Journal of Language Modelling

VOLUME 1 ISSUE 2
DECEMBER 2013

Institute of Computer Science
Polish Academy of Sciences
Warsaw

Journal of Language Modelling

VOLUME 1 ISSUE 2
DECEMBER 2013

JOURNAL OF
LANGUAGE MODELLING

ISSN 2299-8470 (electronic version)
ISSN 2299-856X (printed version)
http://jlm.ipipan.waw.pl/

MANAGING EDITOR
Adam Przepiórkowski IPI PAN

SECTION EDITORS
Elżbieta Hajnicz IPI PAN
Agnieszka Mykowiecka IPI PAN
Marcin Woliński IPI PAN

STATISTICS EDITOR
Łukasz Dębowski IPI PAN

Published by IPI PAN
Instytut Podstaw Informatyki
Polskiej Akademii Nauk
Institute of Computer Science
Polish Academy of Sciences
ul. Jana Kazimierza 5, 01-248 Warszawa, Poland

Layout designed by Adam Twardoch.
Typeset in XeLaTeX using the typefaces: *Playfair Display*
by Claus Eggers Sørensen, *Charis SIL* by SIL International,
JLM monogram by Łukasz Dziedzic.

All content is licensed under
the Creative Commons Attribution 3.0 Unported License.
http://creativecommons.org/licenses/by/3.0/

EDITORIAL BOARD

Steven Abney University of Michigan, USA

Ash Asudeh Carleton University, CANADA;
University of Oxford, UNITED KINGDOM

Chris Biemann Technische Universität Darmstadt, GERMANY

Igor Boguslavsky Technical University of Madrid, SPAIN;
Institute for Information Transmission Problems,
Russian Academy of Sciences, Moscow, RUSSIA

António Branco University of Lisbon, PORTUGAL

David Chiang University of Southern California, Los Angeles, USA

Greville Corbett University of Surrey, UNITED KINGDOM

Dan Cristea University of Iași, ROMANIA

Jan Daciuk Gdańsk University of Technology, POLAND

Mary Dalrymple University of Oxford, UNITED KINGDOM

Darja Fišer University of Ljubljana, SLOVENIA

Anette Frank Universität Heidelberg, GERMANY

Claire Gardent CNRS/LORIA, Nancy, FRANCE

Jonathan Ginzburg Université Paris-Diderot, FRANCE

Stefan Th. Gries University of California, Santa Barbara, USA

Heiki-Jaan Kaalep University of Tartu, ESTONIA

Laura Kallmeyer Heinrich-Heine-Universität Düsseldorf, GERMANY

Jong-Bok Kim Kyung Hee University, Seoul, KOREA

Kimmo Koskenniemi University of Helsinki, FINLAND

Jonas Kuhn Universität Stuttgart, GERMANY

Alessandro Lenci University of Pisa, ITALY

Ján Mačutek Comenius University in Bratislava, SLOVAKIA

Igor Mel'čuk University of Montreal, CANADA

Glyn Morrill Technical University of Catalonia, Barcelona, SPAIN

Stefan Müller Freie Universität Berlin, GERMANY

Reinhard Muskens Tilburg University, NETHERLANDS

Mark-Jan Nederhof University of St Andrews, UNITED KINGDOM

Petya Osenova Sofia University, BULGARIA

David Pesetsky Massachusetts Institute of Technology, USA

Maciej Piasecki Wrocław University of Technology, POLAND

Christopher Potts Stanford University, USA

Louisa Sadler University of Essex, UNITED KINGDOM

Ivan A. Sag † Stanford University, USA

Agata Savary Université François Rabelais Tours, FRANCE

Sabine Schulte im Walde Universität Stuttgart, GERMANY

Stuart M. Shieber Harvard University, USA

Mark Steedman University of Edinburgh, UNITED KINGDOM

Stan Szpakowicz School of Electrical Engineering
and Computer Science, University of Ottawa, CANADA;
Institute of Computer Science,
Polish Academy of Sciences, Warsaw, POLAND

Shravan Vasishth Universität Potsdam, GERMANY

Zygmunt Vetulani Adam Mickiewicz University, Poznań, POLAND

Aline Villavicencio Federal University of Rio Grande do Sul,
Porto Alegre, BRAZIL

Veronika Vincze University of Szeged, HUNGARY

Yorick Wilks Florida Institute of Human and Machine Cognition, USA

Shuly Wintner University of Haifa, ISRAEL

Zdeněk Žabokrtský Charles University in Prague, CZECH REPUBLIC

Populating a multilingual ontology of proper names from open sources

Agata Savary[1]*, Leszek Manicki*[2,3]*, and Małgorzata Baron*[1,2]
[1] Université François Rabelais Tours, Laboratoire d'informatique, France
[2] Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland
[3] Poleng, Poznań, Poland

ABSTRACT

Even if proper names play a central role in natural language processing (NLP) applications they are still under-represented in lexicons, annotated corpora, and other resources dedicated to text processing. One of the main challenges is both the prevalence and the dynamicity of proper names. At the same time, large and regularly updated knowledge sources containing partially structured data, such as Wikipedia or GeoNames, are publicly available and contain large numbers of proper names. We present a method for a semi-automatic enrichment of Prolexbase, an existing multilingual ontology of proper names dedicated to natural language processing, with data extracted from these open sources in three languages: Polish, English and French. Fine-grained data extraction and integration procedures allow the user to enrich previous contents of Prolexbase with new incoming data. All data are manually validated and available under an open licence.

