction is performed by a recursive call so another tree-to-language edit distance matrix is computed. For instance, coming from cell $(0, 0)$ to $(1, 1)$ we concatenate

the empty sequence $nos_\emptyset$ with the operation sequence os_1=$\langle(relabel, 0, b), (delete, 0.1, /)\rangle$ which results from correcting the subtree $\{(\epsilon, a), (0, c), (1, d)\}$ at position 0 in t to a valid subtree with root b. The subtree that we obtain is $\{(\epsilon, b), (0, c)\}$. The cost of os_1 is $2 \leq th$ so we can add the resulting operation sequence set which contains os_1 itself to cell $(1, 1)$. The matrix which is computed for correcting the subtree $\{(\epsilon, a), (0, c), (1, d)\}$ into $\{(\epsilon, b), (0, c)\}$ is shown in Fig. 5.5. Note that os_1 stems from the sequence obtained here in cell $(2, 1)$, prefixed with position 0.

(iii) **Deleting a subtree** (denoted by $\downarrow$): coming from the upper cell we concatenate its operation sequences with a deletion of a subtree in t. For instance coming from cell $(0, 1)$ to $(1, 1)$ we concatenate the operation sequence $\langle(add, 0, b), (add, 0.0, c)\rangle$ having cost 2 with the operation sequence $os_2 = \{\langle(delete, 0.1, /), (delete, 0.0, /), (delete, 0, /)\rangle\}$ allowing us to delete the subtree at position 0 in t. However, the cost of this deletion is 3 and its concatenation with $(0, 1)$ yields a sequence with cost 5, which exceeds the threshold 2. Thus we don't have, for the cell $(1, 1)$, any operation sequence coming from $(0, 1)$.

M'	0 b	1 c
0 a	$\{\langle(relabel, \epsilon, b)\rangle\}$	$\{\langle(relabel, \epsilon, b),$ $(insert, 0, c))\rangle\}$
1 c	$\{\langle(relabel, \epsilon, b),$ $(delete, 0, /))\rangle\}$	$\{\langle(relabel, \epsilon, b)\rangle\}$
2 d	$\emptyset$	$\{\langle(relabel, \epsilon, b),$ $(delete, 1, /)\rangle\}$

Figure 5.5: New matrix computed by a recursive call

For the other cells of the matrix in Fig. 5.4, we use the transition (q_1, b, q_1). If the word formed by the column tags is in $L(FSA_{root})$ (i.e. we reach a final state), the bottom cell of the current column contains possible solutions. Since $bbb \in L(FSA_{root})$, cell $(3, 3)$ contains an operation sequence capable of transforming t into a valid tree $t'_i \in L(S)$. When we apply this operation sequence, i.e. os_1=$\langle(relabel, 0, b), (delete, 0.1, /)\rangle$, on the tree t, we obtain the tree t'_1 in Fig. 5.6.

All the cells of the last column ($j = 4$) of the matrix in Fig. 5.4 are empty, which means that we can not have an operation sequence with a cost less than $th = 2$ for a word with the prefix *bbbb*. In this situation we backtrack by deleting the last column and try another transition. In this example we will delete all columns except the first one. After backtracking to q_0 it is possible to follow the transition (q_0, a, q_2) for computing the second column of the matrix in Fig. 5.7. The other columns of this matrix are computed by following the transition (q_2, b, q_2) until we reach another empty column. Note that the node operation sequence contained in cell $(3, 4)$ in Fig. 5.7 may be expressed as a single higher level operation on subtrees, namely as inserting a subtree $\{(\epsilon, b), (0, c)\}$ at position 3.

We backtrack again and use the transition (q_2, c, q_3). The cells of this current column (for $j = 4$) are shown in Fig. 5.8.

The word *abbc* formed by the tags of the current columns is in $L(FSA_{root})$ and the bottom cell of the current column contains a sequence with cost no higher than the threshold. Therefore we obtain a new correction t'_2 depicted in Fig. 5.6. In the state q_3 we don't have any outgoing

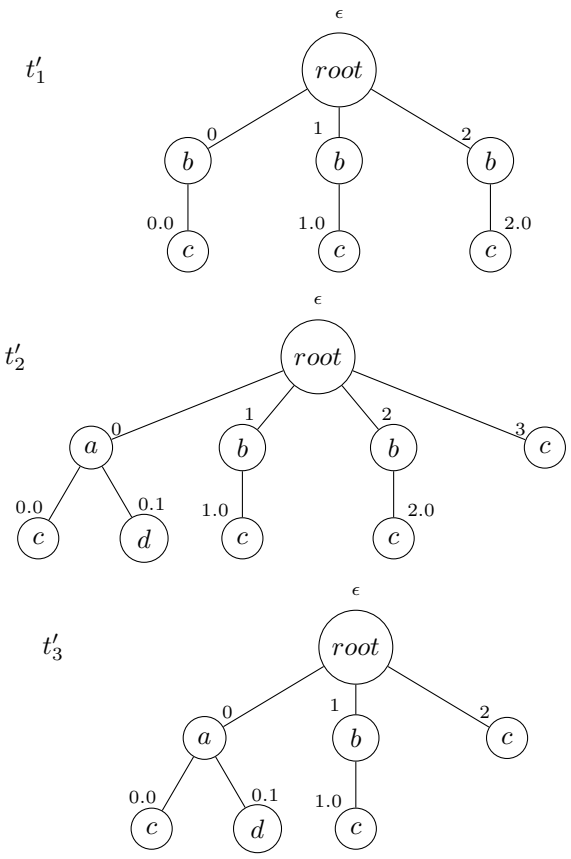


Figure 5.6: Three possible corrections t'_1, t'_2 and t'_3 for the tree t in Fig. 5.2

.

M	0 root	1 a	2 b	3 b	4 b	**5** ***b***
0 root	$\{nos_\emptyset\}$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\boldsymbol{\emptyset}$
1 a	$\emptyset$	$\{nos_\emptyset\}$	$\{\langle(add, 1, b),$ $(add, 1.0, c)\rangle\}$	$\emptyset$	$\emptyset$	$\boldsymbol{\emptyset}$
2 b	$\emptyset$	$\{\langle(delete, 1.0, /),$ $(delete, 1, /)\rangle\}$	$\{nos_\emptyset\}$	$\{\langle(add, 2, b),$ $(add, 2.0, c)\rangle\}$	$\emptyset$	$\boldsymbol{\emptyset}$
3 b	$\emptyset$	$\emptyset$	$\{\langle(delete, 2.0, /),$ $(delete, 2, /)\rangle\}$	$\{nos_\emptyset\}$	$\{\langle(add, 3, b),$ $(add, 3.0, c)\rangle\}$	$\boldsymbol{\emptyset}$

Figure 5.7: Content of the matrix M for $u = abbbb$ (after backtracking from state q_1 in FSA_{root})

transition so we backtrack, then we try the word *abc*. Fig. 5.9 shows the corresponding matrix, with a sequence in its bottom-right cell whose cost is not higher than th. This sequence is obtained with a new matrix computed by a recursive call in order to correct the subtree $\{(\epsilon, b), (0, c)\}$ at position 2 into $\{(\epsilon, c)\}$. The resulting correction t'_3 is depicted in Fig. 5.6. After that, we will have no more possibilities to find other corrections than t'_1, t'_2 and t'_3 within the threshold $th = 2$.

M	0 root	1 a	2 b	3 b	**4** ***c***
0 root	$\{nos_\emptyset\}$	$\emptyset$	$\emptyset$	$\emptyset$	$\boldsymbol{\emptyset}$
1 a	$\emptyset$	$\{nos_\emptyset\}$	$\{\langle(add, 1, b),$ $(add, 1.0, c)\rangle\}$	$\emptyset$	$\boldsymbol{\emptyset}$
2 b	$\emptyset$	$\{\langle(delete, 1.0, /),$ $(delete, 1, /)\rangle\}$	$\{nos_\emptyset\}$	$\{\langle(add, 2, b),$ $(add, 2.0, c)\rangle\}$	$\emptyset$
3 b	$\emptyset$	$\emptyset$	$\{\langle(delete, 2.0, /),$ $(delete, 2, /)\rangle\}$	$\{nos_\emptyset\}$	$\boldsymbol{\{\langle(add, 3, c)\rangle\}}$

Figure 5.8: Content of the matrix M after backtracking

M	0 root	1 a	2 b	**3** ***c***
0 root	$\{nos_\emptyset\}$	$\emptyset$	$\emptyset$	$\boldsymbol{\emptyset}$
1 a	$\emptyset$	$\{nos_\emptyset\}$	$\{\langle(add, 1, b),$ $(add, 1.0, c)\rangle\}$	$\boldsymbol{\emptyset}$
2 b	$\emptyset$	$\{\langle(delete, 1.0, /),$ $(delete, 1, /)\rangle\}$	$\{nos_\emptyset\}$	$\boldsymbol{\{\langle(add, 2, c)\rangle\}}$
3 b	$\emptyset$	$\emptyset$	$\{\langle(delete, 2.0, /),$ $(delete, 2, /)\rangle\}$	$\boldsymbol{\{\langle(relabel, 2, c),}$ $\boldsymbol{(delete, 2.0, /)\rangle\}}$

Figure 5.9: Content of the matrix M after the next backtracking

5.2.2 Properties, Experiments and State-of-the-Art Comparison

In Amavi et al. (2013) we formally define the objects used by our approach (an XML tree, a schema, a subtree, a partial tree, a tree language) and their properties (validity, local validity, partial validity), operations on nodes (relabeling, adding and deleting) and on subtrees (inserting and removing), operation sequences, their equivalence and cost. In this context the *distance* between a tree t and a tree language L denotes the minimum tree-to-tree distance between t and any tree in L. The *tree correction set* is the set of all valid trees (i.e. belonging to L) whose distance from t is no greater than the threshold th.

The algorithm takes four parameters:

- t: an XML tree to be corrected,
- S: a structure description (a DTD),
- th: a natural threshold,
- c: an intended root tag of the resulting trees.

It returns the set of node-edit operation sequences allowing to get the tree correction set from t.

We prove the termination, soundness and completeness of the algorithm, i.e. we show that it always terminates, that each returned operation sequence is a valid correction (within th), and that each valid correction (within th) is returned.

The time complexity is equal to $O((f_t+1) \times (f_S)^{|t|+th} \times 6 \times |\Sigma| \times (|t|+th))^{th}$, where f_t is the maximum fan-out of t (the maximum number of children of any node in t), f_S is the maximum fan-out of all states in all finite-state automata in the schema S, $|t|$ is the size of t (the number of its nodes), and $|\Sigma|$ is the size of the alphabet in the schema S.

Several experiments were conducted in order to examine the performances of our algorithm on real-life data in function of different parameters: (i) the document size, (ii) the threshold value, (iii) the number of errors, (iv) the position of an error, (v) the nature of the DTD. We have used a large XML file in the TEI format stemming from the named entity annotation level of the National Corpus of Polish (cf. Section 4.3). The corresponding DTD seemed appropriate for the experiments since it defines elements concerned by a varying degree of flexibility. Six testing scenarios were designed so as to test:

1. the correction time and the number of correction candidates found in function of the document size,

2. the correction time in function of the distance threshold th,

3. the correction time in function of the number of errors introduced in the corrected file,

4. the correction time and the number of correction candidates in function of the position of the error in the corrected file, and of the nature of the DTD,

5. the correction time and the number of candidates in function of the distance threshold th, when the corrected document is empty,

6. the algorithm's behavior when only minimal-cost corrections are calculated.

The results show that, despite its theoretical exponential time complexity, our algorithm shows a behavior which is rather polynomial in function of the threshold value and of the document size. Surprisingly enough, the higher the distance of the corrected document from the schema, the shorter the correction time. This is probably due to the fact that errors appearing close to the beginning of the document rapidly reduce the correction time, which results from the left-to-right processing of each siblings' level. Understandingly, if errors appear in the parts of the file which are concerned by optionality, alternative, and unbounded repetitions of elements in the schema, their correction is more time consuming than in the non-ambiguous part. Finally, the correction time is closely correlated with the number of correction candidates found, which on its turn directly results from the above-mentioned factors: the size of the input document, the threshold value, the position of an error and the nature of the schema. This correlation between the correction time and the number of candidates might explain, at least partly, why the tree-to-language correction problem, as defined in our approach, is more difficult to solve than in some other works discussed in Section 5.1.3. Namely, this problem is frequently reduced in the literature to finding the tree-to-language distance only, without proposing a particular correction sequence, or to proposing a fixed number of minimal sequences only. If completeness of the correction set is required, the correction time grows accordingly.

We also performed a contrastive study of the existing tree-to-language correction algorithms, including ours, with respect to the problem definition, the informativeness of the documentation and the availability issues. The following aspects were taken into account:

- the elementary edit operations admitted on trees (node relabeling, insertion or deletion, possibly restricted to leafs or non-root nodes),

- the validity aspects (well-formedness, structural validity and correctness of attributes),

- the algorithm's output (tree-to-language distance; minimal, k closest or all corrections; edit sequences),

- the schema type (a DTD, an XML schema, an extended DTD) and model (a tree automaton, a set of regular expressions, a hedge automaton, an ordered labeled tree, a pushdown automaton, etc.),
- the model of the XML document to be corrected (a ranked or unranked ordered labeled tree, possibly serialized into a word of tags),
- time and space complexity,
- availability of proofs (for correctness, completeness, termination and complexity),
- the nature and the availability of the experimental data,
- the availability of the binaries and of the source code.

This contrastive study show the diversity of the approaches with respect to the problem definition, which makes their direct comparison hard to perform. There is an interesting correlation between how different approaches view the XML document, and which schema model and elementary operations they select.

- If the XML document is seen as a **tree**, the schema must obviously be a tree grammar (local or single-type), sometimes represented in a particular way, e.g. as a tree Tekli et al. (2007); Bertino et al. (2004, 2008) or as a hedge grammar Xing et al. (2006). The well-formedness is not an issue here since an ill-formed document is not a tree. In this case, i.e. in Xing et al. (2006), Staworko & Chomicki (2006), Tekli et al. (2007), Svoboda (2010); Svoboda & Mlýnková (2011) and in our proposal, the most natural elementary operations (except node relabeling) seem to be those concerning leaves rather than internal node. An exception to this rule is Suzuki (2007), and possibly also Boobna & de Rougemont (2004).
- The remaining approaches (Staworko et al., 2008; Thomo et al., 2008) view an XML document as a **word** of opening and closing tags and the schema is transformed to a pushdown automaton on words. This view offers a rather natural framework for well-formedness issues (e.g. correcting a missing closing tag comes down to inserting a character in a word). But most importantly, also elementary edit operations on internal tree nodes seem to be rather natural in this context.

5.3 Incremental Algorithms on Words and Trees

Another part of my contributions concerns the dynamic and incremental setting of string and tree algorithms. Recall that the Boytsov (2011) in his state-of-the-art survey of string-to-language correction algorithms concentrates on methods dedicated to infrequently updated dictionaries, which are used primarily for retrieval. In some applications however the **instability of the vocabulary**, i.e. the necessity of adding, deleting or modifying the list of valid words, can occur on a regular basis. The challenge is then to efficiently update the corresponding data structures and preserve their properties with respect to the unchanged part of the vocabulary.

5.3.1 Incremental String and Tree Validation and Correction

Under these premises, in (Cheriat et al., 2005) we introduced the problem of **incremental string-to-language correction**, in the sense that an invalid word w to be corrected results from another valid word w' by the application of a sequence S of updates (elementary edit operations). The aim is to correct w in such a way that the resulting words are not only close

to w but also can be reached from w' by an edit sequence similar to S. In this way, priority is given to solutions which take the user's intension to modify w' into account.

The motivation for the incremental string-to-string correction comes from the area of XML-document validation and correction (cf. Section 5.1.3). The validity of each node in such a document is described by a regular expression (in case of a DTD) or by a set of regular expressions (in case of an XML schema). When a user wishes to modify a valid document but performs a set of invalid updates (i.e. leading to an invalid tree) we may start with *locally* validating and correcting the nodes concerned by the updates, together with their closest neighborhood: fathers, siblings, and sons. Since each set of siblings may locally be viewed as a string, we reduce a part of the tree correction to the string-to-string correction problem. Thus, we may often obtain our first valid correction candidates without even touching good parts of the whole tree (those that remain unchanged with respect to the initially valid XML tree) which allows to spare computation time and space.

In Bouchou et al. (2006b,a) we extend and implement these ideas of incremental validation and correction of XML documents. When user's updates are applied to a valid XML document, an incremental validator verifies whether the updated document complies with the schema, by re-validating only the parts of the document involved in the updates. If the re-validation fails for a node at position p, a correction routine is called for the subtree rooted at p. This routine assumes that p's label l is correct and considers corrections over its descendants, according to, roughly, the tree-to-language correction algorithm decribed in Section 5.2. This approach does not guarantee the completeness of the correction set since valid trees (within the threshold) can also be obtained by modifying the root label of the locally corrected subtree. This may extend the correction to the siblings and ancestors of position p, thus to the whole tree. Restraining the correction to the local subtree gives, however, priority to the user's updates since it does not go far beyond the positions which the user intended to modify.

5.3.2 Handling Dynamic Vocabularies in Finite-State Automata

Some of my contributions address incrementality issues concerning natural language vocabularies and finite-state automata representing them. An example of an incremental construction of a minimal acyclic finite-state automaton is presented by Daciuk et al. (2000). In this approach words are added to the FSA in the lexicographic order in such a way that inserting a new word requires a modification of the states belonging to the path followed during the insertion of the previous word. Thus, after each addition the FSA remains minimal with no need of applying the minimization algorithm to the whole set of states. When the input data are unsorted the incremental algorithm is similar up to an additional step of cloning all states on the path followed while adding the current word. Here again the necessary modifications are local and the FSA remains minimal after each newly added word.

In (Daciuk et al., 2005b) we address a similar problem: the one of an **incremental and semi-incremental construction of pseudo-minimal automata**. A pseudo-minimal automaton is a minimal acyclic automaton that has a proper element (a transition or a state) for each word belonging to the language of the automaton. That proper element is not shared with any other word, and it can be used for implementing a function on words belonging to the language, for instance perfect hashing (see below). A pseudo-incremental construction is a one in which the intermediate automaton is non necessarily minimal but factorized with respect to suffixes and prefixes of the added words. We propose three algorithms to solve the abovementioned problem. They can be used with lexicographically sorted data, unsorted data, or data sorted on decreasing length, and result from slight modifications of known algorithms for the incremental and semi-incremental construction of minimal deterministic acyclic automata, including the one in (Daciuk et al., 2000). The modification consists mainly in verifying the following property:

in a pseudo-minimal automaton, there is no path on which a divergent state (i.e. a state with more than one outgoing transition) follows a convergent state (i.e. a state with more than one incoming transition). Experiments on word lists extracted from English, French and Polish corpora show that our algorithms are slower than the one by Revuz (1991), the latter however needs an unusual sorting of input data (in reverse lexicographic order).

In (Daciuk et al., 2005a) we address the problem of perfect hashing, i.e. of a mapping between a set of n unique words and n consecutive numbers. It can apply, for instance, to information retrieval, where numbers associated to words are used as pointers to lists of positions of this word in a text (i.e. they form the inverted text). This enables an efficient search and a high compression. In FSA-based static perfect hashing this mapping usually corresponds to the alphabetical order of the words of the FSA's language. It is implemented by adding a weight to each transition (or to each state)[5]. When a word is searched for in this FSA, the sum of the weights on the followed transitions results in the hash number for the word. With this solution, adding a new word may unfortunately change the mapping of many previous words.