Keywords:
proper names, named entities, multilingual ontology population, Prolexbase, Wikipedia, GeoNames, Translatica

1 INTRODUCTION

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 but they are still under-represented in lexicons, annotated corpora, and other resources dedicated to text pro-

cessing. One of the main challenges is both the prevalence and the dynamicity of proper names. New names are constantly created for new institutions, products and works. New individuals or groups of people are brought into focus and their names enter common vocabularies.

At the same time, large knowledge sources become publicly available, and some of them are constantly developed and updated by a collaborative effort of large numbers of users, Wikipedia being the most prominent example. The data contained in these sources are partly structured, which increases their usability in automatic text processing.

In this paper our starting point is Prolexbase (Krstev *et al.* 2005; Tran and Maurel 2006; Maurel 2008), an open multilingual knowledge base dedicated to the representation of proper names for NLP applications. 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*, a tool meant for a semi-automatic population of Prolexbase from Wikipedia and, to a lesser extent, from GeoNames.

Figure 1 shows the data flow in our Prolexbase population process. The three main data sources are: (i) Polish, English and French Wikipedia, (ii) Polish names in GeoNames, (iii) Polish inflection resources in Translatica, a machine translation software. Automatically selected relevant classes in Wikipedia and in GeoNames are manually mapped on Prolexbase typology. The data belonging to the mapped classes are automatically extracted and their popularity (or frequency) is estimated. Inflection rules are used to automatically predict inflected forms of both simple and multi-word entries from Wikipedia. The resulting set of candidate names is fed to ProlexFeeder, which integrates them with Prolexbase in two steps. Firstly, a candidate is automatically checked to see if it represents an entity which is already present in Prolexbase. Secondly, the entry, together with its translations, variants, relations and inflected forms is manually validated by an expert lexicographer.

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 high quality, i.e.

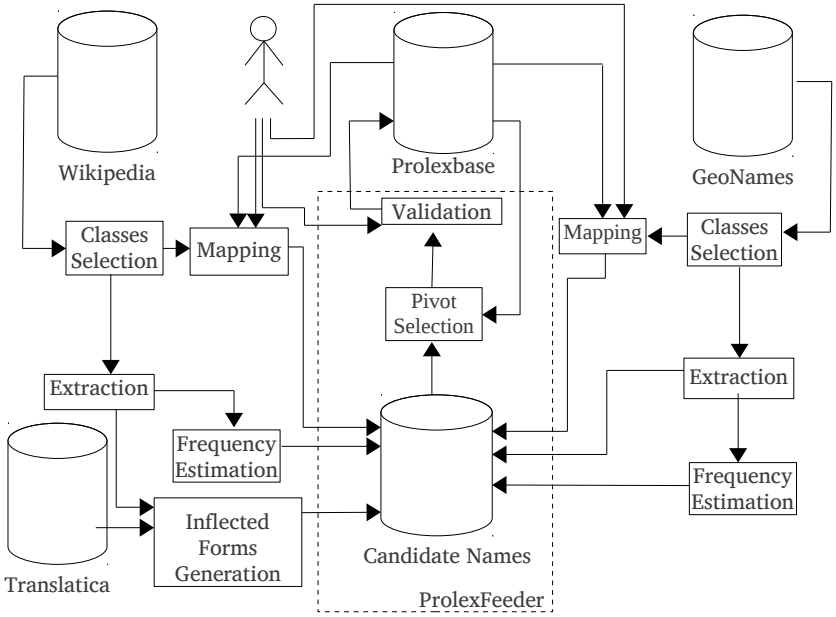


Figure 1: Data flow in Prolexbase population via ProlexFeeder.

manually validated, incremental resource dedicated to NLP. This implies:

- A rather labour-intensive activity, thus a reduced scope of the result. This requires a definition of appropriate selection criteria that allow us to retain only the most relevant, popular and stable names. In this paper we exploit criteria based on: (i) the popularity of the corresponding Wikipedia articles, (ii) systematic lists of some major categories found in GeoNames.
- Thorough data integration techniques allowing us to avoid duplication of data during an enrichment process (as opposed to extraction from scratch) in which previously validated data can be merged or completed with new incoming data.
- NLP-targeted features, particularly with respect to highly inflected languages such as Polish, which are non-existent in traditional ontologies. Prolexbase was designed with such languages in mind, notably Serbian (Krstev *et al.* 2005), which belongs, like Polish, to the family of Slavic languages. This allows us to account for rich word formation, variation and inflection processes within the same model.

Prolexbase might correspond to the *kernel NE lexicon*, i.e. the common shared NE vocabulary appearing in texts of differents 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-occurence with the kernel NEs.

This paper is organized as follows. Section 2 summarizes the major features of Prolexbase and of the input data sources relevant to the population process. Section 3 describes data integration issues. In Section 4 we briefly address the human validation interface. Section 5 is dedicated to evaluation of the population process and of a named entity recognition system using the resulting Prolexbase resources. Section 6 contains a detailed discussion of related work. Finally, Section 7 concludes our contributions, and Section 8 summarizes some perspectives for future work and mentions possible applications of the rich Prolexbase model and data.

2 INPUT KNOWLEDGE SOURCES

2.1 *Prolexbase*

Prolexbase (Krstev *et al.* 2005; Tran and Maurel 2006; Maurel 2008) offers a fine-grained multilingual model of proper names whose specificity is both concept-oriented and lexeme-oriented. Namely, it comprises a language-independent ontology of concepts referred to by proper names, 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** and 34 **types**, cf. (Agafonov *et al.* 2006):

1. *Anthroponym* is the supertype for individuals – *celebrity*, *first name*, *patronymic*, *pseudo-anthroponym* – and collectives – *dynasty*, *ethnonym*, *association*, *ensemble*, *firm*, *institution*, and *organization*.
2. *Toponym* comprises territories – *country*, *region*, *supranational* – and other locations – *astronym*, *building*, *city*, *geonym*, *hydronym*, and *way*.

3. *Ergonym* includes *object*, *product*, *thought*, *vessel*, and *work*.
4. *Pragmonym* contains – *disaster*, *event*, *feast*, *history*, and *meteorology*.

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 in the state-of-the-art approaches discussed in Section 6. Secondly, these concepts, called **pivots**, include 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*, understood as a collective anthroponym). **Accessibility** means that one referent is accessible through another, generally better known, referent (Tran and 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

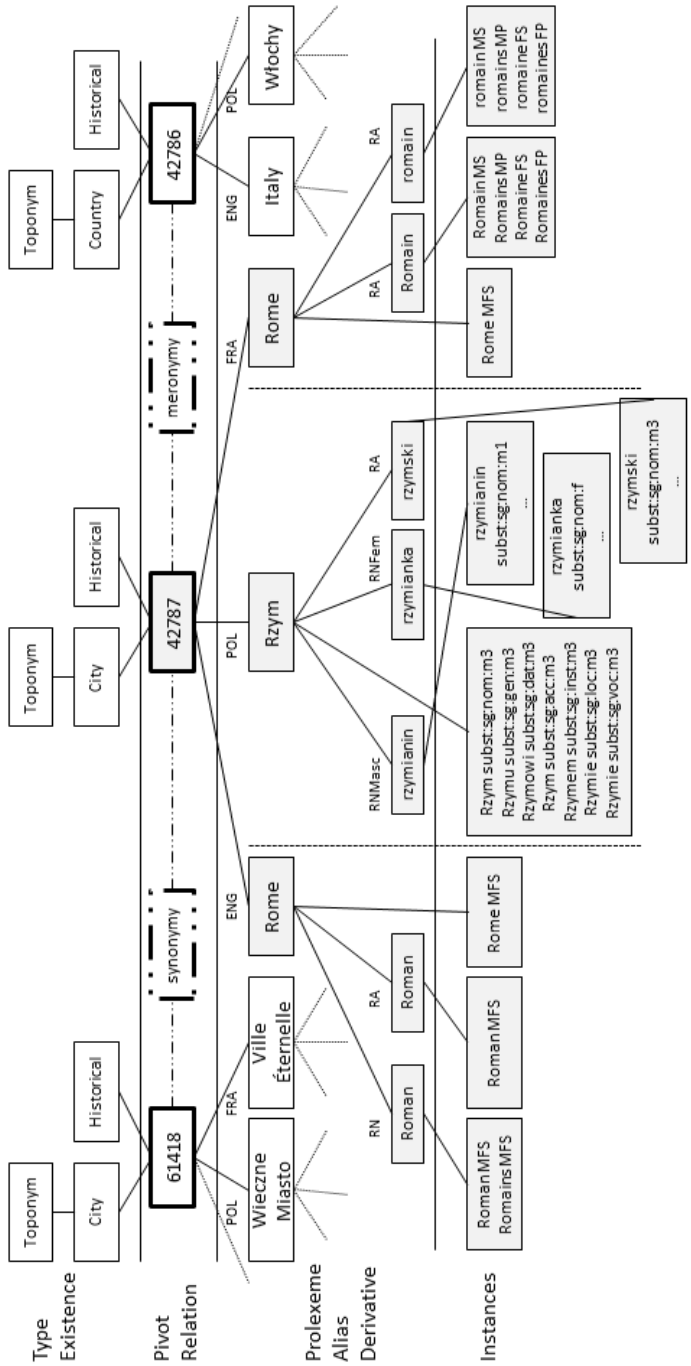


Figure 2: Extract of the intended contents of Prolexbase with four levels and three prolexemes.

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**[1] 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 2, 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.

In order to adapt Prolexbase to being populated with Wikipedia data in an automated way several minor changes in the original Prolexbase structure have been made. Notably, the **Wikipedia link** attribute has been added to the description of prolexemes in every language. Furthermore, since intense searches of prolexemes, aliases and instances are frequently performed by ProlexFeeder, indices have been created on appropriate data.

2.2 *Wikipedia*

Wikipedia is a constantly growing project grouping a set of open-source online encyclopaedia initiatives run by the MediaWiki software and filled with content by volunteer authors. Polish is the sixth largest

[1] Note that Prolexbase terminology is non-standard with respect to WordNet (Miller 1995). Notably, in Prolexbase hypernyms of entities referred to by proper names are *metaconcepts*, entities are *concepts* (represented by pivot identifiers), and the inflected forms of names are called *instances*. In WordNet, hypernyms of entities are *concepts* while the surface forms of the entities themselves are called *instances*. See also Section 6 for a discussion on the instance-to-concept mapping which we perform, as opposed to the concept-to-concept mapping standard in the related work.

Wikipedia project with its 900,000 articles. We use the dump containing all Polish articles of the current release available at the beginning of 2011. The data extraction process described in Section 3.1.1 may be iteratively repeated using a newer Wikipedia dump release in order to add new entries to Prolexbase and complete the existing ones.

Wikipedia articles describing the same topic in different languages can be connected by *interwiki* links. We used this interconnection shorthand feature to automatically extract translations for titles of Polish articles.

Categories and *infobox templates* are two possible means of classifying Wikipedia articles. Both are language-specific and user-defined. No mechanism is provided for ensuring a compatibility of category hierarchies in different languages. As a result, a Polish entry and its English or French equivalent may be assigned to non-equivalent categories or incompatible category hierarchies. Moreover, categories are often used by Wikipedia users to group related articles rather than to create a hierarchical structure of data. Thus, some categories may include both individual entities and general domain-related terms. For instance, the category *powiaty* 'counties' in Polish Wikipedia contains the list of Polish counties but also terminological items such as *powiat grodzki* 'city county' (a county type in the current Polish administrative division system) or *powiaty i gminy o identycznych nazwach* 'Polish homonymous counties and communes' (containing the list of homonymous Polish administrative division units). Conversely, infoboxes are usually added to articles that only cover individual entities, not general domain-related terms. For this reason, we used infobox templates as the main criteria for extracting and classifying proper names from Wikipedia, as described in Section 3.1.1.