We propose three methods for FSA-based **dynamic perfect hashing**, i.e. the one in which adding a new word does not change the mapping for the previous words. The first method uses static perfect hashing with a minimal automaton (MA), accompanied by a translation vector, to compensate for the change. In the second method no additional vector is used but the weights on transitions are adapted so as to reflect the word addition order. As a result some transitions have to be duplicated (in order to get different weights) and the resulting weighted automaton (WA) may not be minimal. The third solution uses a pseudo-minimal automaton (PA, see above) in that the number assigned to a word is placed on the proper transition of this word. A WA and a PA can theoretically be exponentially bigger than the corresponding minimal automaton. However, experimental tests on corpora of different sizes (7 thousand, 1.5 million and 44 million words), languages (English, Polish and French) and types (technical, literary and press) showed that: (i) for small vocabularies of up to 100 words the WA and the PA have the same size as the minimal automaton, (ii) with a growing corpus size, the $|PA|/|MA|$ ($|WA|/|MA|$) ratio grows continually up to 1.75, and then decreases slightly when the corpus size rises above 30-million words.

5.4 Contributions and Perspectives

In the light of the contrastive state-of-the-art analysis of the tree-to-language correction mentioned in Section 5.2, we believe that our correction algorithm documented in (Amavi et al., 2013) offers the first full-fledged study of the document-to-schema correction problem. Not only do we measure the distance between a document and a schema but also find the candidate correction trees. We do not limit ourselves to finding the minimal solution but find all solutions within a threshold instead. Thus, we consider the correction as an enumeration problem rather than a decision problem, contrary to most other approaches. Our documentation is one of the few which includes the complexity, correctness and completeness proofs. Our contribution also seems to be the only one that offers, in addition to the executable and the source code, also the user's guide and the set of testing data used to obtain the experimental results. Consequently, it seems to be the only reproducible one. Last but not least, our source code is the only one to be distributed under a known license, namely the open license GNU LGPL v3[6].

Some recent approaches such as Suzuki (2007), Staworko et al. (2008) and Svoboda & Mlýnková (2011) shed new light on the document-to-schema correction problem in that they

[5]This weight is equal to the ordinal number of the first word recognized by traversing this transition minus the sum of weights on the preceding transitions.

[6]http://www.info.univ-tours.fr/ savary/English/xmlcorrector.html

introduce edit operations acting on internal nodes, extend the schema's expressive power to XML schemas instead of DTDs only, and offer optimizations of data structures via graph-based modeling. One of our perspectives is to examine how these proposals can be integrated with ours so as to propose a more universal framework in which different variants of the correction problem might be solved most efficiently.

As far as incremental algorithms of finite-state-tools are concerned, I wish to experiment with their applicability to various natural language data. Notably, the minimal, weighted and pseudo-minimal automata implementing dynamic perfect hashing might be used as compression techniques for electronic dictionaries containing multi-word expressions. Note that compiling a MWE list into a finite-state automaton may result in a much lower compression rate than in case of single words (Savary, 2000) due to inflection of non-final components of MWEs. I think that the compression efficiency might be enhanced if a hashing automaton is used to encode single MWE components, and if a MWEs are then represented as sequences of hash numbers.

Chapter 6

Research Framework and Management

Research nowadays is an extremely challenging activity. Researchers are expected to show versatile and complementary skills, which in other domains are covered by several distinct professions. Scientific competence, even if considered crucial, is only a part of these abilities, which also include teaching, strategic planning, project management, event organization, evaluation and self-evaluation, software development, foreign language proficiency, international relations, human resource management, reporting, accountancy, and many others. To be a senior researcher capable of playing a leading role in the community is to be able to achieve most of these abilities, often through self-training. Moreover, quality requirements and professional ethics, make us face challenges which may come into conflict, e.g. if competitiveness and scientific rigorousness are to be considered simultaneously.

Since this dissertation is meant to plead my accreditation to lead and supervise research activities, I dedicate this chapter to describing my major contributions which helped me acquire some of the skills mentioned above.

6.1 Natural Language Processing Research in Blois and Tours

The local framework of my scientific activity through the past 11 years has been the *BDTLN* (*Bases de Données et Traitement Automatique des Langues* 'Databases and Natural Language Processing') research team of the *Laboratoire d'infomatique* (*LI*, 'Computer Science Laboratory') at the *Université François Rabelais Tours* in France. The large majority of our team's activity, including my own, is located in *Blois*, a city of about 50,000 inhabitants, which hosts a university campus of about 1,500 students, 60 kilometers away from the university headquarters in Tours.

BDTLN offers a diverse expertise ranging from information systems, data mining and data warehouses, through XML, semantic web and web services, through natural language processing, language resources and tools, corpus linguistics, human-machine interfaces and finite-state algorithmics. This is a very stimulating context promoting open-mindedness and crossing barriers between domains, which resulted in innovative interdisciplinary proposals, such as XML document correction or OLAP session recommendation, inspired from spelling correction assets in NLP. At the same time, the relatively small size of the team, and of its subset dedicated to the NLP activity in particular (roughly 4–5 permanent members and 1–3 PhD and post-doctoral students), as well as its geographical distance from the LI lab's headquarters in Tours, are challenging enough to make intensive external collaborations even more inevitable than in large scientific centers.

6.2 External Collaborations

At the international level, for obvious reasons of my origins and mother tongue, I have established particularly strong links with Poland, and notably with the *Institute of Computer Science of the Polish Academy of Sciences in Warsaw, Poland* (*IPIPAN*). This collaboration was initiated by a bilateral French-Polish EGIDE Polonium project (now called a PHC project, *Projet Hubert-Curien*), co-proposed and co-directed by Marcin Woliński and myself. In 2009-2010 I was a visiting researcher at the IPIPAN, where I directed three workpackages in an ERDF project (NEKST) and a national project (NKJP). I have also participated in three other Polish national projects coordinated by the IPIPAN: a Polish spin-off of the European LUNA project, CORE and CLARIN.PL. I was a subcontractor in the European CESAR project. This intensive collaboration yielded several language resources and tools for Polish, available under open licenses via the *Computational Linguistics in Poland*[1] portal, as well as 18 international and national publications, which I co-authored with Małgorzata Baron, Marta Chojnacka-Kuraś, Monika Czerepowicka, Katarzyna Głowińska, Celina Heliasz, Mateusz Kopeć, Aleksandra Krawczyk-Wieczorek, Michał Lenart, Filip Makowiecki, Leszek Manicki, Małgorzata Marciniak, Maciej Ogrodniczuk, Jakub Piskorski, Adam Przepiórkowski, Piotr Sikora, Danuta Skowrońska, Paweł Śliwiński, Jakub Waszczuk, Anna Wesołek, Joanna-Rabiega Wiśniewska, Marcin Woliński, Bartosz Zaborowski and Magdalena Zawisławska. In 2012, Adam Przepiórkowski was the official proposer of the PARSEME COST action and is now the Vice-Chair of its Management Committee, while the IPIPAN plays the role of the Grant Holder.

I have also been involved in collaborations with 3 Polish universities. In 2004 I initiated and organized a 2-month scientific stay of a visiting professor, Jan Daciuk, from the *Gdańsk University of Technology*, funded by the French ministry (*Contingent national des professeurs*). This collaboration yielded 2 common international publications. Since 2010, I have been collaborating with Krzysztof Jassem and Filip Graliński from the *University of Poznań*. We have co-authored one international publication, co-proposed a collaborative European project (SHAGRALER) and co-developed language resources and tools within an ERDF project (NEKST) and a European project (CESAR). Within the NEKST project I have also established a close collaboration with Monika Czerepowicka from the *University of Olsztyn*. We have co-authored SEJF, an electronic dictionary of Polish phraseology and an international publication. Monika is now the Polish representative a the Management Committee of the PARSEME COST action.

My links with the *University of Belgrade* in Serbia have been initiated via the bilateral French-Serbian EGIDE project Pavle Savic (now called a PHC project, *Projet Hubert-Curien*). I collaborated with Cveana Krstev and Ranka Stanković for the integration of Multiflex in LeXimir, a lexicographic framework for the creation and management of electronic dictionaries (cf. Section 3.3.5). I co-authored two publications with Cvetana Krstev and Duško Vitas. Now, Cvetana and Ranka are the Serbian representatives at the Management Committee of the PARSEME COST action, and Cvetana is also its Steering Committee member in charge of short-term scientific missions.

In 2011, when organizing the CIAA/FSMNLP conference in Blois, I initiated the invitation of over 10 participants via the national ACCES fund offered by the French Ministry of Higher Education and Research to researchers from Central and Eastern Europe participating in conferences organized in France. In this way, I met Nina Yevtushenko and Natalia Kushik from the *Tomsk State University* in Russia. In 2012 I was a visiting professor at this university during one weak.

At the national level, I have been involved in a long-lasting collaboration with the *Laboratoire d'informatique Gaspard Monge* (LIGM), at the Université Paris Est Marne-la-Vallée,

[1] http://clip.ipipan.waw.pl/LRT

from which I graduated with a PhD degree in 2000. I am notably a co-developer of *Unitex*[2], a MWE-aware multilingual corpus processor, distributed under the LGPL license. As a result of a subcontracting work within the national RNTL Outilex project, Multiflex (cf. Section 3.3) was integrated with Unitex as a module for the creation and inflection of electronic lexicons for contiguous Multi-Word Expressions. Recently, I have also been involved in a close collaboration with Matthieu Constant in organizing the CIAA/FSMNLP-2011 conference in Blois, in PARSEME COST action management (Matthieu is the French representative at PARSEME's Management Committee) and in setting up the ANR PARSEME-FR project proposal.

At the regional level, the LI laboratory in Tours/Blois has been recently building a scientific federation with its counterpart in Orléans, the *Laboratoire d'Informatique Fondamentale d'Orléans* (LIFO). Since Blois is geographically located between Tours and Orléans, the BDTLN team has been playing an active role in this process. My personal involvement is based on tight links with the *CA* (*Contraintes et Apprentissage*, 'Constraints and Machine Learning') team, whose activity is partly dedicated to NLP. I was a co-proposer, with the LIFO as coordinator, of a collaborative European project (SHAGRALER) in 2012, to which I brought two Polish partners (IPIPAN Warsaw and the University of Poznań). I have been closely collaborating with Yannick Parmentier in setting up the PARSEME COST action proposal, in its evaluation by the COST office, and in its current management (Yannick is the French representative at PARSEME's Management Committee and the leader of one of the 4 working groups). Since last October, Denis Maurel, Yanick and myself are co-supervising a PhD thesis by Jakub Waszczuk, dedicated to Multi-Word Expressions and parsing.

6.3 Bibliometrics

As far as quantitative productivity data are concerned, since my PhD I have authored and co-authored 41 publications, including:

- 9 papers in international peer-reviewed journals,
- 4 book and collection chapters,
- 19 papers in the proceedings of international and national peer-reviewed conferences, 8 of which appeared as Springer special issues (Lecture Notes in Computer Science, Lecture Notes in Artificial Intelligence, or Advances in Soft Computing),
- 4 papers in the proceedings of international peer-reviewed workshops,
- 5 technical reports.

Most of these publications (34) are written in English, 5 in French and 2 in Polish. I was a unique author of 7 of them, and I co-authored the other 34 of them with 42 French, Polish and Serbian collaborators.

Even if various selectivity and impact factors are not always reliable in research evaluation several examples of such statistical data are cited in Table 6.1 (mostly according to the websites of the corresponding journals). The selected papers are mostly available via the URL links cited in the bibliography.

The interdisciplinary nature of my research is visible through the publication venues in the domains of computer science, computational linguistics, and linguistics. The DBLP Computer Science Bibliography contains 17 of my publications. The citation lists from the: (i) ACM Guide to Computing Literature, (ii) Microsoft Academic Search, and (iii) Google Scholar, index

[2] `http://www-igm.univ-mlv.fr/~unitex/`

Table 6.1: Visibility and impact factors of my selected publications

Publication	Journal/Conference	Impact Factor or Indexation
(Amavi et al., 2013)	The Computer Journal	IF=0.755, 5-year-IF=0.954; A* in CORE[3]
(Waszczuk et al., 2013)	International Journal of Data Mining, Modelling and Management	Scopus (Elsevier), Academic OneFile (Gale), DBLP Computer Science Bibliography, Expanded Academic ASAP (Gale), Cabell's Directory of Publishing Opportunities, and Excellence in Research for Australia (ERA): Journal list 2012
(Savary & Piskorski, 2011)	Control & Cybernetics	h-index=22; SJR=0.35; IF=0.38; C in CORE; SciSearch®, Research Alert®, CompuMath Citation Index®, and Current Contents®/ Engineering, Computing & Technology
(Savary, 2009)	CIAA 2009	B in CORE[4]
(Agafonov et al., 2006)	META	ISI®(Arts & Humanities, Arts & Humanities Citation Index), Scopus, IBZ (Internationale Bibliographie der Geistes- und Sozialwissenschaftlichen Zeitschriftenliteratur), IBR (Internationale Bibliographie der Rezensionen Geistes- und Sozialwissenschaftlicher Literatur), Francis, Google Scholar, INIST, MLA International Bibliography, Repère, Bibliographic Index (Active), Linguistic Bibliography, ERIH, AHCI-SSCI, AERES, and Association canadienne des revues savantes (ACRS)
(Tran et al., 2006)	Lingvisticæ Investigationes	Cultures, Langues, Textes; ESF European Reference Index for the Humanities - Linguistics; Germanistik; IBR/IBZ; Language Abstracts; Linguistic Bibliography/Bibliographie Linguistique; LLBA; MLA Bibliography; and TSA Online
(Savary, 2005)	Archives of Control Sciences	C in CORE; Scopus, EBSCO HOST, Mathematical Reviews, Zentralblatt MATH, VERSITA and BazTech
(Daciuk et al., 2005b)	CIAA 2005	B in CORE[5]

(i) 10, (ii) 18, and (iii) 23 of my publications, respectively, as well as (i) 16, (ii) 35, (iii) over 150 citations to them in publications which I did not co-author. My automatically generated Google Scholar h-index (including auto-citations and erroneous references) is equal to 9.

6.4 Software Development

I am the author of Multiflex (cf. Section 3.3), a formalism and its implementation for an automatic inflection of Multi-Word Units. It has been developed in C and is integrated into 3 linguistic platforms:

- Unitex[6], a multiligual corpus processor available under the GNU LGPL v3 license,
- Toposław[7], a Polish lexicographic framework distributed under the GNU GLP v3 license,
- LeXimir[8], a Serbian tool for lexical resource management and query expansion, available from the author.

I co-authored XMLCorrector[9], a Java implementation of an algorithm for correcting an XML document with respect to a DTD (cf. Section 5.2). It is distributed under the GNU LGPL v3 license.

[6] `http://www-igm.univ-mlv.fr/~unitex/`
[7] `http://zil.ipipan.waw.pl/Toposlaw/`
[8] `http://korpus.matf.bg.ac.rs/soft/LeXimir.html`
[9] `http://www.info.univ-tours.fr/~savary/English/xmlcorrector.html`

6.5 Project Development and Management

Research nowadays is increasingly funded via short or mid-term projects at all administrative levels (European, national, and regional). The challenge is to be able to define challenges and propose innovative solutions in a way which appeals to the funding bodies and political decision makers. Selected projects require huge administrative effort, a strict project management, tight deadlines and precise deliverables.

Table 6.2 contains basic data about projects in which I played a crucial role by setting up project proposals and/or by project coordination. Table 6.3 resumes projects in which I participated as a workpackage leader, collaborator or subcontractor. In total I have contributed to 14 collaborative projects at the European, national (in France and in Poland) and regional (in Région Centre, France and in Mazowsze, Poland) level. In the following section I describe my roles in 4 selected projects.

6.5.1 PARSEME

My major, and the most recent, project experience is strictly related to PARSEME (PARSing and Multi-word Expressions)[10], the IC1207 COST action[11].

COST[12] is an inter-governmental framework (founded in 1971) dedicated to the coordination of nationally funded European research. Its budget stems from the 7th Framework Program via the European Science Foundation. Each COST action follows a bottom-up approach (the scientific challenges are defined by researchers themselves) and its objectives are to overcome research fragmentation issues by creating networks of experts working on related topics in different countries. COST supports cooperation and dissemination (meetings, workshops, short-term missions, training schools, etc.) but no direct research funding in provided. COST action members are countries (not institutions) and each action is supposed to be open to all members from the participating countries all through the action's lifetime. Typically a large number of countries is involved (about 20), and up to 4 institutions from non-COST countries are admitted. Important roles are given to early-stage researchers (researchers with no more than 8 years of experience after their PhD). A yearly budget amounts to about 129,000–156,000 € per year for all partners. Despite this relatively low funding, COST actions are very competitive with a proposal selection rate around 6-10%.

Under these premises, the initiative to build a European COST network dedicated to parsing and MWEs was undertaken by Adam Przepiórkowski (IPIPAN, Warsaw, the official proposer), Agnieszka Patejuk (IPIPAN), Yannick Parmentier (LIFO, Orléans, France) and myself. I was the main author of the written proposal, which consisted in describing the action's background and motivation, defining its objectives and its scientific program structured around 4 working groups (cf. Section 7.3), as well as setting up management structures and a preliminary timetable. The initial consortium gathered by the 4 proposers consisted of several dozens of representatives from 20 countries. It was constituted so as to account for the variety of European languages, as well as for most major linguistic theories (HSPG, LFG, TAG, etc.) and methodologies (knowledge-based and data-driven). In late 2013, i.e. 8 months after the action started, the network amounts to over 100 members from 28 countries, including 75 official representatives in the Management Committee.

Since the action's kick-off meeting I have been playing the role of the Management Committee (MC) chair. It consists in coordinating and implementing the action via:

[10] www.parseme.eu

[11] http://www.cost.eu/domains_actions/ict/Actions/IC1207

[12] http://www.cost.eu/

Table 6.2: My development and coordination of funded collaborative projects

Project	**Dates**	**Budget**	**Coordinator**	**Funding**	**Topics**	**My contributions**	**% research time**
PARSEME	2013–2017	approx. 680,000 €	A. Savary	COST[a]	Parsing and Multi-Word Expressions	Main proposal author, chair of the Management Committee	50%
PARSEME-FR	2013	submitted	Université Paris Est	ANR[b]	Syntactic parsing and multiword expressions in French	Proposal co-author, coordinator for the UFRT	5%
SHAGRALER	2011	non selected	Université d'Orléans	European Commission, STREP project, call FP7-ICT-2011-SME-DCL	Sharing grammatical and lexical resources	Proposal co-author, coordinator for the UFRT	20%
Polonium	2007–2008	6,070 €	LI UFRT[c] & IPIPAN[d]	PHC EGIDE[e]	Inflexional morphology and variation of Polish proper names, automating the construction of language resources	French coordinator: project proposal, adapting Multiflex to Polish	25%

[a] `http://www.cost.eu/`, funded by the European Science Foundation
[b] Agence Nationale de la Recherche 'National Research Agency', funding short-term research contracts
[c] Laboratoire d'informatique, Université François Rabelais Tours
[d] Institute of Computer Science, Polish Academy of Sciences
[e] Bilateral project co-funded by the French and Polish Ministries of Higher Education and Research

Table 6.3: My participation in funded collaborative projects

Project	**Dates**	**Budget**	**Coordinator**	**Funding**	**Topics**	**My contributions**	**% research time**
CORE	2011–2014	120,000 €	IPIPAN[a]	NCN[b]	Coreference annotation and resolution	Consultant in corpus annotation	5% since 2011
CESAR	2011–2013		Hungarian Academy of Sciences	European Commission (CIP-ICT-PSP-271022)	Open linguistic infrastructure for Central and South-East European resources	Subcontractor: Polish resources for proper names, supervision of a 6-person team	30 % in 2012
NEKST	2009–2014	3,500,000 €	IPIPAN & PWr[c]	ERDF[d]	Information retrieval, morphological and syntactic analysis, language resources, QA systems, opinion and text mining	Leader of 2 workpackages, supervision of a 5-person team	20% in 2009–2011
CODEX	2009–2012	68,336 €	INRIA[e] Saclay	ANR[f]	Efficiency, evolution and composition for XML: models, algorithms and systems	Implementation of an algorithm for XML document correction, co-supervision of 2 student projects and 1 research engineer	4 months full time
NKJP	2007–2010	600,000	IPIPAN	MNSW[g]	Construction of the National Corpus of Polish	Leader of the named entity annotation workpackage	50% in 2009–2011
LUNA.PL	2008–2009		IPIPAN	MNSW	Polish spin-off of the European IST-033549 LUNA project. Proper names in spoken dialogues.	Adapting Multiflex to Polish, proper name resource creation	40 %
EmotiRob	2007–2009	85,200 €	UUB[h]	ANR	Companion robot for weakened children	Emotion annotation, compositional model for emotion computing	10%
Pavle Savic	2004–2005	5,500 €	LI UFRT & Belgrade University	PHC EGIDE	Construction and application of a French-Serbian relational dictionary of proper names	Extending Multiflex to Serbian	40%
Outilex	2002–2006	7,700 € (for UFRT)	Université de Marne-la-Vallée	RNTL	Finite-state language processing platform, interoperability of language resources and tools	Subcontractor: integrating Multiflex under Unitex	40%
NomsPropres	2003–2005	94,000 €	LI UFRT	RNTL[i]	Construction and application of a multilingual relational dictionary of proper names	PhD co-supervision (M. Tran)	40%

[a]Institute of Computer Science, Polish Academy of Sciences
[b]Narodowe Centrum Nauki 'National Science Center' in Poland
[c]Wrocław University of Technology
[d]European Regional Development Fund
[e]Institut national de recherche en informatique et en automatique 'National Research Institute in Computer Science and Automation'
[f]Agence Nationale de la Recherche 'National Research Agency', funding short-term research contracts
[g]Ministerstwo Nauki i Szkolnicta Wyższego 'Ministry of Research and Higher Education' in Poland
[h]Université Européenne de Bretagne 'European University of Bretagne'
[i]Réseau National en Technologies Logicielles 'National Software Technology Network', a national body supporting industry/academia collaboration

- convening and chairing the MC meetings, preparing the meeting agenda and validating the minutes,
- preparing the action's annual work and budget plan and negotiating it with the COST office,
- initiating and managing MC votes,
- communicating the MC expenditure approvals to the grant holder,
- defining which participants are entitled to reimbursement for their activities within the action,
- approving payments,
- reviewing and approving yearly financial reports,
- representing the action at the COST Annual Progress Conference of all action chairs.

An efficient day-to-day action management and the preparation of MC decisions are done by the Steering Committee (of 9 most active members) which I convene and chair. I also represent the action to external bodies and partners.

PARSEME is a very exciting initiative and it receives a growing attention from the international community. We have organized the first general scientific meeting in Warsaw in September 2013 with over 50 participants. The second meeting will take place in March 2014 in Athens. PARSEME also endorses and co-organizes the well established annual Multi-Word Expressions workshop[13], which will be co-located with the EACL conference in Gothenburg, Sweden in April 2014. We are also currently funding 6 short-term scientific missions of 1 to 10 weeks, mainly supporting early-stage researchers.

My personal assets from chairing PARSEME are enormous. I have the chance to establish close contacts with a large scientific community centered around one of my major scientific interests. I carry on a daily collaboration with prominent researchers. I participate in meetings and discussions which contribute to fundamental research in NLP. I receive many incentives to continually increase my organization and management skills. I love PARSEME's highly multilingual context, in which the richness of the European linguistic heritage is represented and discussed, and which brings more balance with respect to the traditionally dominant role of English in the NLP research.

6.5.2 National Corpus of Polish

The National Corpus of Polish (*Narodowy Korpus Języka Polskiego*, NKJP)[14] was a 3-year Polish national project (2007–2010), involving a consortium of four main Polish corpus creators, coordinated by the IPIPAN, Warsaw (Przepiórkowski et al., 2008, 2010). The aim of the project was to create a 1.5-billion (10^9) word corpus of Polish annotated at various levels, with a 300-million word balanced subcorpus and a number of annotation tools. A 1-million word gold standard subcorpus was manually annotated at all annotation levels.

I joined the project in 2009 during my sabbatical visit to IPIPAN, and I became the leader of the workpackage (WP) dedicated to the named entity annotation level (cf. Section 4.3). My duties included: (i) the conceptual definition and documentation of the task, (ii) defining the annotation methodology, (iii) setting up and adjusting the WP budget, (iv) recruiting, training

[13] `http://multiword.sourceforge.net/PHITE.php?sitesig=CONF`
[14] `http://nkjp.pl/`

and supervising a team of 5 linguist annotators, (v) developing language resources and tools for automatic pre-annotation of the gold standard corpus, (vi) supervising the development of the annotator's workbench and of machine-learning tools for automatic annotation of the whole 1.5-billion word corpus.

The WP management was performed jointly with the WP dedicated to annotating syntactic words and groups. This gave me an insight to syntactic annotation and its links with named entities, which is now useful in the context of PARSEME (cf. Section 7.3). Methodological principles in annotating NKJP were defined so as to maximize the objectivity of judgment. Namely, each corpus text was annotated by two independent annotators who knew nothing about each other's choices, except what they could learn via the discussion list. Annotation conflicts were then adjudicated, whereas an adjudicator never reviewed a text that he or she has previously annotated.

This project gave me a large experience in linguistic modeling via corpus annotation. Studying corpus examples on a regular basis resulted in a good understanding of the linguistic phenomena in named entities. It was also my first large experience in managing a research team of, mostly, young researchers and students. I gained a conviction that a proper management of a linguistic annotation project consists in putting the linguistic inquiries (rather than the future benefit for training machine-learning tools) in the heart of the decision making processes. I also understood that the annotation task, which is frequently considered as tedious by NLP researchers, can become an exciting and motivating experience for linguists if only they are associated to decision making and convinced about the priority of the language modeling objectives.

6.5.3 CESAR

CESAR[15], carried out in 2011–2013, was a collaborative CIP ICT-PSP European project, a part of the META-NET network of excellence dedicated to fostering the technological foundations of a multilingual European information society.

CESAR was dedicated to creating open linguistic infrastructures for Central and South-East European resources. While I was only a remote subcontractor of this project, I really appreciated its idea and outcome for the Polish computational linguistics. The project was meant to enhance, upgrade, standardize and cross-link a wide variety of language resources and tools and make them available via META-SHARE[16], a platform for sharing language resources and tools (LRTs) in a uniform and documented manner. CESAR funded in particular several LRTs which I co-developed (cf. Section 7.2). They are now available under open licenses and referenced in META-SHARE: NKJP, PCC, PNEG, Prolexbase, PNET, SAWA, SEJF, SEJFEK, and SEJFEK4Spejd.

6.5.4 CODEX

The ANR CODEX[17] project, carried out in 2009-2012, was dedicated to new challenges from the rapidly evolving domain of XML processing. It gathered a consortium of several major French actors in the field and yielded over 60 publications.

In this project I could enlarge my very interesting local collaboration with Béatrice Bouchou and Mírian Halfeld Ferrari Alves (now in Orléans), originated from our cross-domain discussions on links between XML and NLP. Within the PhD thesis by Ahmed Chériat we proposed an original algorithm for incremental XML document correction with respect to a DTD, inspired from the spelling correction problem in NLP. Together with Béatrice I later co-supervised the

[15] http://www.meta-net.eu/projects/cesar
[16] `http://www.meta-net.eu/meta-share`
[17] http://codex.saclay.inria.fr/

re-implementation of the algorithm in a more general framework by Alexandre Borel (a Bachelor student) and by Joshua Amavi (a post-Master researcher). This collaboration led to a large paper in an influential computing journal (Amavi et al., 2013) and to the publication of the algorithm's implementation and its experimental data under an open license[18]. I greatly appreciated the long-lasting nature of this scientific work, the rigorous and patient attitude of by colleagues towards complex formal and practical problems, and the good work organization.

6.6 Research Supervision

In my project management experience, mentioned in the preceding section, I have gained an experience in research team supervision (3 teams of 14 persons in total).

I also have a more limited experience in co-supervision of PhD and Master students, summarized in Table 6.4.

My collaboration with Jakub Waszczuk has started 4 years ago within the NKJP project. Jakub was the main programmer in the named entity and syntax annotation tasks. We co-authored 4 publications dedicated to this work. In 2013, Jakub was recruited as a PhD student in Blois for a PhD dissertation placed in the heart of the PARSEME COST action (cf. Section 7.3).

The PhD work by Mickaël Tran was dedicated to the initial development of Prolexbase (cf. Section 4.5). His role was to conceive the Prolexbase model and perform its implementation, as well as to develop collaborative lookup and edition tools and data exchange formats. My contribution to this work was related to multilingual lexicographic sorting principles necessary for the lexicographer's workbench, as well as to state-of-the-art studies in lexicographic knowledge bases.

Ionas Michailidis carried out a part-time, mostly remote, research activity, under my co-supervision, dedicated to named entity recognition (NER) methods for modern Greek. He published several articles in international and national conferences and workshops. Despite the fact that his PhD was not completed due to the lack of funding, and that my contribution was too modest to justify co-authoring of Ionas' publications, this early supervision experience initiated me into the domain of machine-learning NER which was beneficial in my later investigations.

The Master thesis in lingusitics by Małgorzata Spędzia initiated our close 2-year collaboration on an automated method for feeding Prolexbase from multilingual open data (cf. Section 4.5), documented in 2 common publications. Małgorzata was the main lexicographer in this task, funded by the CESAR and NEKST projects. She also contributed to the evaluation of the SEJFEK MWE lexicon via corpus annotation (Savary et al., 2012b) and to the PNET lexicon of NE triggers (cf. Section 7.1).

The Master thesis by Med El Amine Fahmi was the starting point for implementing the XML document correction algorithm addressed in Section 5.3. It was further pursued in another co-supervised Bachelor computing project by Alexandre Borel, and completed within my collaboration with Béatrice Bouchou and Joshua Amavi in the CODEX project.

The Master thesis by Abdesselem Beghriche led to his initiation in computational lexicography, notably by a thorough state-of-the-art study.

Let me also mention that I have informally contributed to three other theses in computer science which I did not officially supervise: the PhD thesis by Marc Le Tallec (Tallec et al., 2009, 2010b,a), the PhD thesis by Ahmed Chériat (Cheriat et al., 2005; Bouchou et al., 2006b,a), and the Master thesis by Piotr Sikora (Woliński et al., 2009; Marciniak et al., 2009b).

[18] http://www.info.univ-tours.fr/ savary/English/xmlcorrector.html

Table 6.4: My supervision of PhD and Master students

Student	**Dates**	**Affiliation**	**Diploma**	**Funding**	**Thesis**	**Supervision**	**Common papers**
Jakub WASZCZUK	since 2013	UFRT[a]	PhD in computer science	Ministry PhD Grant	Parsing and MWEs	Denis Maurel 5%, Yannick Parmentier 25%, Agata Savary 70% (estimation)	(Savary et al., 2010; Waszczuk et al., 2010, 2013; Savary & Waszczuk, 2012)
Michaël TRAN	2003–2006	UFRT	PhD in computer science	RNTL project	Prolexbase. A multilingual relational dictionary of proper names. Conception, implementation and on-line management.	Denis Maurel 90%, Agata Savary 10%	(Tran et al., 2006)
Ionas MICHAILIDIS	2003–2008 (uncompleted)	UFRT[b], TEIT[c], ECDN[d]	PhD in computer science	unfunded	Named entity recognition in modern Greek	Denis Maurel 10%, Nathalie Friburger 45%, Agata Savary 45%	
Małgorzata SPĘDZIA	2011	UFRT	Master in linguistics	CESAR, NEKST	Developing a Polish module for Prolexbase. A case study of toponyms.	Denis Maurel 50%, Agata Savary 50%	Savary et al. (2013b,a)
Med El Amine FAHMI	2009	UFRT	Master in computer science		Automatic correction of XML documents	Béatrice Bouchou 50%, Agata Savary 50 %	
Abdesselem BEGHRICHE	2004	UFRT	Master in computer science		Automatic tagging of an MWE lexicon	Agata Savary 100%	

[a]Université François Rabelais Tours
[b]Université François Rabelais Tours
[c]Technological Education Institute of Thessaloniki, Greece
[d]European Commission Delegation in Nigeria, Abuja

6.7 Research Evaluation

Peer reviewing is a fundamental quality assurance principle in research. My experience in this matter is manifold. As far as reviewing scientific publications is concerned:

- I am an editorial board member of the *Journal of Language Modelling*[19].
- I was a Scientific Committee member of three journal special issues:
  - *ACM Transactions on Speech and Language Processing*, 10(2-3) – Special Issue on *Multiword Expressions: from Theory to Practice and Use*[20],
  - *Traitement Automatique des Langues*, 54(2) – Numéro spécial *Entités Nommées*[21] 'Special Issue on Named Entities',
  - *Traitement Automatique des Langues*, 52(3) – Numéro spécial *Ressources Linguistiques Libres*[22] 'Special Issue on Open Language Resources'.
- I was a Program Committee member of international and national conferences and workshops:
  - 9th International Conference on Natural Language Processing *PolTAL 2014*[23], Warsaw, Poland,
  - The First Joint Conference on Lexical and Computational Semantics **SEM 2012*[24], Montreal, Canada,
  - International Conference on Language Resources and Evaluation *LREC 2012*[25], Istanbul, Turkey,
  - Workshop on Multiword Expressions *MWE-ACL 2011*[26] and *MWE-NAACL-2013*[27],
  - Computational Linguistics – Applications Conference *CL-A 2011*, Jachranka, Poland,
  - International Workshop on Balto-Slavonic Natural Language Processing *BSNLP 2007, 2009, 2011 and 2013*[28],
  - *NOOJ/Intex Workshop 2004*[29], Tours, France.

Since 2013 I serve as a European expert. I participated in the remote and central evaluation of 7 project proposals submitted to the *FP7-SME-2013* call[30]. Currently I am a reviewer of one project selected in the FP7-SME-2012 call.

In 2011 I was nominated as a member of the computer science (27th) section of the National University Council (*Conseil national des universités*, CNU), which is the central French academia evaluation agency. Three of the main CNU tasks are: (i) granting accreditations (*qualifications*) for candidates to Assistant Professor (*Maître de conférence*) and Professor positions at French

[19] http://nlp.ipipan.waw.pl/ojs/index.php/JLM/
[20] http://multiword.sourceforge.net/PHITE.php?sitesig=SPECIAL
[21] http://www.atala.org/NAMENAMED-ENTITY-NAMED-ENTITY
[22] http://www.atala.org/-Ressources-Linguistiques-Libres-
[23] http://poltal.ipipan.waw.pl/
[24] http://ixa2.si.ehu.es/starsem/
[25] http://www.lrec-conf.org/lrec2012
[26] http://multiword.sourceforge.net/mwe2011/
[27] http://multiword.sourceforge.net/mwe2013/
[28] http://nlp.pwr.wroc.pl/BSNLP11/
[29] http://tln.li.univ-tours.fr/Tln_Colloques/JIntex2004/Appel.html
[30] https://ec.europa.eu/research/participants/portal/page/capacities?callIdentifier=FP7-SME-2013

universities, (ii) granting promotions on merit to assistant professors and professors, (iii) assigning authorizations to sabbatical leaves (*Congé de Recherche et Conversion Thématique*, CRCT) on merit. I evaluated 7 promotion applications in 2012, and 35 qualification dossiers in 2013.

Finally, I was involved in 3 PhD juries: of Hyun Gue Huh[31] at the Université Marne-la-Vallée in 2005, as well as of Marc Le Tallec[32] and Mickaël Tran (cf. Section 6.6) at the Université François Rabelais Tours in 2012 and 2006.

6.8 Event Organization

My major experience in research event organization is related to the CIAA-FSMNLP[33] conference which took place in Blois in 2011. For the first time these two scientifically close communities met in a joint event: the 16th International Conference on Implementation and Application of Automata, and the 9th International Workshop on Finite State Methods and Natural Language Processing. I played the role of the Organizing Committee co-chair (with Matthieu Constant) of this event. I was in charge of most coordination tasks related to fund raising, budget planning, expenditure, internal and external communication, calls for papers, website development, logistics, accommodation and meals, support to invited speakers, invitations, social events and reporting.

This large experience is now beneficial in the PARSEME COST action management, which mainly consists in meeting organization (at a rate of 3–4 large meetings per year). In particular, I am a co-organizer of the 10th edition of the annual MWE workshop, which will be co-located with the EACL 2014 conference in Gothenburg, Sweden. It will include a special track dedicated to PARSEME topics.

6.9 Teaching and Administration

Last but not least, as an academia member, I dedicate a large part of my professional activity to teaching and associated administrative tasks. In 2002 I was recruited as an assistant professor the University Institute of Technology (*Institut Universitaire de Technologie*, IUT) in Blois, part of the Université François Rabelais Tours. Since then I have been teaching computer science in different IUT departments. I have been in charge of computer architecture, operating system, algorithmics and programing lectures, tutorials and labs at the Networking and Communications department (*Réseaux et Télécommunication*) and at the Communication Services and Networks department (*Services et Réseaux de Communication*, SRC), which deliver 2-year undergraduate technological diplomas. Since 2006 I have also been specializing in the security of operating systems within a Professional Bachelor course (*License Professionnelle*) dedicated to the Quality and Security of Information Systems (QSSI). I also supervise several student projects and work placements yearly.

I occasionally teach in Master's programs, including the Erasmus Mundus IT4BI Master's program (Information Technologies for Business Intelligence)[34] hosted at the Computer Science Department of the UFRT in Blois, jointly with the Université Libre de Bruxelles, the Universitat Politècnica de Catalunya in Barcelona, the Ecole Centrale Paris and the Technische Universität Berlin. This very selective program welcomes outstanding students from all over the world.

[31] *Délimitation et étiquetage des morphèmes en coréen par ressources linguistiques* 'Delimitation and tagging of Korean morphemes with language ressources'

[32] *Compréhension de parole et détection des émotions pour robot compagnon* 'Understanding speech and emotion mining for a companion robot'

[33] `http://ciaa-fsmnlp-2011.univ-tours.fr`

[34] http://it4bi.univ-tours.fr/

Since 2013 I am in charge of lectures and tutorials dedicated to NLP techniques for information retrieval.

As far as my administrative teaching-related activities are concerned, I was the head of international relations at the IUT Blois in 2003–2012. In this period I coordinated work placements and study periods abroad for several dozens of IUT students.

Chapter 7

General Conclusions and Perspectives

In this thesis, meant to validate my capacity of and maturity for directing research activities, I have presented a panorama of several topics in computational linguistics, linguistics and computer science.

Over the past decade, I was notably concerned with the phenomena of compositionality and variability of linguistic objects. I illustrated the advantages of a compositional approach to the language in the domain of emotion detection and I explained how some linguistic objects, most prominently multi-word expressions, defy the compositionality principles. I tried to demonstrate that the complex properties of MWEs, notably variability, are partially regular and partially idiosyncratic. This fact places the MWEs on the frontiers between different levels of linguistic processing, such as lexicon and syntax.