Like categories, *redirects* belong to special classes of Wikipedia articles. They allow one to automatically access an article whose title is not identical with the query. Redirects may be used for various purposes including orthography and transcription variants, official names (1), eliptical variants, acronyms and abbreviations, diachronic variants (2), pseudonyms, common spelling errors and names with extra disambiguating data (3).

(1) Main Polish entry: *Wielka Brytania* 'Great Britain'
Redirects: *Zjednoczone Królestwo* 'United Kingdom', *Zjednoczo-*

ne Królestwo Wielkiej Brytanii i Irlandii Północnej 'United Kingdom of Great Britain and Northern Ireland'

(2) Main Polish entry: *Plac Powstańców Warszawy w Warszawie* 'Warsaw Uprising Square in Warsaw'
Redirects: *Plac Napoleona* 'Napoleon Square', *Plac Warecki* 'Warka Square'

(3) Main English entry: *Sierra Blanca* (settlement in Texas)
Redirects: *Sierra Blanca (TX)*, *Sierra Blanca, TX*

2.3 *GeoNames*

GeoNames is a database of geographical names collected from various publicly available and official sources such as American National Geospatial-Intelligence Agency (NGA), U.S. Geological Survey Geographic Names Information System or British Ordnance Survey. The database contains over 10 million records related to over 8 million unique features. It stores toponym names in different languages but also some encyclopaedic and statistical data such as elevation, population, latitude and longitude. Information on administrative subdivision is also provided for numerous entries. Entries are categorized into 9 main classes which in turn divide into 645 highly fine-grained subcategories.[2] For instance, code *S.CAVE* refers to the subcategory *cave* of the main class *spot*. All the data are freely available under the Creative Commons Attribution license[3], both through the GeoNames web interface and through numerous programming libraries (APIs). As GeoNames exploits heterogeneous sources and the quality of its contents may vary, a wiki-like interface is provided for users in order to correct and expand the data.

2.4 *Translatica*

As described in Section 2.1, Prolexbase entries in any language are supposed to be supplied with their inflected forms called *instances*. Neither GeoNames, nor Wikipedia contain explicit inflection or grammat-

[2] `http://www.geonames.org/export/codes.html`
[3] `http://creativecommons.org/licenses/by/3.0/`

ical data. Due to the limited inflection system of English and French proper names, we do not automatically generate inflected forms of entries in these languages. Polish, however, has a rich inflection system and instances have to be suggested automatically if the human validation is to be efficient. We use the inflection routine, based on dictionary lookup and guessing, developed for Translatica (Jassem 2004), a Polish-centred machine translation system. For over 260,000 extracted Wikipedia entries almost 2 million instances have been collected in this way. We used the *Morfologik* dictionary[4] as a source of inflected forms both for single entries and for components of multiword units. All Polish instances were further manually validated and corrected before their addition to Prolexbase (cf. Section 4).

3 DATA INTEGRATION

3.1 *Data selection*

Wikipedia and GeoNames were used as the main sources of new entries for Prolexbase enrichment. In this section we describe the process of extracting structured data from both sources.

3.1.1 Data selection from Wikipedia

Since Wikipedia is a general-purpose encyclopaedia, the first challenge was to select only those Wikipedia articles whose titles represent proper names. Initially, Wikipedia categories seemed to provide natural selection criteria. Some previous attempts, such as (Toral *et al.* 2008), are based, indeed, on mapping WordNet synsets onto Wikipedia categories and on applying capitalisation rules for retaining only virtual proper names from a set of entries. However the high number of Wikipedia categories (1,073 low-level in Polish, 73,149 in total) and their heterogeneous nature explained in Section 2.2 made us turn to primarily using **infoboxes**, similarly to DBpedia (Bizer *et al.* 2009).

We extracted the list of all infobox templates used in Polish Wikipedia and manually selected those which seemed related to proper names. As a result we obtained 340 relevant templates. We extracted all Polish entries containing infoboxes built upon these templates.

[4] `http://morfologik.blogspot.com`

Each entry was assigned a class based on the name of the corresponding infobox template. English and French translations of Polish entities (if any) were extracted via interwiki links. Thus, we obtained a trilingual list of classified named entities, henceforth called *initWikiList*.

The Polish version of Wikipedia, unlike e.g. the English version, contains rather few infobox templates referring to people. Even if several specific classes, like *żołnierz* 'soldier', *piłkarz* 'football player' or *polityk* 'politician' do exist, the major part of people-related articles contain a *biogram* 'personal data' infobox, consisting only of basic personal data (date of birth, nationality, etc.). The *initWikiList* contained numerous Polish entries with an infobox of the *biogram* class. We noticed that such entries often belong to fine-grained Wikipedia **categories**, e.g. *niemieccy kompozytorzy baroku* 'German Baroque composers'. These categories turned out to be rather homogeneous in terms of including only actual named entities, and not general domain-related terms (cf. Section 2.2). Moreover, many articles belonging to these categories had no infobox attached.