I have shown the highly heterogeneous nature of MWEs by citing their two existing taxonomies. After an extensive state-of-the art study of MWE description and processing, I have summarized Multiflex, a formalism and a tool for lexical high-quality morphosyntactic description of MWUs. It uses a graph-based approach in which the inflection of a MWU is expressed in function of the morphology of its components, and of morphosyntactic transformation patterns. Due to unification the inflection paradigms are represented compactly. Orthographic, inflectional and syntactic variants are treated within the same framework. The proposal is multilingual: it has been tested on six European languages of three different origins (Germanic, Romance and Slavic), I believe that many others can also be successfully covered. Multiflex proves interoperable. It adapts to different morphological language models, token boundary definitions, and underlying modules for the morphology of single words. It has been applied to the creation and enrichment of linguistic resources, as well as to morphosyntactic analysis and generation. It can be integrated into other NLP applications requiring the conflation of different surface realizations of the same concept.

Another chapter of my activity concerns named entities, most of which are particular types of MWEs. Their rich semantic load turned them into a hot topic in the NLP community, which is documented in my state-of-the art survey. I have presented the main assumptions, processes and results issued from large annotation tasks at two levels (for named entities and for coreference), parts of the National Corpus of Polish construction. I have also contributed to the development of both rule-based and probabilistic named entity recognition tools, and to an automated enrichment of Prolexbase, a large multilingual database of proper names, from open sources.

With respect to multi-word expressions, named entities and coreference mentions, I pay a special attention to nested structures. This problem sheds new light on the treatment of complex linguistic units in NLP. When these units start being modeled as trees (or, more generally, as acyclic graphs) rather than as flat sequences of tokens, long-distance dependencies, discontinu-

ities, overlapping and other frequent linguistic properties become easier to represent. This calls for more complex processing methods which control larger contexts than what usually happens in sequential processing. Thus, both named entity recognition and coreference resolution comes very close to parsing, and named entities or mentions with their nested structures are analogous to multi-word expressions with embedded complements.

My parallel activity concerns finite-state methods for natural language and XML processing. My main contribution in this field, co-authored with 2 colleagues, is the first full-fledged method for tree-to-language correction, and more precisely for correcting XML documents with respect to a DTD. We have also produced interesting results in incremental finite-state algorithmics, particularly relevant to data evolution contexts such as dynamic vocabularies or user updates.

Multilinguality is the leitmotif of my research. I have applied my methods to several natural languages, most importantly to Polish, Serbian, English and French. I have been among the initiators of a highly multilingual European scientific network dedicated to parsing and multi-word expressions. I have used multilingual linguistic data in experimental studies. I believe that it is particularly worthwhile to design NLP solutions taking declension-rich (e.g. Slavic) languages into account, since this leads to more universal solutions, at least as far as nominal constructions (MWUs, NEs, mentions) are concerned. For instance, when Multiflex had been developed with Polish in mind it could be applied as such to French, English, Serbian and Greek. Also, a French-Serbian collaboration led to substantial modifications in morphological modeling in Prolexbase in its early development stages. This allowed for its later application to Polish with very few adaptations of the existing model. Recall also that other researchers stress the advantages of NLP studies on highly inflected languages (cf. Section 3.2.3) since their morphology encodes much more syntactic information than is the case e.g. in English.

In this thesis I was also supposed to demonstrate my ability of playing an active role in shaping the scientific landscape, on a local, national and international scale. I described my: (i) various scientific collaborations and supervision activities, (ii) roles in over 10 regional, national and international projects, (iii) responsibilities in collective bodies such as program and organizing committees of conferences and workshops, PhD juries, and the National University Council (CNU), (iv) activity as an evaluator and a reviewer of European collaborative projects. It is up to the habilitation jury to assess these contributions, judge the maturity of the candidate, and make critical remarks and recommendations.

In the following sections I sketch scientific perspectives resulting from my experience, putting a special impact on links among various domains and communities.

7.1 Enhancing and Extending the Existing Language Resources and Tools

Hand-crafted electronic lexicons and annotated corpora belong to precious language resources crucial for automated support of linguistic studies on the one hand, and for the development of supervised language technology tools on the other hand. Efforts towards extension and enhancement of these resources should be pursued. As already mentioned in Section 3.6, we need an extended Multiflex model for lexical representation of MWEs in which dependencies with respect to external elements, as well as grouping of flexemes into lexemes, would be taken into account. Advanced automated dictionary-based and corpus-based graph prediction facilities are needed to speed up the lexicographer's work. Finally, an extension of the descriptive power to discontinuous, notably verbal, MWEs would pave the way for their recognition and parsing.

NKJP is already a reference corpus for Polish but would benefit from extensions (cf. Section 4.8). The NE annotation level might cover new NE types (products, events, quantities, etc.). A more fine-grained annotation of temporal expressions and metonymy would be needed.

On the level of mentions, a possible linking to external ontologies such as Prolexbase, WordNet or DBpedia would help build future entity linking and disambiguation tools. First steps towards this task, on the level of deep parsing trees, have already been taken by Hajnicz (2013). Finally, an extensive annotation of multi-word expressions should be addressed in a new NKJP annotation level.

Open source named entity recognition tools, such as NERF (cf. Section 4.4.2), deserve further investigation. Integration of external lexical resources in the observation schema has already been achieved but should further be tuned and optimized. Automatic lemmatization of names remains a challenge and should be addressed, possibly by exploiting both lexicons, such as SAWA, PNEG (cf. Section 3.5) or Prolexbase, and the NE lemmas available in NKJP. Another exciting perspective is to extend NERF functionalities to entity linking and disambiguation, notably via morphosyntactic data in Prolexbase possibly integrated into the Linked Open Data. NERF was conceived in a modular and relatively language-independent way and its adaptation to other languages would be an interesting challenge. Finally, NERF's ability to tag tree-like structures makes it a good candidate to address probabilistic parsing of multi-word expressions (cf. Section 7.3).

7.2 Integrating Fine-Grained Language Data into the Linked Data

In recent years the Semantic Web (SW) has been meeting natural language processing. Linked Open Data open exciting opportunities for information processing in natural language texts. Ontological representation of entities in the LOD provides, notably, a solid base for resolving the ambiguity of reference in named entity recognition, categorization, linking, disambiguation, relation detection and translation, as well as in coreference annotation. URIs can for instance partly replace dominant expressions in the coreference clusters discussed in Section 4.6.1, since they have good chances to be more informative than most text mentions.

Let us, however, moderate our enthusiasm for LOD. Obviously, the URIs are no ultimate solution to the problem of reference representation. The set of referents existing in all possible discourse universes largely exceeds what could possibly be represented by URIs, despite the impressive size of the Semantic Web. According to cognitive linguists (Fauconnier, 2003), the discourse world (mental space) is proper to each discourse, like each terminology is not only domain- but also text-dependent (Bourigault & Slodzian, 2000). Under these circumstances, a universal representation of the set of all possible "real-world" objects and concepts is probably an utopia.

If we still wish to benefit from the Semantic Web at least partly in some NLP application, e.g. to capture the *majority opinion*, as Langacker (1986) put it, we also need to solve the old problem of NE variability in texts. Indeed, NEs, especially the multi-word ones, still rarely occur in a corpus in the same surface forms as they do in ontologies and in other knowledge bases. Therefore, the efforts of bringing the SW and NLP together should be pursued.

It seems that the community of Entity Linking, for instance, is still more distant from the one of NER, coreference resolution and other NLP tasks than would be natural. According to (Hachey et al., 2013), most methods in NE disambiguation take candidate search (in a knowledge base) for granted (e.g. they rely on exact match against Wikipedia titles only) and focus on complex candidate ranking algorithms instead. Experiments show, however, that substantial enhancements can be achieved precisely in the candidate search stage from solving coreference and lexical variability issues, e.g. acronyms. As shown in the state-of-the-art survey in Section 4.2, only one current entity linking system covers a morphologically rich language with a complex declension system in nouns and adjectives (Russian). It is obvious that such languages, including Polish, will be particularly challenging to entity linking for the same reasons as they

are to NER and MWU extraction tasks. Namely, the inflectional variation alone accounts for a big percentage of NE and MWE occurrences in corpora, as opposed to their canonical citation forms in lexicons and ontologies. Recall for instance that (cf. page 110) 23% of dominant expressions occurring in Polish coreference clusters appear in inflected forms different from their base forms. None of these occurrences can be recognized via exact match techniques e.g. using Wikipedia titles.

As Hachey et al. (2013) put it, *named entities have interesting internal structures that a NE disambiguation system might want to exploit.* This brings us back to variant conflation problems (discussed in Section 2), to annotating MWUs and NEs with nested structures (Section 3.5 and 4.3), to coreference-annotated corpora (Sections 4.6), to nested NE recognition tools (Section 4.4), etc. In "the best of the two worlds", where semi-structured data aggregate the advantages of both natural language texts and structured data, such resources and tools should be leveraged for a more accurate semantic NLP. Ideally, linguistic units and features contained therein should also have their unique identifiers, referenced in such (possibly open and collaboratively created) language resources, and interconnected with identifiers of objects and concepts that they name in LOD ontologies.

This idea is being pursued by the Open Linguistics Working Group, which federates the efforts towards Linguistic Linked Open Data (LLOD)[1] (Chiarcos et al., 2012). The LLOD cloud currently contains DBpedia, YAGO, Wiktionary, verb nets, frame nets, wordnets, annotated corpora, classifications of morphological categories, etc. As far as lexical resources dedicated to NEs are concerned, notably JRC Names (Steinberger et al., 2011) is indirectly interlinked with DBpedia. More detailed morphosyntactic resources, such as those belonging to our contributions, especially for highly inflected languages, might contribute to these interlinked data.

We gain an initial understanding of a possible nature of such LOD/LLOD data linking from Prolexbase. Recall that it contains both the language-independent and the language-specific level. At the former, conceptual proper names are represented by pivots, roughly equivalent to URIs in LOD. They are attached to types and supertypes and related to other pivots, which is equivalent e.g. to DBpedia facts. At the latter, prolexemes represent canonical labels for proper names, like titles of Wikipedia articles or DBpedia entries. The added value from Prolexbase is, in particular, to structure a whole range of (inflectional, syntactic and semantic) variants of these canonical names into a linguistically sound hierarchy (aliases, derivatives, instances, etc.), rather than a flat list of equivalents (like redirects in Wikipedia). Moreover, important efforts have been made towards standardizing lexical NE terminology in Prolexbase via ISOCat[2], and defining a standard NE-oriented LMF interchange format (Bouchou & Maurel, 2008). Thus, we might say that, in a way, Prolexbase is a LOD/LLOD interface in a nutshell and it clearly deserves interlinking with LOD and LLOD simultaneously.

As far as all resources addressed in this thesis are concerned, we can already measure their maturity for LOD/LLOD interlinking by the Linked Data standards. As shown in Table 7.1, most of our resources are halfway to LOD integration as they correspond mostly to the 3-star class in the 5-star LOD classification[3]. The remaining tasks that would allow these resources to obtain the total integration stage are: (i) using URIs to identify entities, facts and features, (ii) inserting links towards other linked data.

[1] `http://linguistics.okfn.org/resources/llod/`
[2] `http://www.isocat.org/`
[3] `http://5stardata.info/`

Table 7.1: Language resources with our contribution and their maturity for Linked Open Data integration

Resource	**Description**	**License**	**Download area**	**Format**	**LOD class** (5-star scale)	**LOD class description**
SAWA	Grammatical Lexicon of Warsaw Urban Proper Names	CC-BY SA	http://zil.ipipan.waw.pl/SAWA	text	⋆⋆⋆	downloadable, machine-readable, non-proprietary format
SEJF	Grammatical Lexicon of Polish Phraseology	CC-BY SA	http://zil.ipipan.waw.pl/SEJF	text	⋆⋆⋆	
SEJFEK	Grammatical Lexicon of Polish Economic Phraseology	CC-BY SA	http://zil.ipipan.waw.pl/SEJFEK	text	⋆⋆⋆	
PNEG	Gazetteer for Polish Named Entities	2-clause BSD	http://clip.ipipan.waw.pl/Gazetteer	LMF, text	⋆⋆⋆	
PNET	Triggers for Polish Named Entities	2-clause BSD	http://zil.ipipan.waw.pl/PNET	text	⋆⋆⋆	
NKJP	National Corpus of Polish	GNU GPL v.3	http://clip.ipipan.waw.pl/LRT?action=AttachFile&do=view&target=NKJP-PodkorpusMilionowy-1.1.tgz	TEI P5	⋆⋆⋆	
PCC	Polish Coreference Corpus	CC BY v.3	http://zil.ipipan.waw.pl/PolishCoreferenceCorpus	TEI P5, MMAX, BRAT	⋆⋆⋆	
Prolexbase	Multilingual Ontology of Proper Names	CC BY-SA	http://zil.ipipan.waw.pl/Prolexbase http://www.cnrtl.fr/lexiques/prolex/	SQL, partly text, partly LMF	⋆⋆/⋆⋆⋆ /⋆⋆⋆⋆⋆	downloadable, machine-readable, partly non-proprietary format, partly linked with Wikipedia

7.3 Towards Deep Parsing of Multi-Word Expressions

Natural Language Processing applications nowadays face three essential challenges: (i) linguistic precision of methods and results (reflecting, at least partly, the richness and creativity of human language), (ii) specificities of particular languages and language families, (iii) computational efficiency in the context of large amounts of (possibly noisy) data to be processed rapidly. Seminal works such as (Sag et al., 2002) consider that one of the key problems to be overcome in order to meet all of these requirements simultaneously are multi-word expressions. As shown in Section 3.2, substantial progress in MWE understanding and processing has already been achieved in lexicographic and computational frameworks, for instance with respect to the problem of automatic extraction of MWEs from corpora and their lexical description. However, the resulting methods and tools still face serious limitations. On the one hand, they mostly concern either shallow linguistic processing (morphological analysis, shallow parsing, etc.). On the other hand, when deep, especially probabilistic, parsing is concerned, only contiguous MWEs are addressed. Since MWEs show idiosyncratic behavior at different levels, the integration of a full range of MWEs is necessary in deep processing. Moreover, morphological and syntactic specificities of different European languages, especially the highly inflected ones, call for a common multilingual framework.

Under these premises, the COST IC1207 action PARSEME (PARSing and Multi-word Expressions) has gathered a consortium of multidisciplinary experts from 27 countries, representing 26 languages and 6 dialects from 8 language families: Celtic (Gaelic), Germanic (British/American English, Danish, Dutch, German, Icelandic, Norwegian, Swedish), Finno-Ugric (Estonian, Hungarian), Hellenic (Greek), Romance (Swiss /France French, Italian, European/Brazilian Portuguese, Spanish), Semitic (Hebrew, Maltese), Slavic (Bulgarian, Croatian, Czech, Polish, Serbian, Slovak, Slovenian, Macedonian), and Turkic (Turkish). The main objectives of this network is to go beyond the state of the art in MWE understanding and processing by integrating MWEs in deep parsing and advanced applications such as machine translation. PARSEME's activity is organized in 4 working groups:

1. **Lexicon/grammar interface**. The challenges here are to: (i) better understand the linguistic properties of MWEs, in particular at the lexical and syntactic level, in different languages, (ii) enhance the usability of MWE lexicons and valence dictionaries in parsing, (iii) pave the way towards interoperability of lexicons and a reduction of their production cost. A possible contribution from previous assets would be, for instance, to adapt MWE extraction tools to the existing MWE lexicons so that automated enrichment of these resources is possible rather than extraction from scratch, and that the resulting descriptions show the internal structure of the extracted units and may thus be directly applicable to parsing.

2. **Parsing techniques for multi-word expressions**. The challenges are to: (i) better understand the potential of different linguistic frameworks (LFG, HPSG, TAG, etc.) with respect to parsing MWEs, (ii) to enhance the coverage of the existing grammars with respect to MWEs, (iii) to enhance parsing efficiency, e.g. by eliminating spurious ambiguities in MWEs, (iv) to reduce the cost of grammar production and enhance its interoperability, e.g. by offering abstract compact representation formalisms (meta-grammars) which could be compiled into different grammatical formalisms.

3. **Hybrid parsing of multi-word expressions**. We address here the problems such as: (i) the difficulty of integrating external language resources in probabilistic and hybrid parsing, (ii) going beyond contiguous MWEs by taking distant dependencies in a sentence into account, (iii) supplementing the costly and scarce annotated data by unannotated

data, whose quantity is practically unlimited, (iv) putting forward recommendations and best practices for enhancing knowledge-based parsing of MWEs with probabilistic scores.

4. **Annotating Multi-Word Expressions in Treebanks**. Treebanks are the major resources for linguistic modeling of syntax and semantics, and for training and evaluating probabilistic parsers, but MWE annotation is still in its early stage. In most cases, only contiguous MWEs are annotated and/or their deep syntactic structure is not described. The aims are to: (i) provide annotation guidelines for representing MWEs in constituency and dependency treebanks, (ii) re-annotate existing treebanks according to new needs and recent discoveries, (iii) put forward recommendations on how to use current and future treebanks to automatically extract lexicons and probability scores addressed in other working groups.

7.4 On the Cross-Roads of MWE Processing and Tree-to-Language Correction

As extensively discussed in Section 2.3 and in Chapter 3, orthographic, morphological, syntactic and semantic variability is among the major properties of MWEs. When lexical resources of MWEs are available, one of the challenges is to be able to identify their occurrences in corpora despite their variability with respect to the base forms. For contiguous MWEs, a possible solution is to generate extensional lexicons enumerating all possible variants, such as those generated by Multiflex (cf. Section 3.3). However, when a non contiguous, especially verbal, MWE is concerned, all its grammatically correct instantiations correspond to a possibly infinite (due to admitting unconstrained nominal group complements, adverbial modifiers, etc.) set of syntactic subtrees.

We think that such a set of potential variants of a MWE could be encoded in a lexicon as a tree language. A particular occurrence of this MWE in a treebank, in its turn, would be seen as a syntactic (sub)tree. In this context, identifying MWEs in a treebank could be modeled as an instance of the tree-to-language correction problem. Namely, each syntax tree fragment whose leaves satisfy the minimal lexical constraints for a particular MWE would be corrected with respect to the tree language of this MWE. With a similarity threshold equal to 0, fully grammatical occurrences only would be recognized. With a small positive threshold, partial (but not too huge) ungrammaticality would be allowed, which may help process noisy data, e.g. in spontaneous speech, social networks, etc.

7.5 Towards a Unified Approach to Tree-to-Language Correction

In (Amavi et al., 2013) we have presented a contrastive state-of-the-art study of the tree-to-language correction. This analysis shows that, despite the size of the tree-to-language correction community and the richness of the proposed techniques, there is a rather weak reproducibility of the published results due to unavailability of the implementations, documentation, source codes and experimental data. Therefore, we think that it would be interesting and valuable to perform an in-depth survey of the domain similarly to (Boytsov, 2011). The existing methods might be classified within a taxonomy, as well as implemented and tested within a common platform and on different data sets. This would allow us for direct performance comparisons and complexity, completeness and soundness proofs.

Research nowadays faces a lot of challenges stemming from its increasingly specialized nature. Scientific communities get fractioned into narrow subfields and specialities. Although huge amounts of scientific findings and achievements are accumulated, it is hard to make them visible and understandable by the whole community, thus many similar problems are being addressed by several communities, sometimes in parallel and similar solutions are proposed but described by specific domain-dependent terminologies. A researcher nowadays should make more and more efforts to be aware of the already existing approaches and results before he/she defines new problems and puts forward novel methods to solve them. I think that this evolution should further be promoted towards establishing solid links among different fields and communities.

Another property of the current research is its increasingly project-oriented funding. While this is sound for developing versatile skills in researchers, for feeding applied research and for research dissemination, it also creates many risks for the long term future. The very nature of fundamental research assumes unpredictability of results and of the time needed to obtain them. Project-oriented research activity requires, conversely, tight schedules and precise deliverables. Let us hope that a balance between these two types of scientific activity can be achieved both at the national and at the international level.

Finally, let us stress again that benefits from balancing the NLP assets with respect to language representativity. In this dissertation I tried to show that simultaneously taking different languages of different nature into account helps achieve more universal and insightful solutions. Further promotion of this language variety is one of my major perspectives.

Bibliography

Abeillé, Anne, Lionel Clément & François Toussenel. 2003. *Building a treebank for French* 165–187. Kluwer Academic Publishers.

Abeillé, Anne & Yves Schabes. 1989. Parsing idioms in lexicalized tags. In Harold L. Somers & Mary McGee Wood (eds.), *Proceedings of the 4th Conference of the European Chapter of the ACL, EACL'89, Manchester*, 1–9. The Association for Computer Linguistics. `http://dblp.uni-trier.de/db/conf/eacl/eacl1989.html#AbeilleS89`.

Abney, S. 1996. Partial Parsing via Finite-state Cascade. In *Proceedings, Workshop on Robust Parsing, 8th European Summer Schol in Logic, Language and Information*, 8–15. Prague, Czech Republic.

Abramowicz, Witold, Agata Filipowska, Jakub Piskorski, Krzysztof Wecel & Karol Wieloch. 2006. Linguistic Suite for Polish Cadastral System. In *Proceedings of the LREC'06*, 53–58. Genoa, Italy.

Acedański, Szymon. 2010. A Morphosyntactic Brill Tagger for Inflectional Languages. In Hrafn Loftsson, Eiríkur Rögnvaldsson & Sigrún Helgadóttir (eds.), *Advances in Natural Language Processing*, vol. 6233 Lecture Notes in Computer Science, 3–14. Springer.

Agafonov, Claire, Thierry Grass, Denis Maurel, Nathalie Rossi-Gensane & Agata Savary. 2006. La traduction multilingue des noms propres dans PROLEX. *Mεta* 51(4). 622–636. `http://www.erudit.org/revue/meta/2006/v51/n4/014330ar.html`. Les Presses de l'Université de Montréal.

Aho, Alfred, John Hopcroft & Jeffrey Ullman. 1980. *Strutures de données et algorithmes*. Paris: InterEditions.

Al-Haj, Hassan & Shuly Wintner. 2010. Identifying multi-word expressions by leveraging morphological and syntactic idiosyncrasy. In *Proceedings of the 23rd International Conference on Computational Linguistics* COLING '10, 10–18. Stroudsburg, PA, USA: Association for Computational Linguistics. `http://dl.acm.org/citation.cfm?id=1873781.1873783`.

Alegria, Iñaki, Olatz Ansa, Xabier Artola, Nerea Ezeiza, Koldo Gojenola & Ruben Urizar. 2004. Representation and Treatment of Multiword Expressions in Basque. In *Proceedings of the ACL'04 Workshop on Multiword Expressions*, 48–55.

Alex, Beatrice, Barry Haddow & Claire Grover. 2007. Recognising nested named entities in biomedical text. In *Bionlp '07: Proceedings of the workshop on bionlp 2007*, 65–72. Morristown, NJ, USA: Association for Computational Linguistics.

Amavi, Joshua, Béatrice Bouchou & Agata Savary. 2013. On Correcting XML Documents with Respect to a Schema. *The Computer Journal* doi:10.1093/comjnl/bxt006. `http:`

//comjnl.oxfordjournals.org/content/early/2013/02/13/comjnl.bxt006.abstract. Preprint: http://www.info.univ-tours.fr/~savary/English/papersASgb.html#TheComputerJournal.

Andreewsky, A., Fathi Debili & Christian Fluhr. 1977. Computational Learning of Semantic Lexical Relations for the Generation and Automatic Analysis of Content. In *Proceedings, IFIP Congress*, 667–73. Toronto: IFIP.

Appelt, Douglas E., Jerry R. Hobbs, John Bear, David Israel, Megumi Kameyama, Andy Kehler, David Martin, Karen Myers & Mabry Tyson. 1995. SRI international FASTUS system MUC-6 test results and analysis. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, 237–248. NIST, Morgan-Kaufmann Publishers.

Arampatzis, A. T., C. H. A. Koster & T. Tsoris. 1997. IRENA: Information retrieval engine based on natural language analysis. In *Proceedings, Intelligent Multimedia Information Retrieval Systems and Management (RIAO'97)*, 159–75. Montreal: CID, Paris.

Arampatzis, A. T., T. Tsoris, C. H. A. Koster & Th. P. van der Weide. 1998. Phrase-based Information Retrieval. *Information Processing and Management* 34(6). 693–707.

Attia, Mohammed A. 2006. Accommodating multiword expressions in an arabic LFG grammar. In *Proceedings of the 5th international conference on Advances in Natural Language Processing* FinTAL'06, 87–98. Berlin, Heidelberg: Springer-Verlag. doi:10.1007/11816508_11. http://dx.doi.org/10.1007/11816508_11.

Baggio, Giosuè, Michiel van Lambalgen & Peter Hagoort. 2012. *The processing consequences of compositionality* chap. The processing consequences of compositionality. Oxford University Press.

Baldwin, Timothy & Aline Villavicencio. 2002. Extracting the Unextractable: A Case Study on Verb-particles. In *Proceedings of the 6th Conference on Natural Language Learning (CoNLL-2002)*, 98–104.

Bański, Piotr & Adam Przepiórkowski. 2009. Stand-off TEI annotation: the case of the National Corpus of Polish. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III) at ACL-IJCNLP 2009*, 64–67. Singapore.

Becker, Markus, Witold Drożdżyński, Hans-Ulrich Krieger, Jakub Piskorski, Ulrich Schäfer & Feiyu Xu. 2002. SProUT - Shallow Processing with Typed Feature Structures and Unification. In *Proceedings of ICON 2002, Mumbay, India*, .

Beesley, Kenneth R. & Lauri Karttunen. 2003. *Finite State Morphology*. CSLI.

Bejcek, Eduard, Pavel Straňák & Daniel Zeman. 2011. Influence of Treebank Design on Representation of Multiword Expressions. In Alexander F. Gelbukh (ed.), *Computational Linguistics and Intelligent Text Processing - 12th International Conference, CICLing 2011, Tokyo, Japan, February 20-26, 2011. Proceedings, Part I*, vol. 6608 Lecture Notes in Computer Science, 1–14. Springer. doi:http://dx.doi.org/10.1007/978-3-642-19400-9_1.

Bejček, Eduard & Pavel Straňák. 2010. Annotation of multiword expressions in the Prague dependency treebank. *Language Resources and Evaluation* 44(1–2). 7–21.

Benveniste, Emile. 1974. *Fondements syntaxiques de la composition nominale. Formes nouvelles de la composition nominale* 145–176. Gallimard, Paris.

Bertino, Elisa, Giovanna Guerrini & Marco Mesiti. 2004. A Matching algorithm for measuring the structural similarity between an XML documents and a DTD and its applications. *Information Systems* 29. 23–46.

Bertino, Elisa, Giovanna Guerrini & Marco Mesiti. 2008. Measuring the structural similarity among XML documents and DTDs. *Journal of Intelligent Information Systems* 30. 55–92.

Bień, Janusz. 1991. *Koncepcja słownikowej informacji morfologicznej i jej komputerowej weryfikacji*. Warszawa: Wydawnictwa Uniwersytetu Warszawskiego.

Bizer, Christian, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak & Sebastian Hellmann. 2009. DBpedia - A crystallization point for the Web of Data. *J. Web Sem.* 7(3). 154–165.

Blanc, Olivier & Matthieu Constant. 2005. Lexicalization of Grammars with Parameterized Graphs. In *Proceedings of RANLP'05*, Borovets, Bulgaria.

Bollacker, Kurt, Patrick Tufts, Tomi Pierce & Robert Cook. 2007. A Platform for Scalable, Collaborative, Structured Information Integration. In *Proceeding of the Sixth International Workshop on Information Integration on the Web*, .

Boobna, Utsav & Michel de Rougemont. 2004. Correctors for XML Data. In *Proceedings of XSym 04, Toronto, Canada*, vol. 3186 Lecture Notes in Computer Science, 97–111. Springer.

Bouchou, Béatrice, Ahmed Cheriat, Mirian Halfeld Ferrari Alves & Agata Savary. 2006a. Integrating Correction into Incremental Validation. In *Proceeding of BDA 06, Lille, France*, .

Bouchou, Béatrice, Ahmed Cheriat, Mirian Halfeld Ferrari Alves & Agata Savary. 2006b. XML Document Correction: Incremental Approach Activated by Schema Validation. In *Proceedings of IDEAS 06, Delhi, India*, 228–238. IEEE Computer Society.

Bouchou, Béatrice, Denio Duarte, Mirian Halfeld Ferrari Alves, Dominique Laurent & Martin A. Musicante. 2004. Schema Evolution for XML: A Consistency-Preserving Approach. In *Proceedings of MFCS 04, Prague, Czech Republic*, vol. 3153 Lecture Notes in Computer Science, 876–888. Springer.

Bouchou, Béatrice & Denis Maurel. 2008. Prolexbase et LMF : vers un standard pour les ressources lexicales sur les noms propres. *TAL* 49(1). 61–88.

Bourigault, Didier. 1993. An Endogeneous Corpus-Based Method for Structural Noun Phrase Disambiguation. In *Proceedings, Sixth Conference of the European Chapter of the Association for Computational Linguistics (EACL'93)*, 81–86. Utrecht: ACL.

Bourigault, Didier. 1994. *LEXTER un Logiciel d'EXtraction de TERminologie. Application à l'extraction des connaissances à partir de textes.* Paris: École des Hautes Études en Sciences Sociales Thèse en mathématiques, informatique appliquée aux sciences de l'homme.

Bourigault, Didier. 1996. LEXTER, a Natural Language Tool for Terminology Extraction. In *Proceedings, Seventh EURALEX International Congress*, 771–79. Göteborg: EURALEX.

Bourigault, Didier & Monique Slodzian. 2000. Pour une terminologie textuelle. *Terminologies Nouvelles* 19. 29–32.

Boytsov, Leonid. 2011. Indexing Methods for Approximate Dictionary Searching: Comparative Analysis. *ACM Journal of Experimental Algorithmics* 16(1).

Brants, Thorsten. 2003. Natural Language Processing in Information Retrieval. In Bart Decadt, Véronique Hoste & Guy De Pauw (eds.), *CLIN*, vol. 111 Antwerp papers in linguistics, University of Antwerp.

Breidt, Elisabeth, Frédérique Segond & Guiseppe Valetto. 1996. Formal Description of Multi-Word Lexemes with the Finite-State Formalism IDAREX. In *Proceedings of COLING-96, Copenhagen*, 1036–1040.

Broda, Bartosz, Michał Marcińczuk, Marek Maziarz, Adam Radziszewski & Adam Wardyński. 2012. KPWr: Towards a Free Corpus of Polish. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of LREC'12*, Istanbul, Turkey: ELRA.

Burnard, Lou & Syd Bauman (eds.). 2008. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Oxford.

Böhmová, Alena, Jan Hajič, Eva Hajičová & Barbora Hladká. 2003. The Prague Dependency Treebank: Three-level annotation scenario. In Anne Abeillé (ed.), *Treebanks: Building and using parsed corpora*, vol. 20 Text, Speech and Language Technology, 103–127. Dordrecht: Kluwer.

Cadiot, Pierre. 1992. A entre deux noms : vers la composition nominale. *Lexique* 11. 193–240.

Calzolari, Nicoletta, Charles J. Fillmore, Ralph Grishman, Nancy Ide, Alessandro Lenci, Catherine MacLeod & Antonio Zampolli. 2002. Towards Best Practice for Multiword Expressions in Computational Lexicons. In *LREC'02*, 1934–1940.

Cheriat, Ahmed, Agata Savary, Béatrice Bouchou & Mirian Halfeld Ferrari Alves. 2005. Incremental string correction: Towards correction of XML documents. In Jan Holub & Milan Simánek (eds.), *Stringology*, 201–215. Department of Computer Science and Engineering, Faculty of Electrical Engineering, Czech Technical University.

Chiarcos, Christian, Sebastian Hellmann, Sebastian Nordhoff, Steven Moran, Richard Littauer, Judith Eckle-Kohler, Iryna Gurevych, Silvana Hartmann, Michael Matuschek & Christian M. Meyer. 2012. The Open Linguistics Working Group. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012), Istanbul, Turkey, May 23-25, 2012*, 3603–3610. European Language Resources Association (ELRA).

Chinchor, Nancy. 1997. MUC-7 Named Entity Task Definition. In *Proc. of muc-7*, .

Chrobot, Agata. 1998a. Flexion automatique des mots composés. *Cahiers de l'Institut de Linguistique de Louvain* 145–159.

Chrobot, Agata. 1998b. Fonctionnalités INTEX dans l'outil d'aide à la traduction : LexPro CD Databank. *Lingvisticae Investigationes* 22. 311–325.

Chrobot, Agata. 1999. Enrichissement terminologique en anglais fondé sur des dictionnaires généraux et spécialisés. *Terminologies Nouvelles* 19.

Church, Kenneth. 2011. A Pendulum Swung Too Far. *Linguistic Issues in Language Technology* 6(5). 1–27.

Cohen, Jacob. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20. 37–46.

Constant, Matthieu, Joseph Le Roux & Anthony Sigogne. 2013. Combining compound recognition and PCFG-LA parsing with word lattices and conditional random fields. *ACM Trans. Speech Lang. Process.* 10(3). 8:1–8:24. doi:10.1145/2483969.2483970. `http://doi.acm.org/10.1145/2483969.2483970`.

Constant, Matthieu, Anthony Sigogne & Patrick Watrin. 2012. Discriminative strategies to integrate multiword expression recognition and parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1* ACL '12, 204–212. Stroudsburg, PA, USA: Association for Computational Linguistics. `http://dl.acm.org/citation.cfm?id=2390524.2390554`.

Copestake, Ann, Fabre Lambeau, Aline Villavicencio, Francis Bond, Timothy Baldwin, Ivan A. Sag & Dan Flickinger. 2002. Multiword expressions: linguistic precision and reusability. In *Proceedings of LREC 2002*, .

Corbin, Danielle. 1992. Hypothèses sur les frontières de la composition nominale. *Cahiers de grammaire* 17. 26–55. Université de Toulouse Le Mirail.

Courtois, Blandine & Max Silberztein (eds.). 1990. Les dictionnaires électroniques du français. *Langue française* 87.

Czerepowicka, Monika. 2011. „Toposław" jako narzędzie znakowania jednostek wieloczłonowych. In Iza Matusiak-Kempa & Sebastian Przybyszewski (eds.), *Nowe zjawiska w języku, tekście, komunikacji. Kontekst a komunikacja*, 28–35. Olsztyn, Polska: Centrum Badań Europy Wschodniej UWM.

Czerepowicka, Monika. submitted. SEJF – Słownik elektroniczny jednostek frazeologicznych. *Język Polski* .

Czerepowicka, Monika & Iwona Kosek. 2011. Problemy opisu związków frazeologicznych w formalizmie „Multifleks" (na przykładzie rodzaju wyrażeń frazeologicznych). In Mirosław Bańko & Dorota Kopcińska (eds.), *Różne formy, różne treści*, 117–126. Warszawa, Polska: Wydział Polonistyki Uniwersytetu Warszawskiego.

Daciuk, J., S. Mihov, B. Watson & R. Watson. 2000. Incremental Construction of Minimal Acyclic Finite State Automata. *Computational Linguistics* 26(1). 3–16.

Daciuk, Jan, Denis Maurel & Agata Savary. 2005a. Dynamic Perfect Hashing with Finite-State Automata. In Mieczyslaw A. Klopotek, Slawomir T. Wierzchon & Krzysztof Trojanowski (eds.), *Intelligent Information Systems* Advances in Soft Computing, 169–178. Springer.

Daciuk, Jan, Denis Maurel & Agata Savary. 2005b. Incremental and Semi-incremental Construction of Pseudo-Minimal Automata. In Jacques Farré, Igor Litovsky & Sylvain Schmitz (eds.), *Implementation and Application of Automata, 10th International Conference, CIAA 2005, Sophia Antipolis, France, June 27-29, 2005, Revised Selected Papers*, vol. 3845 Lecture Notes in Computer Science, 341–342. Springer. Preprint: `http://www.info.univ-tours.fr/~savary/English/papersASgb.html#CIAA05`.

Daciuk, Jan & Dawid Weiss. 2011. Smaller Representation of Finite State Automata. In Béatrice Bouchou-Markhoff, Pascal Caron, Jean-Marc Champarnaud & Denis Maurel (eds.), *CIAA*, vol. 6807 Lecture Notes in Computer Science, 118–129. Springer.

Daiber, Joachim, Max Jakob, Chris Hokamp & Pablo N. Mendes. 2013. Improving Efficiency and Accuracy in Multilingual Entity Extraction. In *Proceeding of the 9th International Conference on Semantic Systems (I-SEMANTICS 2013), Graz, Austria*, .

Daille, Béatrice. 1994. *Approche mixte pour l'extraction de terminologie : Statistique lexicale et filtres linguistiques*. Paris: Université de Paris 7 Thèse en informatique fondamentale.

Daille, Béatrice. 1996. Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. In Judith L. Klavans & Philip Resnik (eds.), *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, 49–66. Cambridge: MIT Press.

Damerau, Fred J. 1964. A technique for computer detection and correction of spelling errors. *Communications of the ACM* 7(3). 171–176. `http://doi.acm.org/10.1145/363958.363994`.

David, Sophie & Pierre Plante. 1990a. De la nécessité d'une approche morpho-syntaxique dans l'analyse de textes. *Intelligence Artificielle et Sciences Cognitives au Québec* 3(3). 140–54.

David, Sophie & Pierre Plante. 1990b. Le progiciel TERMINO : de la nécessité d'une analyse morphosyntaxique pour le dépouillement terminologique des textes. In *Colloque International sur les Industries de la Langue: Perspectives des Années 1990*, 71–88. Montréal: Office de la Langue Fran caise et Société des Traducteurs du Quebec.

David Barnard and Gwen Clarke and Nicholas Duncan. 1995. Tree-to-tree Correction for Document Trees. Tech. Rep. 95-372 Department of Computing and Information Science, Queen's University Kingston, Ontario.

Davis, Anthony R. & Leslie Barrett. 2013. Lexical semantic factors in the acceptability of english support-verb-nominalization constructions. *ACM Trans. Speech Lang. Process.* 10(2). 5:1–5:15. doi:10.1145/2483691.2483694. `http://doi.acm.org/10.1145/2483691.2483694`.

Delpech, Estelle, Béatrice Daille, Emmanuel Morin & Claire Lemaire. 2012. Extraction of domain-specific bilingual lexicon from comparable corpora: compositional translation and ranking. In *Proceedings of the 24th International Conference on Computational Linguistics*, à paraître. Mumbai, Inde. `http://hal.archives-ouvertes.fr/hal-00743807`. ANR-08-CORD-013 FP7/2007-2013, Grant Agreement no 248005.

Derwojedowa, M. & M. Rudolf. 2003. Czy Burkina to dziewczyna i co o tym sądzą ich królewskie mości, czyli o jednostkach leksykalnych pewnego typu. *Poradnik językowy* 5. 39–49.