This observation led us to extending the coverage of the extraction process. We collected the list of 676 person-related categories containing entries from *initWikiList*. Then we expanded *initWikiList* with all entries from these categories that did not contain an infobox. Each entry from the resulting trilingual list was assigned: (i) its Wikipedia URLs in Polish, English and French (if any) (ii) its *Wikipedia class*, i.e. its Polish infobox class (if its article contained an infobox) or its Polish category (if the entry belonged to a person-related Wikipedia category). After filtering out some evident errors we obtained the final list of candidate proper names and their associated data to be added to Prolexbase. The list contained 262,124 Polish entries with 255,835 English and 139,770 French translations.

As mentioned in Section 2.2, Wikipedia **redirects** may be valuable sources of aliases and synonyms for the retrieved entries but they are heterogeneous in nature. Orthography and transcription variants, official names (1), eliptical variants, acronyms and abbreviations represent aliases in terms of Prolexbase. Diachronic variants (2) and pseudonyms correspond to diachronic and diastratic synonyms, respectively. Spelling errors and variants with disambiguating data (3) are irrelevant. We could automatically remove only redirects of type (3), as well as those pointing at article subsections rather than articles

themselves. The elimination of spelling errors, as well as the distinction between virtual aliases and synonyms had to be left for further manual validation stage (cf. Section 4). The final resulting collection contained 33,530 redirects to Polish, 297,377 to English, and 92,351 to French Wikipedia articles.

3.1.2 Data selection from GeoNames

As the amount of entries in GeoNames is huge it is hardly feasible to validate all of them manually before adding them to Prolexbase. Thus, it was necessary to select a well-defined subset of these data. We have used only the country names[5], all Polish names[6], as well as alternate names[7]. We have examined several category-dependent selection criteria based on numerical data accessible in GeoNames such as the height of a mountain or the population of a city. Such criteria proved hard to apply in a general case: some well-known mountains or cities are low or have few inhabitants. We finally decided to treat GeoNames as complementary to Wikipedia as far as the selection criteria are concerned. Namely, Wikipedia entries were sorted by their *frequency* value based on the popularity of the corresponding articles in Wikipedia, as discussed in Section 3.3. Conversely, GeoNames was used as a source of systematic lists of names belonging to some major categories. Thus far, the following GeoNames categories have been selected: (i) all countries and their capitals, (ii) all first-order (*województwo*) and second-order (*gmina*) administrative division units in Poland and their chief towns, (iii) all first-order administrative division units in other European countries and their chief towns. Other GeoNames entries were extracted only if they referred to entities located in Poland. The total number of entries selected from GeoNames according to these criteria was equal to 42,376.

3.2 *Ontology mapping*

Merging different ontologies into a common structure is a well-known problem, as discussed in Section 6. In most approaches, the aim is to propose a unified framework in which one ontology is mapped onto another and the granularity of both can be fully conserved.

[5] `http://download.geonames.org/export/dump/allCountries.zip`
[6] `http://download.geonames.org/export/dump/PL.zip`
[7] `http://download.geonames.org/export/dump/alternateNames.zip`

In our work, the aim of ontology mapping is different. We aim at creating a named entity resource, whose typology size is balanced with respect to NLP tasks such as named entity recognition (NER), machine translation, etc. This requires usually several dozens of types at most. Thus, we wish to map the types of our source ontologies (Wikipedia and GeoNames) on types and relations of Prolexbase so that only the typology of the latter resource is conserved. This mapping has been manually performed, as described in this section.

3.2.1 Mapping Wikipedia onto Prolexbase ontology

All Polish Wikipedia classes (340 infobox classes or 676 person-related categories, cf. Section 3.1.1) proved appropriate for a rather straightforward mapping onto Prolexbase types and existence values (historical, fictitious or religious). For instance, the Wikipedia infobox class *Postać telenowela* ('Soap opera character') could be mapped on Prolexbase type *Celebrity* and *fictitious* existence.

Moreover, numerous Wikipedia classes were specific enough to allow a global assignment of other relations as well. A (language-independent) meronymy relation with a toponym was the most frequent one. For example, the Wikipedia category *Władcy Blois* ('Counts of Blois') was mapped on Prolexbase type *Celebrity*, *historical* existence, and *accessibility* relation with *Blois* with the *leader* subject file.

The mapping and the selection of related pivots was done manually. As a result, each Wikipedia entry was automatically assigned the Prolexbase type, existence, meronymy and/or accessibility on which its Wikipedia class was mapped. Rare erroneous assignments that might result for individual entries from this global mapping were to be fixed in the human validation stage.

3.2.2 Mapping GeoNames onto Prolexbase ontology

A mapping was also necessary between GeoNames and Prolexbase typologies. In most cases global assignment of GeoNames main categories to Prolexbase types was sufficient. However, several GeoNames subcategories refer to different Prolexbase types than their parent main categories, e.g. the subcategory *S.CAVE* (cave, caves) corresponds to the Prolexbase type *geonym* although its parent category *S* (spot, building, farm) is mapped on type *building*.

3.3 *Frequency code estimation*

As mentioned in the Section 2.1, every language-specific entry (prolexeme) in Prolexbase obtains one of three *frequency* labels which describes how popular the given prolexeme is:

1. commonly used,
2. infrequently used,
3. rarely used.

Since Wikipedia does not indicate any similar measure for its articles we based our estimation on monthly statistics of *Wikipedia hits*[8] from January 1st, 2010 to December 31st, 2010. We split Wikipedia entries into 4 subclasses: cities (that made about a half of all entries that we had collected), people (celebrities – approx. 25% of all entries), works and other entries. Hit count thresholds of frequency groups were rather arbitrarily[9] set for every subclass separately:

- for *celebrity* and *work* subclasses: 10% of entries with the highest number of visits received code 1 (*commonly used*), next 30% got code 2 (*infrequently used*) and the rest was assigned code 3 (*rarely used*),
- for *city* and *other* subclasses: the first 4% received code 1, next 16% – code 2, and the rest – code 3.