Desmet, Bart & Véronique Hoste. 2010. Towards a Balanced Named Entity Corpus for Dutch. In *Proceedings of LREC'10*, Valletta, Malta.

Dillon, Martin & Ann S. Gray. 1983. FASIT : A Fully Automatic Syntactically Based Indexing System. *Journal of the American Society for Information Science* 34(2). 99–108.

Dinarelli, Marco & Sophie Rosset. 2012. Tree representations in probabilistic models for extended named entities detection. In *Eacl '12*, 174–184. Stroudsburg, PA, USA: Association for Computational Linguistics. `http://dl.acm.org/citation.cfm?id=2380816.2380840`.

Doddington, George R., Alexis Mitchell, Mark A. Przybocki, Lance A. Ramshaw, Stephanie Strassel & Ralph M. Weischedel. 2004. The Automatic Content Extraction (ACE) Program - Tasks, Data, and Evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004, May 26-28, 2004, Lisbon, Portugal*, European Language Resources Association.

Downing, Pamela. 1977. On the Creation and Use of English Compound Nouns. In *Proceedings of CICLING-2002*, vol. 53 4, 810–842. Linguistic Society of America.

Drożdżyński, Witold, Hans-Ulrich Krieger, Jakub Piskorski, Ulrich Schäfer & Feiyu Xu. 2004. Shallow Processing with Unification and Typed Feature Structures – Foundations and Applications. *Künstliche Intelligenz* 1/04.

Du, M. W. & S. C. Chang. 1992. A model and a fast algorithm for multiple errors spelling correction. *Acta Informatica* 29. 281–302.

Ehrmann, Maud. 2008. Les entitées nommées, de la linguistique au TAL : Statut théorique et méthodes de désambiguïsation. PhD Thesis. Université Paris 7.

Enguehard, Chantal & Laurent Pantera. 1995. Automatic Natural Acquisition of a terminology. *Journal of Quantitative Linguistics* 2(1). 27–32.

Evans, David A., Kimberly Ginther-Webster, Mary Hart, Robert G. Lefferts & Ira A. Monarch. 1991. Automatic Indexing Using Selective NLP and First-Order Thesauri. In *Proceedings, Intelligent Multimedia Information Retrieval Systems and Management (RIAO'91)*, 624–43. Barcelona: CID, Paris.

Fabre, Cecile & Pascale Sébillot. 1996. Interprétation automatique des composés nominaux anglais hors domaine: quelles solutions. In *10e Congrès Reconnaissance des Formes et Inelligence Artificielle (RFIA'96), Rennes*, 71–79.

Fagan, Joel L. 1987. Automatic Phrase Indexing for Document Retrieval: An Examination of Syntactic and Non-Syntactic Methods. In *Proceedings, Tenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'87)*, 91–101. ACM.

Farmakiotou, Dimitra, Vangelis Karkaletsis, John Koutsias, George Sigletos, Constantine D. Spyropoulos & Panagiotis Stamatopoulos. 2000. Rule-Based Named Entity Recognition For Greek Financial Texts. In *In Proceedings of the Workshop on Computational Lexicography and Multimedia Dictionaries (COMLEX 2000)*, 75–78.

Fauconnier, Gilles. 2003. *Encyclopedia of cognitive science* chap. Cognitive Linguistics. London: Macmillan.

Fernando, Samuel & Mark Stevenson. 2012. Mapping WordNet synsets to Wikipedia articles. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey.

Finkel, Jenny Rose & Christopher D. Manning. 2009a. Joint Parsing and Named Entity Recognition. In *Hlt-naacl*, 326–334. The Association for Computational Linguistics.

Finkel, Jenny Rose & Christopher D. Manning. 2009b. Joint Parsing and Named Entity Recognition. In *Proceedings of NAACL-2009*, 326–334. Boulder, Colorado, USA: Association for Computational Linguistics.

Finkel, Jenny Rose & Christopher D. Manning. 2009c. Nested Named Entity Recognition. In *Proceedings of EMNLP-2009*, 141–150. Singapore: Association for Computational Linguistics.

Fort, Karën, Maud Ehrmann & Adeline Nazarenko. 2009. Vers une méthodologie d'annotation des entités nommées en corpus ? In *Actes de la 16ème Conférence sur le Traitement Automatique des Langues Naturelles 2009*, Senlis, France. `http://hal.archives-ouvertes.fr/hal-00402321`. Quaero.

Fort, Karën & Benoît Sagot. 2010. Influence of pre-annotation on POS-tagged corpus development. In *Proceedings of the Fourth Linguistic Annotation Workshop* LAW IV '10, 56–63. Stroudsburg, PA, USA: Association for Computational Linguistics. `http://dl.acm.org/citation.cfm?id=1868720.1868727`.

Foufi, Vassiliki. 2013. Les noms composés A(A)N du Grec Moderne et leurs variantes. In Fryni Kakoyianni Doa (ed.), *Penser le lexique-grammaire : perspectives actuelles*, Paris, France: Editions Honoré Champion.

Freitas, Claudia, Cristina Mota, Diana Santos, Hugo Goncalo Oliveira & Paula Carvalho. 2010. Second harem: Advancing the state of the art of named entity recognition in portuguese. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, 3630–3637. ELRA.

Friburger, N. & D. Maurel. 2001. Finite-State Transducer Cascade to Extract Proper Nouns in Texts. In *Proceedings, 6th Conference on Implementations and Applications of Automata*, 97–106. Pretoria, South Africa.

Friburger, Nathalie & Denis Maurel. 2004. Finite-state transducer cascade to extract named entities in texts. *Theoretical Computer Science* 313. 94–104.

Gaál, T. 2001. Is this finite-state transducer sequentiable? In *Proceedings, 6th Conference on Implementations and Applications of Automata*, 107–115. Pretoria, South Africa.

Gaizauskas, Robert, T. Wakao, Kevin Humphreys, Hamish Cunningham & York Wilks. 1995. University of Sheffield: Description of the LaSIE System as Used for MUC-6. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, 207–220. NIST, Morgan-Kaufmann Publishers.

Galicia-Haro, Sofía N. & Alexander Gelbukh. 2009. Complex named entities in Spanish texts. In Satoshi Sekine & Elisabete Ranchhod (eds.), *Named Entities. Recognition, classification and use*, 71–96. John Benjamins.

Galliano, Sylvain, Guillaume Gravier & Laura Chaubard. 2009. The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts. In *Proceedings of Interspeech 2009*, 2583–2586. International Speech Communication Association (ISCA).

Głowińska, Katarzyna & Adam Przepiórkowski. 2010. The Design of Syntactic Annotation Levels in the National Corpus of Polish. In *Proceedings of LREC 2010, Malta*, .

Graliński, Filip, Krzysztof Jassem & Michał Marcińczuk. 2009a. An Environment for Named Entity Recognition and Translation. In *Proceedings of the 13th Annual Meeting of the European Association for Machine Translation (EAMT'09)*, 88–96. Barcelona.

Graliński, Filip, Krzysztof Jassem, Michał Marcińczuk & Paweł Wawrzyniak. 2009b. Named Entity Recognition in Machine Anonymization. In *Recent Advances in Intelligent Information Systems*, 247–260. Warsaw: Exit.

Graliński, Filip, Agata Savary, Monika Czerepowicka & Filip Makowiecki. 2010. Computational Lexicography of Multi-Word Units: How Efficient Can It Be? In *Proceedings of the COLING-MWE'10 Workshop, Beijing, China*, 2–10.

Gravier, Guillaume, Gilles Adda, Niklas Paulson, Matthieu Carré, Aude Giraudel & Olivier Galibert. 2012. The ETAPE corpus for the evaluation of speech-based TV content processing in the French language. In *International Conference on Language Resources, Evaluation and Corpora*, na. Turquie. `http://hal.archives-ouvertes.fr/hal-00712591`.

Green, Spence, Marie-Catherine de Marneffe, John Bauer & Christopher D. Manning. 2011. Multiword Expression Identification with Tree Substitution Grammars: A Parsing tour de force with French. In *EMNLP*, 725–735. ACL.

Green, Spence, Marie-Catherine de Marneffe & Christopher D. Manning. 2013. Parsing Models for Identifying Multiword Expressions. *Computational Linguistics* 39(1). 195–227.

Grégoire, Nicole. 2010. DuELME: a Dutch electronic lexicon of multiword expressions. *Language Resources and Evaluation* 44(1-2).

Gross, Gaston. 1988. Degré de figement des noms composés. *Langages* 90. 57–71. Paris : Larousse.

Gross, Gaston. 1990. Définition des noms composés dans un lexique-grammaire. *Langue Française* 87.

Gross, Gaston. 1996. *Les expressions figées en français. noms composés et autres locutions.* Paris: Ophrys.

Gross, Maurice & Jean Senellart. 1998. Nouvelles bases statistiques pour les mots du français. In *Proceedings of JADT'98, Nice 1998*, 335–349.

Habert, Benoît & Christian Jacquemin. 1993. Noms composés, termes, dénominations complexes : Problématiques linguistiques et traitements automatiques. *Traitement automatique des langues* 34(2). 5–42.

Hachey, Ben, Will Radford, Joel Nothman, Matthew Honnibal & James R. Curran. 2013. Evaluating Entity Linking with Wikipedia. *Artif. Intell.* 194. 130–150. doi:10.1016/j.artint.2012.04.005. `http://dx.doi.org/10.1016/j.artint.2012.04.005`.

Haghighi, Aria & Dan Klein. 2009. Simple coreference resolution with rich syntactic and semantic features. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3* EMNLP '09, 1152–1161. Stroudsburg, PA, USA: Association for Computational Linguistics. `http://dl.acm.org/citation.cfm?id=1699648.1699661`.

Hajnicz, Elzbieta. 2013. Mapping Named Entities from NKJP Corpus to Skadnica Treebank and Polish Wordnet. In Mieczyslaw A. Klopotek, Jacek Koronacki, Malgorzata Marciniak, Agnieszka Mykowiecka & Slawomir T. Wierzchon (eds.), *Language Processing and Intelligent Information Systems - 20th International Conference, IIS 2013, Warsaw, Poland, June 17-18, 2013. Proceedings*, vol. 7912 Lecture Notes in Computer Science, 92–105. Springer.

Hall, Patrick A. & Geoff R. Dowling. 1980. Approximate String Matching. *Computing Surveys* 12(4). 381–402.

Hendrickx, Iris, Gosse Bouma, Walter Daelemans, Veronique Hoste, Geert Kloosterman, Anne-Marie Mineur, Joeri Van, Der Vloet & Jean-Luc Verschelde. 2008. A Coreference Corpus and Resolution System for Dutch. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, 144–149. Marrakech, Morocco: European Language Resources Association (ELRA).

Hinrichs, Erhard, Sandra Kübler, Karin Naumann & Heike Zinsmeister. 2005a. Recent Developments in Linguistic Annotations of the TüBa-D/Z Treebank. In *27th Annual Meeting of the German Linguistic Association*, Cologne, Germany.

Hinrichs, Erhard W., Sandra Kübler & Karin Naumann. 2005b. A Unified Representation for Morphological, Syntactic, Semantic, and Referential Annotations. In *Proceedings of the ACL Workshop on Frontiers In Corpus Annotation II: Pie In The Sky*, 13–20. Ann Arbor, Michigan, USA.

Hobbs, Jerry R., Douglas Appelt, John Bear, David Israel, Megumi Kameyama, Mark Stickel & Mabry Tyson. 1997. FASTUS: A Cascaded Finite-State Transducer for Extracting Information from Natural-Language Text. In Emmanuel Roche & Yves Schabes (eds.), *Finite-State Language Processing*, 383–406. Cambridge: MIT Press.

Hoffart, Johannes, Fabian Suchanek, Klaus Berberich & Gerhard Weikum. 2013. YAGO2: A Spatially and Temporally Enhanced Knowledge Base from Wikipedia. *Artificial Intelligence* 194. 28–61.

Hoffart, Johannes, Fabian M. Suchanek, Klaus Berberich, Edwin Lewis-Kelham, Gerard de Melo & Gerhard Weikum. 2011. YAGO2: exploring and querying world knowledge in time, space, context, and many languages. In *Proceedings of the 20th International Conference on World Wide Web, WWW 2011, Hyderabad, India, March 28 - April 1, 2011 (Companion Volume)*, 229–232. ACM.

Hopcroft, John E. 1971. An $n \log n$ algorithm for minimizing the states of in a finite automaton. In Z. Kohavi & A. Paz (eds.), *The Theory of Machines and Computations*, 189–96. New York: Academic Press.

Hopcroft, John E. & Jeffrey D. Ullman. 1979. *Introduction to Automata Theory, Languages, and Computation*. Reading: Addison-Wesley.

Huang, Chu-Ren & Dan Jurafsky (eds.). 2010. *COLING 2010, 23rd International Conference on Computational Linguistics, Posters Volume, 23-27 August 2010, Beijing, China*. Chinese Information Processing Society of China.

Iida, Ryu, Mamoru Komachi, Kentaro Inui & Yuji Matsumoto. 2007. Annotating a Japanese Text Corpus with Predicate-Argument and Coreference Relations. In *Proceedings of the Linguistic Annotation Workshop (LAW 2007)*, 132–139. Stroudsburg, PA, USA: Association for Computational Linguistics.

ISO/TC 37/SC 4. 2007. Language resource managment-Lexical markup framework (LMF), ISO DIS 24613:2007. http://lirics.loria.fr/documents.html.

Itai, Alon & Shuly Wintner. 2013. Hebrew Multiword Expressions. Lexical Representation and Morphological Processing. http://typo.uni-konstanz.de/parseme/images/Meeting/2013-09-16-Warsaw-meeting/WG1-Itai-Wintner.pdf.

Jacquemin, Christian. 2001. *Spotting and Discovering Terms through Natural Language Processing.* MIT Press.

Janssen, Theo M. V. 2001. Frege, Contextuality and Compositionality. *Journal of Logic, Language and Information* 10(1). 115–136.

Jassem, Krzysztof. 1996. Elektroniczny słownik dwujęzyczny w automatycznym tłumaczeniu tekstu. PhD thesis. Uniwersytet Adama Mickiewicza. Poznań.

Jassem, Krzysztof. 2004. Applying Oxford-PWN English-Polish dictionary to Machine Translation. In *Proceedings of 9th European Association for Machine Translation Workshop, "Broadening horizons of machine translation and its applications", Malta, April*, 98–105.

Jurafsky, D. & J.H. Martin. 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* Prentice Hall series in artificial intelligence. Pearson Education. `http://books.google.fr/books?id=fZmj5UNK8AQC`.

Justeson, John S. & Slava M. Katz. 1995. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering* 1(1). 9–27.

Kaalep, Heiki-Jaan & Kadri Muischnek. 2008. Multi-Word Verbs of Estonian: a Database and a Corpus. In *Proceedings of the Language Resources and Evaluation Conference (LREC). Linguistic Annotation Workshop*, 48–51. Marrakech, Maroc.

Kaplan, R. & M. Kay. 1994. Regular Models of Phonological Rule Systems. *Computational Linguistics* 20(3).

Karttunen, Lauri. 1993. Finite-State Lexicon Compiler. Tech. Rep. ISTL-NLTT2993-04-02 Xerox PARC.

Karttunen, Lauri, Ronald M. Kaplan & Annie Zaenen. 1992. Two-Level Morphology with Composition. In *Proceedings of the 14 th International Conference on Computational Linguistics (COLING'92), Nantes*, 141–148.

Klebanov, Beata Beigman, Jill Burstein & Nitin Madnani. 2013. Sentiment profiles of multiword expressions in test-taker essays: The case of noun-noun compounds. *ACM Trans. Speech Lang. Process.* 10(3). 12:1–12:15. doi:10.1145/2483969.2483974. `http://doi.acm.org/10.1145/2483969.2483974`.

Kornai, A. 1999. *Extended Finite State Models of Language.* Cambridge, UK: Cambridge University Press.

Korzen, Iorn & Matthias Buch-Kromann. 2011. Anaphoric relations in the Copenhagen Dependency Treebanks. In *Proceedings of DGfS Workshop*, 83–98. Göttingen, Germany.

Koskenniemi, Kimmo. 1983. *Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production.* Helsinki: University of Helsinki PhD dissertation.

Kracht, Marcus. 2007. Compositionality: The very idea. *Research on Language and Computation* 5(3). 287–308. doi:10.1007/s11168-007-9031-5. `http://dx.doi.org/10.1007/s11168-007-9031-5`.

Kravalová, Jana & Zdeněk Žabokrtský. 2009. Czech named entity corpus and SVM-based recognizer. In *Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration* NEWS '09, 194–201. Stroudsburg, PA, USA: Association for Computational Linguistics.

Krstev, Cvetana, Jelena Jaćimović & Duško Vitas. 2012. Recognition and normalization of some classes of named entities in Serbian. In *Proceedings of the Fifth Balkan Conference in Informatics* BCI '12, 52–57. New York, NY, USA: ACM. doi:10.1145/2371316.2371327. http://doi.acm.org/10.1145/2371316.2371327.

Krstev, Cvetana, Ivan Obradović, Ranka Stanković & Duško Vitas. 2013. An Approach to Efficient Processing of Multi-word Units. In Adam Przepiórkowski, Maciej Piasecki, Krzysztof Jassem & Piotr W. Fuglewicz (eds.), *Computational linguistics: Applications*, 109–129. Berlin: Springer-Verlag. http://www.springer.com/engineering/computational+intelligence+and+complexity/book/978-3-642-34398-8.

Krstev, Cvetana, Ranka Stanković, Ivan Obradović, Duško Vitaš & Milos Utvic. 2010. Automatic Construction of a Morphological Dictionary of Multi-Word Units. *LNAI* 6233. 226–237.

Krstev, Cvetana, Ranka Stanković, Duško Vitas & Ivan Obradović. 2006a. WS4LR: A Workstation for Lexical Resources. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006), Genoa, Italy*, 1692–1697.

Krstev, Cvetana, Dusko Vitas, Ivan Obradovic & Milos Utvic. 2011. E-Dictionaries and Finite-State Automata for the Recognition of Named Entities. In Matthieu Constant, Andreas Maletti & Agata Savary (eds.), *Fsmnlp* ACL Anthology, 48–56. Association for Computational Linguistics.

Krstev, Cvetana, Duško Vitas, Denis Maurel & Mickaël Tran. 2005. Multilingual Ontology of Proper Names. In *Proceedings of Language and Technology Conference (LTC'05), Poznań, Poland*, 116–119. Wydawnictwo Poznańskie.

Krstev, Cvetana, Duško Vitas & Agata Savary. 2006b. Prerequisites for a Comprehensive Dictionary of Serbian Compounds. *LNCS* 4139. 552–563.

Kubiak-Sokół, Aleksandra & Marek Łaziński (eds.). 2007. *Słownik nazw miejscowości i mieszkańców*. Warszawa: Wydawnictwo Naukowe PWN.

Kyriacopoulou, Tita, Safia Mrabti & Anastasia Yannacopoulou. 2002. Le dictionnaire électronique des noms composés en grec moderne. *Lingvisticæ Investigationes* 25(1). 7–28.

Langacker, Ronald W. 1986. An Introduction to Cognitive Grammar. *Cognitive Science* 10. 1–40.

Laporte, E. 1997. Rational Transductions for Phonetic Conversion and Phonology. In Emmanuel Roche & Yves Schabes (eds.), *Finite-State Language Processing*, Cambridge: MIT Press.

Laporte, Eric, Takuya Nakamura & Stavroula Voyatzi. 2008a. A French Corpus Annotated for Multiword Expressions with Adverbial Function. In *Proceedings of the Language Resources and Evaluation Conference (LREC). Linguistic Annotation Workshop*, 48–51. Marrakech, Maroc.