Note that these values are defined for prolexemes rather than pivots, e.g. a person may be very well known in Poland, thus it has frequency code 1 in Polish, while it gets code 2 or 3 in French or English.

The definition of frequency values for GeoNames followed the assumption that it was a secondary resource. Thus, each name appearing also in Wikipedia kept the Wikipedia hit-based frequency code. All other names of countries, European and Polish administrative division units, as well as capitals of these entities, were assigned code 1, since we wished to include these major classes on a systematic basis. The

[8] Available via the `http://stats.grok.se/` service

[9] A group of 3 French and Polish native experts examined the list of entries ordered according to the decreasing value of Wikipedia hits. The frequency code was supposed to be 1 as long as at least 2 entries known by at least one of the experts appeared in consecutive windows of about 30-entries. The threshold choice between code 2 and 3 was arbitrary.

remaining names were arbitrarily distributed over the 3 codes – see (Savary *et al.* 2013) for details.

3.4 *Pivot selection*

Data extracted from Wikipedia represent concepts and relations which may already be present in Prolexbase. Thus, the main challenge is to preserve the uniqueness of concepts, i.e. to select the proper (language-independent) pivot if the current concept is already present in Prolexbase, and to create a new pivot otherwise. Working on three languages simultaneously greatly increases the reliability of this process. Recall that Prolexbase originally contained mostly French data. If new Polish or English data were to be examined separately, few hints would be available as to the pre-existence of adequate pivots. For instance, if Prolexbase already contains the prolexeme *Aix-la-Chapelle* with pivot 45579, it is hard to guess that the incoming Polish prolexeme *Akwizgran* should be attached to the same pivot. If, however, all three equivalents – *Aachen* (EN), *Aix-la-Chapelle* (FR) and *Akwizgran* (PL) are extracted from Wikipedia then their matching with pivot 45579 is straightforward.

While selecting the most probable pivot, ProlexFeeder assumes that: (i) the current content of Prolexbase has the validated status, (ii) data added automatically have the non-validated status, (iii) while validating an entry we rely only on the already validated data. Due to homonymy and variation, comparing the Wikipedia entry with a prolexeme is not enough. At least three other sources of evidence may be exploited. Firstly, some homonyms can be distinguished by their type, e.g. the Wikipedia entry *Aleksander Newski* as a work (film) should not be mapped on the pivot of type celebrity. Secondly, a Wikipedia entry may be equal to an alias rather than a prolexeme of an existing pivot. For instance, the main entry in Example (1) ('Great Britain'), is shorter than its alias ('United Kingdom of Great Britain and Northern Ireland') in Wikipedia, conversely to Prolexbase, where the most complete name is usually chosen as the prolexeme. Thirdly, a common URL is a strong evidence of concept similarity.

Consider Table 1 showing a sample set of Wikipedia data resulting (except the *pivot* attribute) from the preprocessing described in the preceding sections. Figure 3 sketches the algorithm of pivot selection for a new data set e. Its aim is to find each pivot p existing

Figure 3: Algorithm for selecting candidate pivots for a new incoming entry.

Function *getPivotCandidates*(e) return *pivotList*
Input e: structure as in Table 1 //incoming entry
Output *pivotList*: ordered list of (p, d) with $p, d \in \mathbb{N}$ //proposed pivots and their distances from e

1. **begin**
2. **for each** $l \in \{PL, EN, FR\}$ **do**
3. $pivots.l \leftarrow \langle\rangle$ //empty list
4. **for each** $p \in allPivots$ **do** //for each existing pivot
5. **for each** $l \in \{PL, EN, FR\}$ **do** //for each language
6. **if** $distance(e, p, l) < 10$ **then**
7. $insertSorted(p, pivots.l)$ //insert the new pivot in the sorted candidates list
//merge three sorted candidate lists into one
8. $pivotList \leftarrow mergeSorted(pivots.PL, pivots.EN, pivots.FR)$
9. **if** $pivotList = \langle\rangle$ **then** //no similar pivot found
10. $pivotList \leftarrow \langle(getNewPivot(), 0)\rangle$ //create a new pivot
11. **return** *pivotList*
12. **end**

Function *distance*(e, p, l) return d
Input e: structure as in Figure 1 //incoming entry
p: pivot
$l \in \{PL, EN, FR\}$ //language
Output $d \in \{0, 1, 2, 3, 10\}$ //distance between e and p

13. **begin**
14. $d = 10$
15. **if** $e.l.lex = p.l.lex$ **then** $d \leftarrow 0$ //same lexeme
16. **else if** $e.l.lex \in p.l.aliases$ **then** $d \leftarrow 1$ //lexeme same as an alias
17. **else if** $e.l.url = p.l.url$ **then** $d \leftarrow 2$ //matching Wiki URL
18. **if** $d \leq 1$ **and** $e.l.url \neq p.l.url$ **and** $e.type \neq p.type$ **then** $d \leftarrow 3$
19. **return** d
20. **end**

in Prolexbase such that, for each language l (PL, EN or FR), the data linked with p (if any) are similar to e. The similarity between e and p grows with the decreasing value of the *distance* function, which compares the lexemes, aliases, URLs and types of e and p in the given language. We assume that e is likely to be similar to p in any of the following cases: (i) e and p share the same lexeme in a particular language (line 15), (ii) e is an alias of p (line 16), (iii) e and p share the same URL (line 17). In the last case, a bi-directional matching of lexemes and aliases between Wikipedia and Prolexbase is not always a good strategy. For instance, the redirects in Example (2) are former names ('Napoleon Square', 'Warka Square') of a square ('Warsaw Uprising Square in Warsaw'). Recall that in Prolexbase such variants are not considered as aliases but refer to different pivots (linked by the diachronic synonymy relation). Finally, we give a penalty if e shares the lexeme with the existing pivot p but either their URL or their type differ (line 18).