Laporte, Eric, Takuya Nakamura & Stavroula Voyatzi. 2008b. A French Corpus Annotated for Multiword Expressions with Adverbial Function. In *Proceedings of the Language Resources and Evaluation Conference (LREC). Linguistic Annotation Workshop*, 48–51. Marrakech, Maroc.

Levenshtein, V. 1966. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. *Soviet Physics-Doklady* 10(8). 707–710.

Levi, Judith N. 1978. *The Syntax and Semantics of Complex Nominals.* Academic Press, New York-London.

Linguistic-Data-Consortium. 2006. ACE (Automatic Content Extraction) Spanish Annotation Guidelines for Entities. Available at `http://projects.ldc.upenn.edu/ace/docs/Spanish-Entities-Guidelines_v1.6.pdf` (accessed on Feb. 18, 2013).

Lowrance, Robert & Roy Wagner. 1975. An extension of the string-to-string correction problem. *Journal of the ACM* 2(22). 177–183.

Lubaszewski, Wiesław. 2007. Information extraction tools for Polish text. In *Proceedings of 3rd Language and Technology Conference, Poznań, Poland*, Wydawnictwo Poznanskie, Fundacja Uniwersytetu im. A. Mickiewicza.

Lubaszewski, Wiesław. 2009. *Słowniki komputerowe i automatyczna ekstrakcja informacji z tekstu*. Kraków: AGH Uczelniane Wydawnictwa Naukowo-Dydaktyczne.

Lyons, John. 1978. *Sémantique linguistique*. Cambridge University Press 1990th edn.

Marcińczuk, Michal & Maciej Piasecki. 2007. Pattern Extraction for Event Recognition in the Reports of Polish Stockholders. In *Proceedings of IMCSIT–AAIA'07, Wisła, Poland*, 275–284.

Marcińczuk, Michał & Maciej Piasecki. 2010. Named Entity Recognition in the Domain of Polish Stock Exchange Reports. In *Proceedings of Intelligent Information Systems 2010, Siedlce, Poland*, 127–140.

Marciniak, Małgorzata, Joanna Rabiega-Wiśniewska, Agata Savary, Marcin Woliński & Celina Heliasz. 2009a. Constructing an Electronic Dictionary of Polish Urban Proper Names. In *Recent Advances in Intelligent Information Systems*, 233–246. Exit.

Marciniak, Malgorzata, Agata Savary, Piotr Sikora & Marcin Wolinski. 2009b. Toposaw - A Lexicographic Framework for Multi-word Units. In Zygmunt Vetulani (ed.), *Ltc*, vol. 6562 Lecture Notes in Computer Science, 139–150. Springer.

Marcińczuk, Michal & Maciej Piasecki. 2011. Statistical Proper Name Recognition in Polish Economic Texts. *Control and Cybernetics* .

Marcińczuk, Michał & Jan Kocoń. 2013. Recognition of Named Entities Boundaries in Polish Texts. In *Proceedings of the Workshop on Balto-Slavonic NLP 2013, ACL 2013, Sofia, Bulgaria)*, Association for Computational Linguistics.

Marcińczuk, Michał, Jan Kocoń & Maciej Janicki. 2013. Liner2 – A Customizable Framework for Proper Names Recognition for Polish. In Robert Bembenik, Lukasz Skonieczny, Henryk Rybinski, Marzena Kryszkiewicz & Marek Niezgodka (eds.), *Intelligent tools for building a scientific information platform*, vol. 467 Studies in Computational Intelligence, 231–253. Springer Berlin Heidelberg. doi:10.1007/978-3-642-35647-6_17. `http://dx.doi.org/10.1007/978-3-642-35647-6_17`.

Marcus, Mitchell P., Mary Ann Marcinkiewicz & Beatrice Santorini. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics* 19(2). 313–330. `http://dl.acm.org/citation.cfm?id=972470.972475`.

Maurel, Denis. 2008. Prolexbase. A multilingual relational lexical database of proper names. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2008), Marrakech, Marocco*, 334–338.

Maurel, Denis, Nathalie Friburger, Jean-Yves Antoine, Iris Eshkol-Taravella & Damien Nouvel. 2011. Cascades de transducteurs autour de la reconnaissance des entités nommées. *Traitement Automatiques des Langues* 52(1). 69–96.

Mazur, Paweł & Robert Dale. 2007. Handling conjunctions in named entities. *Linguisticae Investigationes* 51–70.

Melishar, B. & J. Skryja. 2001. On the size of deterministic finite automata. In *Proceedings, 6th Conference on Implementations and Applications of Automata*, 203–216. Pretoria, South Africa.

de Melo, Gerard & Gerhard Weikum. 2010. MENTA: inducing multilingual taxonomies from wikipedia. In *Proceedings of the 19th ACM Conference on Information and Knowledge Management, CIKM 2010, Toronto, Ontario, Canada, October 26-30, 2010*, 1099–1108. ACM.

Mel'čuk, Igor. 2010. La phraséologie en langue, en dictionnaire et en TALN. In *Actes de la 17ème Conférence sur le Traitement Automatique des Langues Naturelles 2010*, Montréal, Canada.

Mendes, Pablo, Max Jakob & Christian Bizer. 2012. DBpedia: A Multilingual Cross-domain Knowledge Base. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey: European Language Resources Association (ELRA).

Metzler, Douglas P. & Stephanie W. Haas. 1989. The Constituent Object Parser: Syntactic structure matching for Information Retrieval. *ACM Transactions on Information Systems* 7(3). 292–316.

Metzler, Douglas P., Stephanie W. Haas, Cynthia L. Cosic & Charlotte A. Weise. 1990. Conjunction Ellipsis, and Other Discontinuous Constituents in the Constituent Object Parser. *Information Processing and Management* 26(1). 53–71.

Metzler, Douglas P., Stephanie W. Haas, Cynthia L. Cosic & Leslie H. Wheeler. 1989. Constituent Object Parsing for Information Retrieval and similar text processing problems. *Journal of the American Society for Information Science* 40(6). 398–423.

Mikheev, Andrei, Marc Moens & Claire Grover. 1999. Named entity recognition without gazetteers. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics* EACL '99, 1–8. Stroudsburg, PA, USA: Association for Computational Linguistics.

Mohri, Mehryar. 1994. Compact Representations by Finite-State Transducers. In *Proceedings, 32nd Annual Meeting of the Association for Computational Linguistics (ACL'94)*, 204–08. Las Cruces, NM: ACL.

Morin, Emmanuel & Béatrice Daille. 2010. Compositionality and lexical alignment of multi-word terms. *Language Resources and Evaluation* 44(1-2). 79–95.

Muzerelle, Judith, Anaïs Lefeuvre, Jean-Yves Antoine, Emmanuel Schang, Denis Maurel, Jeanne Villaneau & Iris Eshkol. 2013. ANCOR, premier corpus de français parlé d'envergure annoté en coréférence et distribué librement. In *Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)*, 555–563. Les Sables d'Olonne, France.

Mykowiecka, A., M. Marciniak & J. Rabiega-Wiśniewska. 2008. Proper Names in Polish Dialogs. In *Proceedings of the IIS 2008 Workshop on Spoken Language Understanding and Dialogue Systems*, Zakopane, Poland: Springer Verlag.

Müller, Christoph & Michael Strube. 2006. Multi-level annotation of linguistic data with MMAX2. In *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, .

Nadeau, David & Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Linguisticae Investigationes* 30(1). 3–26.

Nedoluzhko, Anna, Jiří Mírovský, Radek Ocelák & Jiří Pergler. 2009. Extended Coreferential Relations and Bridging Anaphora in the Prague Dependency Treebank. In *Proceedings of the 7th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2009)*, 1–16. AU-KBC Research Centre, Anna University, Chennai Goa, India: AU-KBC Research Centre, Anna University, Chennai.

Neviarouskaya, Alena, Helmut Prendinger & Mitsuru Ishizuka. 2010. Recognition of affect, judgment, and appreciation in text. In *Proceedings of the 23rd International Conference on Computational Linguistics* COLING '10, 806–814. Stroudsburg, PA, USA: Association for Computational Linguistics. `http://dl.acm.org/citation.cfm?id=1873781.1873872`.

Nguyen, Hien Thang & Tru Hoang Cao. 2010. Enriching Ontologies for Named Entity Disambiguation. In *Proceedings of the 4th International Conference on Advances in Semantic Processing (SEMAPRO 2010)*, Florence, Italy.

Nivre, Joakim & Jens Nilsson. 2004. Multiword units in syntactic parsing. In *Proceedings of MEMURA 2004 – Methodologies and Evaluation of Multiword Units in Real-World Applications, Workshop at LREC 2004, May 25, 2004, Lisbon, Portugal*, 39–46. Lisbon, Portugal.

Nothman, Joel, Nicky Ringland, Will Radford, Tara Murphy & James R. Curran. 2013. Learning multilingual named entity recognition from Wikipedia. *Artificial Intelligence* 194(0). 151 – 175. doi:10.1016/j.artint.2012.03.006. `http://www.sciencedirect.com/science/article/pii/S0004370212000276`. Artificial Intelligence, Wikipedia and Semi-Structured Resources.

Nouvel, Damien, Jean-Yves Antoine, Nathalie Friburger & Denis Maurel. 2010. An Analysis of the Performances of the CasEN Named Entities Recognition System in the Ester2 Evaluation Campaign. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner & Daniel Tapias (eds.), *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, 523–529. ELRA.

Nouvel, Damien, Jean-Yves Antoine, Nathalie Friburger & Arnaud Soulet. 2013. Fouille de règles d'annotation partielles pour la reconnaissance des entités nommées. In *Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)*, 421–434. Les Sables d'Olonne, France.

Oflazer, Kemal. 1996. Error-tolerant Finite-state Recognition with Applications to Morphological Analysis and Spelling Correction. *Computational Linguistics* 22(1). 73–89.

Oflazer, Kemal, Özlem Çetonoğlu & Bilge Say. 2004. Integrating Morphology with Multi-word Expression Processing in Turkish. In *Second ACL Workshop on Multiword Expressions, July 2004*, 64–71.

Ogrodniczuk, Maciej, Katarzyna Głowińska, Mateusz Kopeć, Agata Savary & Magdalena Zawisławska. 2013a. Interesting Linguistic Features in Coreference Annotation of an Inflectional Language. In *14th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING-2013). Part I.*, vol. 7816 Lecture Notes in Computer Science, 394–407.

Ogrodniczuk, Maciej, Katarzyna Głowińska, Mateusz Kopeć, Agata Savary & Magdalena Zawisławska. 2013b. Interesting Linguistic Features in Coreference Annotation of an Inflectional Language. In Maosong Sun, Min Zhang, Dekang Lin & Haifeng Wang (eds.), *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, vol. 8202 Lecture Notes in Computer Science, 97–108. Springer Berlin Heidelberg. doi: 10.1007/978-3-642-41491-6_10. http://dx.doi.org/10.1007/978-3-642-41491-6_10.

Ogrodniczuk, Maciej, Katarzyna Głowińska, Mateusz Kopeć, Agata Savary & Magdalena Zawisławska. 2013c. Polish Coreference Corpus. In *the 6th language & technology conference (ltc'13)*, .

Ogrodniczuk, Maciej & Mateusz Kopeć. 2011. End-to-end coreference resolution baseline system for Polish. In Zygmunt Vetulani (ed.), *Proceedings of the 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, 167–171. Poznań, Poland.

Osenova, Petya & Sia Kolkovska. 2002. Combining the named-entity recognition task and np chunking strategy for robust pre-processing. In *Proceedings of the Workshop on Treebanks and Linguistic Theories, September 20-21 (TLT02)*, 167–182. Sozopol, Bulgaria: Bulgarian Academy of Sciences.

Osenova, Petya & Kiril Simov. 2004. BTB-TR05: BulTreeBank Stylebook. BulTreeBank Version 1.0. Tech. Rep. BTB-TR05 Linguistic Modelling Laboratory, Bulgarian Academy of Sciences Sofia, Bulgaria.

Pagin, Peter & Dag Westerståhl. 2001a. Compositionality I: Definitions and Variants. *Philosophy Compass* 5. 250—-264.

Pagin, Peter & Dag Westerståhl. 2001b. Compositionality II: Arguments and Problems. *Philosophy Compass* 5. 250—-264.

Pajas, Petr & Jan Štěpánek. 2008. Recent Advances in a Feature-Rich Framework for Treebank Annotation. In *Proceedings of COLING'08, Manchester*, .

Partee, Barbara H., Alice ter Meulen & Robert E. Wall. 1990. *Mathematical Methods in Linguistics*, vol. 30 Studies in Linguistics and Philosophy. Dordrecht: Kluwer.

Paumier, Sébastien. 2008. Unitex 2.1 User Manual.

Pecina, Pavel. 2010. Lexical association measures and collocation extraction. *Language Resources and Evaluation* 44(1-2). 137–158.

Piskorski, Jakub. 2005. Named-Entity Recognition for Polish with SProUT. In *LNCS Vol 3490: Proceedings of IMTCI 2004, Warsaw, Poland.*, .

Piskorski, Jakub, Petr Homola, Małgorzata Marciniak, Agnieszka Mykowiecka, Adam Przepiórkowski & Marcin Woliński. 2004. Information Extraction for Polish Using the SProUT Platform. In *Proceedings of International Conference on Intelligent Information Systems 2004, Zakopane, Poland*, .

Piskorski, Jakub, Marcin Sydow & Anna Kupść. 2007. Lemmatization of Polish Person Names. In *ACL 2007. Proceedings of the Workshop on Balto-Slavonic NLP 2007*, 27–34. Association for Computational Linguistics.

Piskorski, Jakub, Karol Wieloch & Marcin Sydow. 2009. On knowledge-poor methods for person name matching and lemmatization for highly inflectional languages. *Information Retrieval* 12(3). 275–299.

Poesio, Massimo & Ron Artstein. 2008. Anaphoric Annotation in the ARRAU Corpus. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco: European Language Resources Association.

Polański, Kazimierz. 1993. *Encyklopedia językoznawstwa ogólnego*. Wrocław: Ossolineum.

Pradhan, Sameer S., Lance Ramshaw, Ralph Weischedel, Jessica MacBride & Linnea Micciulla. 2007. Unrestricted Coreference: Identifying Entities and Events in OntoNotes. In *Proceedings of the First IEEE International Conference on Semantic Computing (ICSC 2007)*, 446–453. Washington, DC, USA: IEEE Computer Society.

Przepiórkowski, Adam, Mirosław Bańko, Rafał L. Górski & Barbara Lewandowska-Tomaszczyk (eds.). 2012. *Narodowy Korpus Języka Polskiego [Eng.: National Corpus of Polish]*. Warsaw: Wydawnictwo Naukowe PWN.

Przepiórkowski, Adam. 2004. *The IPI PAN Corpus, preliminary version*. Warsaw: Institute of Computer Science.

Przepiórkowski, Adam & Piotr Bański. 2009. Which XML standards for multilevel corpus annotation? In *Proceedings of the 4th Language & Technology Conference*, 245–250. Poznań, Poland.

Przepiórkowski, Adam & Marcin Woliński. 2003. A Flexemic Tagset for Polish. In *Proceedings of the Workshop on Morphological Processing of Slavic Languages, EACL 2003*, 33–40.

Przepiórkowski, Adam. 2008. *Formalizm* ♠ chap. 7. Warsaw: Akademicka Oficyna Wydawnicza EXIT. `http://nlp.ipipan.waw.pl/PPJP/`.

Przepiórkowski, Adam, Rafał L. Górski, Barbara Lewandowska-Tomaszczyk & Marek £aziński. 2008. Towards the National Corpus of Polish. In *Proceedings of LREC 2008*, Marrakech: ELRA.

Przepiórkowski, Adam, Rafał L. Górski, Marek Łaziński & Piotr Pęzik. 2010. Recent Developments in the National Corpus of Polish. In *Proceedings of LREC 2010, Valletta, Malta*, .

Rabiega-Wiśniewska, Joanna. 2006. Formalny opis derywacji w języku polskim. Rzeczowniki i przymiotniki. PhD thesis.

Ramírez-Cruz, Yunior & Aurora Pons-Porrata. 2008. Spanish Nested Named Entity Recognition Using a Syntax-Dependent Tree Traversal-Based Strategy. In Alexander F. Gelbukh & Eduardo F. Morales (eds.), *MICAI 2008: Advances in Artificial Intelligence, 7th Mexican International Conference on Artificial Intelligence, Atizapán de Zaragoza, Mexico, October 27-31, 2008, Proceedings*, vol. 5317 Lecture Notes in Computer Science, 144–154. Springer. doi:http://dx.doi.org/10.1007/978-3-540-88636-5_13.

Ramisch, Carlos, Aline Villavicencio & Christian Boitet. 2010. Web-based and combined language models: a case study on noun compound identification. In Huang & Jurafsky (2010) 1041–1049.

Ramisch, Carlos, Aline Villavicencio & Valia Kordoni. 2013a. Special issue on multiword expressions: From theory to practice and use, part 1. *ACM Transactions on Speech and Language Processing* 10(2).

Ramisch, Carlos, Aline Villavicencio & Valia Kordoni. 2013b. Special issue on multiword expressions: From theory to practice and use, part 2. *ACM Transactions on Speech and Language Processing* 10(3).

Ramshaw, Lance A. & Mitchell P. Marcus. 1995. Text Chunking using Transformation-Based Learning. *CoRR* cmp-lg/9505040.

Rayson, Paul, Scott Piao, Serge Aharoff, Stefan Evert & Bego na Villada Moirón (eds.). 2010. *Multiword expressions: hard going or plain sailing*, vol. 44 Language Resources and Evaluation. Springer.

Recasens, Marta, Eduard Hovy & M. Antònia Martí. 2011. Identity, non-identity, and near-identity: Addressing the complexity of coreference. *Lingua* 121(6).

Recasens, Marta & M. Antònia Martí. 2010. AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan. *Language Resources and Evaluation* 44(4). 315–345.

Revuz, Dominique. 1991. Dictionnaire et Lexiques. Méthodes et Algorithmes. PhD thesis.

Rizzo, Giuseppe, Raphaël Troncy, Sebastian Hellmann & Martin Bruemmer. 2012. NERD meets NIF: Lifting NLP Extraction Results to the Linked Data Cloud. In Christian Bizer, Tom Heath, Tim Berners-Lee & Michael Hausenblas (eds.), *WWW2012 Workshop on Linked Data on the Web, Lyon, France, 16 April, 2012*, vol. 937 CEUR Workshop Proceedings, CEUR-WS.org.

Roche, E. 1997. Parsing with finite state transducers. In Emmanuel Roche & Yves Schabes (eds.), *Finite-State Language Processing*, Cambridge, MA: MIT Press.

Roche, Emmanuel & Yves Schabes. 1997. Deterministic Part-of-Speech Tagging with Finite-State Transducers. In Emmanuel Roche & Yves Schabes (eds.), *Finite-State Language Processing*, 205–40. Cambridge: MIT Press.

Rosset, S., C. Grouin & P. Zweigenbaum. 2011. *Entités nommées structurées: guide d'annotation Quaero* Notes et documents LIMSI. LIMSI-Centre national de la recherche scientifique. `http://books.google.fr/books?id=il9bMwEACAAJ`.

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann Copestake & Dan Flickinger. 2002. Multiword Expresions: A Pain in the Neck for NLP. In *Proceedings of CICLING'02*, Springer.