The *distance* function is used to compare an incoming Wikipedia entry e with each pivot existing in Prolexbase (lines 4–6). For each of the three languages we get a sorted list of pivots which are similar to e (line 7). The three resulting lists are then merged (line 8) by taking two

Table 1: Sample preprocessed Wikipedia data. The attributes represent: Wikipedia lexemes (*PL.lex*, *EN.lex*, *FR.lex*), number of Wikipedia hits in 2010 (*PL.hits*, *EN.hits*, *FR.hits*), frequency (*PL.freq*, *EN.freq*, *FR.freq*), Wikipedia page URL (*PL.url*, *EN.url*, *FR.url*), Wikipedia redirects proposed as aliases (*PL.aliases*, *EN.aliases*, *FR.aliases*), 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*	EN.aliases	*Capital of Italy*, *Castel Fusano*, *Città Eterna*, …	FR.aliases	*Ville Éternelle*, *Ville éternelle*
PL. infl	*Rzymu:sg:gen:m3*, *Rzym:sg:acc:m3*, …	type	city	existence	historical
		meroPivot	none	pivot	42787

factors into account: (i) the rank of a pivot in each of the three lists, (ii) its membership in the intersections of these lists. If no similar pivot was found in any language then a new pivot is proposed (line 9–10).

The actual implementation of this algorithm does not scan all existing pivots for each incoming entry e. The entry is directly compared, instead, to the existing lexemes and aliases in the given language, which is optimal if data are indexed. For instance, if we admit that the database engine running Prolexbase implements indexes on B-trees, and that l, p and a denote the worst-case length of a candidate pivot list, the number of the existing prolexemes and of the existing aliases, respectively, the complexity of our algorithm is of $O(\log p + \log a + l)$. In practice, p was up to four times higher than a, and l was no higher than 10. The candidate pivot searching algorithm proved not to be a bottleneck of our procedure. On average, it takes less than a second to pre-process (off-line) a single Wikipedia entry.

The pivots returned by the algorithm in Figure 3 are proposed to a human validator as possible insertion points for new Wikipedia data, as discussed in Section 4. When the correct pivot has been selected by the lexicographer, ProlexFeeder considers different strategies of merging the new incoming data with the data attached to this selected pivot. For instance, an incoming lexeme may take place of a missing prolexeme or it can become an alias of an existing prolexeme. The values of frequency, URL, aliases, inflected forms, existence, holonym/meronym, and type predicted for the incoming entry (cf. Table 1) may be either complementary or inconsistent with those of the selected pivot. In the latter case, the Prolexbase data are considered as more reliable but the user is notified about the conflict.

As far as the insertion of a GeoNames entry is concerned, the entry is first straightforwardly matched with the extracted Polish Wikipedia entries. If an identical entry is found then its attributes become those of the GeoNames entry (except, possibly, the frequency code, cf. Section 3.3). Otherwise it is considered that the GeoNames entry has no corresponding Wikipedia entry and thus many attributes of its structure shown in Figure 1 become empty. Note that this matching process is less reliable than matching Wikipedia entries with Prolexbase. This is because a significant amount of GeoNames entities do not have translations to other languages, e.g. *Zala*, a Hungarian first-order administrative division unit, is represented in GeoNames with

its Hungarian name only. Although there exist articles describing the same concept in Polish and English Wikipedia (*Komitat Zala* and *Zala County*, respectively) they could not be mapped on *Zala* alone. As a result, both the Wikipedia and the GeoNames entries were suggested as new Prolexbase entries with two different pivots. This problem occurred most often for regions (European administrative division units) extracted from GeoNames, many of which were cited in the holonym country's language only. During the human validation, proper Polish, English and French equivalents were to be found manually for such names, which made the whole procedure highly time-consuming. Therefore, those region names that were hard to identify manually were left for a further stage of the project.

4 HUMAN VALIDATION

The aim of Prolexbase is to offer high-quality lexico-semantic data that have been manually validated. Thus, the results of the automatic data integration presented in Section 3 do not enter Prolexbase directly but are fed to a graphical user interface offered by ProlexFeeder. There, the lexicographer first views new entries proposed by the automatic selection and integration process then validates, completes and/or deletes them. She can also browse the current content of Prolexbase in order to search for possible skipped or mismatched pivots and prolexemes.

Most often, the incoming entries are new to Prolexbase but sometimes they match existing pivots which can be detected by the pivot selection procedure (cf. Section 3.4). In this case, the data coming from external sources complete those already present. Prolexemes in the three languages are proposed together with their Wikipedia URLs (which are usually new to Prolexbase). Some aliases, like *Wieczne Miasto* ('Eternal City') in Table 1, can be transformed into new pivots. Missing relations as well as derivations can be added manually, and the proposed inflected forms of the Polish prolexeme can be corrected or validated.

5 EVALUATION

In order to estimate both the quality of the data integration process and the usability of the human validation interface, samples of Wikipedia entries of three different types were selected: celebrity, work and city, containing 500 entries each. A lexicographer was to process these samples type by type in the GUI, collect the statistics about wrongly proposed pivots and count the time spent on each sample. Table 2 shows the results of this experiment. A true positive is a pivot that has existed in Prolexbase and is correctly suggested for an incoming entry. A true negative happens when there is no pivot in Prolexbase corresponding to the incoming entry and the creation of a new pivot is correctly suggested. A false positive is an existing pivot that does not correspond to the incoming entry but is suggested. Finally, a false negative is an existing pivot which corresponds to the entry but which fails to be suggested (i.e. the creation of a new pivot is suggested instead). Type city has the largest number of true positives since initially Prolexbase contained many French toponyms, some celebrity names and only very few names of works. The true negatives correspond to the newly added concepts. The false positives are infrequent and their detection is easy since the lexicographer directly views the details of the wrongly proposed pivot. False negatives are the most harmful since detecting them requires a manual browsing of Prolexbase in search of prolexemes similar to the current entry. Fortunately, these cases cover only 1.3% of all entries.