Saravanan, K, Monojit Choudhury, Raghavendra Udupa & A Kumaran. 2012. An Empirical Study of the Occurrence and Co-Occurrence of Named Entities in Natural Language Corpora. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey: European Language Resources Association (ELRA).

de Saussure, Ferdinand. 1916. *Cours de linguistique générale*. Paris: Payot (Edition in 2005).

Savary, A. 2001a. Etude comparativee de deux outils d'acquisition de termes complexes. In *Proceedings, Conference Terminologie et Intelligence Artificielle (TIA-2001)*, INIST-CNRS, Nancy.

Savary, Agata. 2000. Recensement et description des mots composés - méthodes et applications. PhD Thesis. Université de Marne-la-Vallée.

Savary, Agata. 2001b. Typographical Nearest-Neighbor Search in a Finite-State Lexicon and Its Application to Spelling Correction. In Bruce W. Watson & Derick Wood (eds.), *CIAA*, vol. 2494 Lecture Notes in Computer Science, 251–260. Springer.

Savary, Agata. 2005. A formalism for the computational morphology of multi-word units. *Archives of Control Sciences* 15(3). 437–449. `http://acs.polsl.pl/index.php?mode=2&show=25#218`. Preprint: `http://www.info.univ-tours.fr/~savary/English/papersASgb.html#ACS`.

Savary, Agata. 2008. Computational Inflection of Multi-Word Units. A contrastive study of lexical approaches. *Linguistic Issues in Language Technology* 1(2). 1–53.

Savary, Agata. 2009. Multiflex: A Multilingual Finite-State Tool for Multi-Word Units. In Sebastian Maneth (ed.), *Implementation and Application of Automata*, vol. 5642 Lecture Notes in Computer Science, 237–240. Springer Berlin Heidelberg. doi:10.1007/978-3-642-02979-0_27. `http://dx.doi.org/10.1007/978-3-642-02979-0_27`. Preprint: `http://www.info.univ-tours.fr/~savary/English/papersASgb.html#CIAA09`.

Savary, Agata. 2012. LMF Format for Polish Named Entity Gazetteer. Tech. rep. LI-François Rabelais University of Tours, France.

Savary, Agata, Marta Chojnacka-Kuraś, Anna Wesołek, Danuta Skowrońska & Paweł Śliwiński. 2012a. Narodowy Korpus Języka Polskiego [Eng.: National Corpus of Polish]. In Przepiórkowski et al. (2012) chap. Anotacja jednostek nazewniczych, 129–167.

Savary, Agata, Blandine Courtois, M. McCarthy-Hammani, Maurice Gross & Katia Zallagui. 1999. Dictionnaire Electronique DELAC anglais : noms composés. Tech. Rep. 59 LADL, Université Paris 7, France.

Savary, Agata & Christian Jacquemin. 2003. Reducing Information Variation in Text. *Lecture Notes in Artificial Intelligence* 2705. 145–181. Springer.

Savary, Agata, Cvetana Krstev & Duško Vitas. 2007. Inflectional Non Compositionality and Variation of Compounds in French, Polish and Serbian, and Their Automatic Processing. *BULAG* 32. 73–93.

Savary, Agata, Leszek Manicki & Małgorzata Baron. 2013a. Populating a multilingual ontology of proper names from open sources. *Journal of Language Modelling* .

Savary, Agata, Leszek Manicki & Małgorzata Baron. 2013b. ProlexFeeder - Populating a Multilingual Ontology of Proper Names from Open Sources. Tech. Rep. 306 Laboratoire d'informatique, François Rabelais University of Tours, France.

Savary, Agata & Jakub Piskorski. 2010. In *Intelligent Information Systems, Siedlce, Poland*, 141–154. Warsaw, Poland: Systems Research Institute, Polish Academy of Sciences.

Savary, Agata & Jakub Piskorski. 2011. Language Resources for Named Entity Annotation in the National Corpus of Polish. *Control and Cybernetics* 40(2). 361–391. http://control.ibspan.waw.pl:3000/contents/export?filename=2011-2-09-Savary_Piskorski.pdf.

Savary, Agata, Joanna Rabiega-Wiśniewska & Marcin Woliński. 2009. Inflection of Polish Multi-Word Proper Names with Morfeusz and Multiflex. *Lecture Notes in Computer Science* 5070. 111–141.

Savary, Agata & Jakub Waszczuk. 2012. Narodowy Korpus Języka Polskiego [Eng.: National Corpus of Polish]. In Przepiórkowski et al. (2012) chap. Narzędzia do anotacji jednostek nazewniczych, 225–252.

Savary, Agata, Jakub Waszczuk & Adam Przepiórkowski. 2010. Towards the Annotation of Named Entities in the Polish National Corpus. In *Proceedings of LREC 10, Valletta, Malta*, European Language Resources Association.

Savary, Agata, Bartosz Zaborowski, Aleksandra Krawczyk-Wieczorek & Filip Makowiecki. 2012b. SEJFEK - a Lexicon and a Shallow Grammar of Polish Economic Multi-Word Units. In *Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon*, 195–214. Mumbai, India: The COLING 2012 Organizing Committee. http://www.aclweb.org/anthology/W12-5116.

Schäfer, Ulrich. 2006. OntoNERdIE—Mapping and Linking Ontologies to Named Entity Recognition and Information Extraction Resources. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-2006)*, 1756–1761. ELRA.

Schuler, William & Aravind Joshi. 2011. Tree-rewriting models of multi-word expressions. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, 25–30. Portland, Oregon, USA: Association for Computational Linguistics. http://www.aclweb.org/anthology/W11-0806.

Schwarz, Christoph. 1989. Content-Based Text Handling. *Information Processing and Management* 26(2). 219–26.

Schwarz, Christoph. 1990. Automatic Syntactic Analysis of Free Text. *Journal of the American Society for Information Science* 41(6). 408–17.

Sekine, Satoshi, Kiyoshi Sudo & Chikashi Nobata. 2002. Extended Named Entity Hierarchy. In *Proceedings of LREC'02*, Canary Island, Spain.

Selkow, S. M. 1977. The Tree-to-Tree Editing Problem. *Information Processing Letters* 6(6). 184–186.

Sheridan, Paraic & Alan F Smeaton. 1992. The Application of Morpho-syntactic Language Processing to Effective Phrase Matching. *Information Processing and Management* 28(3). 349–69.

Shoaran, Maryam & Alex Thomo. 2011. Evolving schemas for streaming XML. *Theoretical Computer Science* 412(35). 4545–4557.

Sikora, Piotr & Marcin Woliński. 2009. Toposław — a Dictionary Creation Tool. In *Recent Advances in Intelligent Information Systems*, Exit.

Silberztein, Max. 1993a. *Dictionnaires électroniques et analyse automatique de textes : Le système INTEX*. Paris: Masson.

Silberztein, Max. 1993b. Les groupes nominaux productifs et les noms composés lexicalisés. *LingvisticæInvestigationes* 17(2).

Silberztein, Max. 2005. NooJ's dictionaries. In *Proceedings of LTC'05, Poznań*, 291–295. Wydawnictwo Poznańskie.

Smadja, Frank. 1992. Xtract: An overview. *Computers and the Humanities* 26(5-6). 399–413.

Smeaton, Alan F. & Paraic Sheridan. 1991. Using Morpho-syntactic Language Analysis in Phrase Matching. In *Proceedings, Intelligent Multimedia Information Retrieval Systems and Management (RIAO'91)*, 415–29. Barcelona: CID, Paris.

Sparck Jones, Karen & John I. Tait. 1984a. Automatic Search Term Variant Generation. *Journal of Documentation* 40(1). 50–66.

Sparck Jones, Karen & John I. Tait. 1984b. Linguistically Motivated Descriptive Term Selection. In *Proceedings, Tenth International Conference on Computational Linguistics (COLING'84)*, 287–90. Stanford: ACL.

Staworko, Slawomir & Jan Chomicki. 2006. Validity-Sensitive Querying of XML Databases. In *Proceedings of EDBT 06, Munich, Germany, Revised Selected Papers*, vol. 4254 Lecture Notes in Computer Science, 164–177. Springer.

Staworko, Slawomir, Emmanuel Filiot & Jan Chomicki. 2008. Querying Regular Sets of XML Documents. In *Proceedings of LiD 08, Rome, Italy*, .

Steinberger, Ralf, Bruno Pouliquen, Mijail Alexandrov Kabadjov, Jenya Belyaeva & Erik Van der Goot. 2011. JRC-NAMES: A Freely Available, Highly Multilingual Named Entity Resource. In *Recent Advances in Natural Language Processing, RANLP 2011, 12-14 September, 2011, Hissar, Bulgaria*, 104–110.

Strzalkowski, Tomek. 1994. Robust Text Processing in Automatic Information Retrieval. In *Proceedings, Fourth Conference on Applied Natural Language Processing (ANLP'94)*, 168–73. Stuttgart: ACL.

Strzalkowski, Tomek. 1995. Natural Language Information Retrieval. *Information Processing and Management* 31(3). 397–417.

Strzalkowski, Tomek & Peter G. N. Scheyen. 1996. Evaluation of the Tagged Text Parser. In Harald Bunt & Masaru Tomita (eds.), *Recent Advances in Parsing Technology*, 201–20. Boston: Kluwer Academic Publisher.

Strzalkowski, Tomek & Barbara Vauthey. 1992. Information Retrieval Using Robust Natural Language Processing. In *Proceedings, 20th Annual Meeting of the Association for Computational Linguistics (ACL'92)*, 104–11. Newark, DE: ACL.

Suchanek, Fabian M., Gjergji Kasneci & Gerhard Weikum. 2007. YAGO: a core of semantic knowledge unifying WordNet and Wikipedia. In *WWW '07: Proceedings of the 16th International World Wide Web Conference*, 697–706. Banff, Canada.

Suzuki, Nobutaka. 2007. Finding K Optimum Edit Scripts between an XML Document and a RegularTree Grammar. In *Proceedings of EROW 07, Barcelona, Spain*, CEUR-WS.org.

Svoboda, Martin. 2010. *Processing of Incorrect XML Data.* Charles University in Prague MA thesis.

Svoboda, Martin & Irena Mlýnková. 2011. Correction of Invalid XML Documents with Respect to Single Type Tree Grammars. In *Proceedings of NDT 11, Macau, China*, vol. 136 Communications in Computer and Information Science, 179–194. Springer.

Szpakowicz, Stan, Francis Bond, Preslav Nakov & Su Nam Kim. 2013. On the semantics of noun compounds. *Natural Language Engineering* 19(3). 289–290.

Szpakowicz, Stanisław. 1986. *Formalny opis składniowy zdań polskich*. Warszawa: Wydawnictwa Uniwersytetu Warszawskiego.

Tai, Kuo-Chung. 1979. The tree-to-tree correction problem. *Journal of the Association for Computing Machinery* 26(3).

Tallec, Marc Le, Jeanne Villaneau, Jean-Yves Antoine, Agata Savary & Arielle Syssau. 2010a. Emologus - A Compositional Model of Emotion Detection Based on the Propositional Content of Spoken Utterances. In Petr Sojka, Ales Horák, Ivan Kopecek & Karel Pala (eds.), *Text, Speech and Dialogue, 13th International Conference, TSD 2010, Brno, Czech Republic, September 6-10, 2010. Proceedings*, vol. 6231 Lecture Notes in Computer Science, 361–368. Springer.

Tallec, Marc Le, Jeanne Villaneau, Jean-Yves Antoine, Agata Savary & Arielle Syssau-Vaccarella. 2009. Détection des émotions à partir du contenu linguistique d'énoncés oraux : application à un robot compagnon pour enfants fragilisés. In *Actes de la 16ème conférence sur le Traitement Automatique des Langues Naturelles (TALN'2009)*, Senlis, France.

Tallec, Marc Le, Jeanne Villaneau, Jean-Yves Antoine, Agata Savary & Arielle Syssau-Vaccarella. 2010b. Détection hors contexte des émotions à partir du contenu linguistique d'énoncés oraux : le système EmoLogus. In *Actes de la 17e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2010)*, Montréal, Canada.

Tekli, Joe, Richard Chbeir, Agma Traina & Caetano Traina. 2011. XML document-grammar comparison: related problems and applications. *Central European Journal of Computer Science* 1. 117–136.

Tekli, Joe, Richard Chbeir & Kokou Yétongnon. 2007. Structural Similarity Evaluation Between XML Documents and DTDs. In *Proceedings of WISE 07, Nancy, France*, 196–211.

Thomo, Alex, Srinivasan Venkatesh & Ying Ying Ye. 2008. Visibly Pushdown Transducers for Approximate Validation of Streaming XML. In *Proceedings of FoIKS 08, Pisa, Italy*, vol. 4932 Lecture Notes in Computer Science, 219–238. Springer.

Tommaso Caselli, Nancy Ide & Roberto Bartolini. 2008. A Bilingual Corpus of Inter-linked Events. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco: European Language Resources Association (ELRA). Http://www.lrec-conf.org/proceedings/lrec2008/.

Toral, Antonio, Sergio Ferrández, Monica Monachini & Rafael Muñoz. 2012. Web 2.0, Language Resources and standards to automatically build a multilingual Named Entity Lexicon. *Language Resources and Evaluation* 46(3). 383–419.

Toral, Antonio, Rafael Muñoz & Monica Monachini. 2008. Named Entity WordNet. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco: European Language Resources Association.

Tran, Mickaël & Denis Maurel. 2006. Prolexbase: Un dictionnaire relatonnel multilingue de noms propres. *Traitement Automatiques des Langues* 47(3). 115–139.

Tran, Mickaël, Denis Maurel & Agata Savary. 2006. Implantation d'un tri lexical respectant la particularité des noms propres. *Lingvisticæ Investigationes* 28(2). 303–323. http://dx.doi.org/10.1075/li.28.2.07tra.

Tran, Mickaël, Denis Maurel, Duško Vitas & Cvetana Krstev. 2005. A French-Serbian Web Collaborative Work on a Multilingual Dictionary of Proper Names. In *Proceedings of the 6th Workshop on Multilingual Lexical Databases (PAPILLON'05), Chiang Rai, Thailand*, .

Tsvetkov, Yulia & Shuly Wintner. 2010. Extraction of Multi-word Expressions from Small Parallel Corpora. In Huang & Jurafsky (2010) 1256–1264.

Vetulani, Zygmunt, Bogdan Walczak, Tomasz Obrębski & Grażyna Vetulani. 1998. *Jednoznaczne kodowanie fleksji rzeczownika polskiego i jego zastosowanie w słownikach elektronicznych - format POLEX*. Poznań: Wydawnictwa Naukowe Uniwersytetu Adama Mickiewicza.

Villavicencio, Aline, Francis Bond, Anna Korhonen & Diana McCarthy. 2005. Special issue on multiword expressions. *Computer Speech & Language* 19(4). http://www.sciencedirect.com/science/journal/08852308/19/4. Special issue on Multiword Expression.

Villavicencio, Aline, Ann Copestake, Benjamin Waldron & Fabre Lambeau. 2004. Lexical Encoding of MWEs. In *ACL Workshop on Multiword Expressions: Integrating Processing, July 2004*, 80–87.

Vincze, Veronika, István Nagy T. & János Zsibrita. 2013. Learning to detect english and hungarian light verb constructions. *ACM Trans. Speech Lang. Process.* 10(2). 6:1–6:25. doi: 10.1145/2483691.2483695. http://doi.acm.org/10.1145/2483691.2483695.

Voutilainen, Atro. 1993. *NPtool*, A detector of English noun phrases. In *Proceedings, Workshop on Very Large Corpora : Academic and Industrial Perspectives*, 48–57. Columbus, Ohio: ACL.

Wacholder, Nina, Yael Ravin & Misook Choi. 1997. Disambiguation of proper names in text. In *Proceedings of the fifth conference on Applied natural language processing* ANLC '97, 202–208. Stroudsburg, PA, USA: Association for Computational Linguistics.

Wagner, Robert A. & Michael J. Fisher. 1974. The String-to-String Correction Problem. *Journal of the Association for Computational Machinery* 21(1). 168–73.

Waszczuk, Jakub, Katarzyna Głowińska, Agata Savary, Adam Przepiórkowski & Michał Lenart. 2013. Annotation Tools for Syntax and Named Entities in the National Corpus of Polish. *International Journal of Data Mining, Modelling and Management* 5(2). 103–122. http://www.inderscience.com/info/inarticle.php?artid=53691". Preprint: http://www.info.univ-tours.fr/~savary/English/papersASgb.html#IJDMMM.

Waszczuk, Jakub, Katarzyna Głowińska, Agata Savary & Adam Przepiórkowski. 2010. Tools and Methodologies for Annotating Syntax and Named Entities in the National Corpus of Polish. In *Proceedings of IMCSIT-CLA'10 Workshop, Wisła, Poland*, 531–539. Polskie Towarzystwo Informatyczne.

Watson, B. 1995. *Taxonomies and Toolkits of Regular Language Algorithms*. Eindhoven, the Netherlands: University of Technology PhD. Thesis.

Wehrli, Eric, Violeta Seretan & Luka Nerima. 2010. Sentence analysis and collocation identification. In *Proceedings of the Workshop on Multiword Expressions: from Theory to Applications (MWE 2010)*, 27–35. Beijing, China: Association for Computational Linguistics. `http://www.aclweb.org/anthology/W/W10/W10-??04`.

Wilcock, Graham. 2009. *Introduction to Linguistic Annotation and Text Analytics*. Morgan & Claypool.

Wolinski, Francis, Frantz Vichot & Bruno Dillet. 1995. Automatic Processing of Proper Names in Texts. In *EACL '95: Proceedings of the seventh conference on European chapter of the Association for Computational Linguistics*, 23–30. Morgan Kaufmann Publishers Inc.

Woliński, Marcin. 2006. Morfeusz – a Practical Tool for the Morphological Analysis of Polish. In *Proceedings of IIS:IIPWM'06*, 503–512. Springer.

Woliński, Marcin, Agata Savary, Piotr Sikora & Małgorzata Marciniak. 2009. Usability improvements in the lexicographic framework Toposław. In *Proceedings of Language and Technology Conference (LTC'09), Poznań, Poland*, 321–325. Wydawnictwo Poznańskie.

Woliński, Marcin. 2001. Rodzajów w polszczyźnie jest osiem. In Włodzimierz Gruszczyński, Urszula Andrejewicz, Mirosław Bańko & Dorota Kopcińska (eds.), *Nie bez znaczenia... Prace ofiarowane Profesorowi Zygmuntowi Saloniemu z okazji jubileuszu 15000 dni pracy naukowej*, 303–305. Białystok: Wydawnictwo Uniwersytetu Białostockiego.

Woliński, Marcin. 2003. System znaczników morfosyntaktycznych w korpusie IPI PAN. *Polonica* XXII–XXIII. 39–55.

Woliński, Marcin. 2009. A Relational Model of Polish Inflection in *Grammatical Dictionary of Polish*. In Zygmunt Vetulani & Hans Uszkoreit (eds.), *Human Language Technology: Challenges of the Information Society*, vol. 5603 Lecture Notes in Artificial Intelligence, 96–106. Berlin: Springer-Verlag.

Xing, Guangming, Chaitanya R. Malla, Zhonghang Xia & Snigdha Dantala Venkata. 2006. Computing Edit Distances Between an XML Document and a Schema and its Application in Document Classification. In *Proceedings of SAC 06, Dijon, France*, 831–835. ACM.

Zaborowski, Bartosz. 2012. *Spejd 1.3.6 - User manual*. `http://zil.ipipan.waw.pl/Spejd`.

Zhang, K. & D. Shasha. 1989. Simple fast algorithms for the editing distance between trees and related problem. *SIAM Journal on Computing* 18(6). 1245–1262.