Table 2: Results of ProlexFeeder on three sets of entries.

Type	Incoming entries	True posit.	True negat.	False posit.	False negat.	Accuracy	Workload
Celebrity	500	87	400	1	12	97.4%	21h30
Work	500	9	472	16	3	96.2%	17h30
City	500	226	264	6	4	98%	16h
All	1500	322	1136	23	19	97.2%	55h

Wrongly selected pivots result mainly from the strict matching algorithm between an incoming lexeme and existing prolexemes and aliases (cf. Figure 3, lines 15–16). For instance, the Polish Wikipedia entry *Johann Sebastian Bach* did not match the Polish prolexeme *Jan Sebastian Bach*, while *The Rolling Stones* appeared in Prolexbase as

Rolling Stones with a collocation link to *The*. Some true homonyms also appeared, e.g. the pivot proposed for *Muhammad Ali* as a boxer represented in fact the pasha of Egypt carrying the same name. The evidence of different French equivalents (*Muhammad Ali* and *Méhémet-Ali*) was not strong enough to allow for the selection of different pivots. Similarly, *Leszno* in the Wielkopolska Province was mistaken for *Leszno* in Mazovia Province.

On average, the processing of an incoming entry takes about 2 minutes. Most of this time is taken by completing and/or correcting the inflected forms of Polish prolexemes (usually 7 forms for each name). Inflecting celebrity names proves the most labour-intensive since Translatica's automatic inflection tool (cf. Section 2.4) makes some errors concerning person names: (i) their gender is wrongly guessed, (ii) the inflection of their components is unknown (thus we get e.g. **Maryla Rodowicza* instead of *Maryli Rodowicz*). Moreover the inflection of foreign family names is a challenge for Polish speakers.

The morphological description of works is easier since they often contain common words (*Skrzynia umarlaka* 'Dead Man's Chest') or they do not inflect at all (*Na Wspólnej* 'On the Wspolna Street'). The main challenge here is to determine the proper gender. For instance *Mistrz i Małgorzata* 'The Master and Margarita' may be used in feminine (while referring to the classifying context *książka* 'the book'), in masculine (the gender of *Mistrz* 'Master'), or even in masculine plural (to account for the coordination dominated by the masculine noun).

Inflecting city names proved relatively easy – most of them contained one word only and their morphology was rather obvious. Notable exceptions were again foreign names for which the application of a Polish inflection paradigm may be controversial (e.g. *w okolicach Viborga/Viborg* 'in the vicinity of Viborg'). Surprisingly enough, the major difficulty for this type came from the fact that almost 50% of the cities already had their pivot in Prolexbase. Since several settlements with the same name frequently occur checking all necessary relations in order to validate the suggested pivot could be non-trivial.

Other problems concerned types and relations. Wrong types were systematically proposed for some groups of Wikipedia entries due to particularities of Wikipedia categories and infobox types. For instance, the names of music bands (*Genesis*) are classified in Wikipedia jointly with individual celebrities, thus changing their Prolexbase type to En-

semble had to be done manually. In samples of type city only one type error appeared (*Trójmiasto* 'Tricity' had to be reclassified as a region), and all works had their type correctly set.

Missing relations are due to the fact that they are not directly deducible from the Wikipedia metadata that were taken into account until now. Thus, the following relations had to be established manually: (i) accessibility between ensembles and their members (*Wilki* and *Robert Gawliński*) or between works and their authors (*Tosca* and *Giacomo Puccini*), (ii) meronymy between celebrities or works and their birth or edition countries (*Kinga Rusin* and *Poland*, the *Wprost* magazine and *Poland*), (iii) meronymy between cities and countries or regions (if several settlements sharing the same name are situated in the same country the meronymy is established with respect to smaller territories allowing for semantic disambiguation). Recall also that derivatives had to be established fully manually.

Prolexbase has already been successfully used for named entity recognition and categorization in French with an extended NE typology (Maurel *et al.* 2011). However, since Prolexbase models both semantic and morphological relations among proper names, we expect the benefit from this resource to be most visible in NLP applications dedicated to morphologically rich languages. The first estimation of this benefit has been performed for Nerf[10], a named entity recognition tool based on linear-chain conditional random fields. Nerf recognizes tree-like NE structures, i.e., containing recursively nested NEs. We used the named entity level of the manually annotated 1-million word National Corpus of Polish, NKJP (Przepiórkowski et al., 2012) 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 indi-

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

cates 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 $8.011e{-}08$.

These results are encouraging, especially given the fact that Nerf's initial performances were rather good due to the big size and the high quality of the training corpus (NKJP), which had been annotated manually by two annotators in parallel, and then adjudicated by a third one.

6 RELATED WORK

Before ProlexFeeder was created, Prolexbase population had been performed mostly manually (Tran *et al.* 2005). Uniqueness of pivots was supported by a rather straightforward method based on a prolexeme match alone. Lists of entries and attributes were crafted in spreadsheet files which were then automatically inserted to Prolexbase provided that pivot identifiers appeared in them explicitly. 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. As a result, Prolexbase contained initially mainly French data. 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.

Prolexbase can be compared to *EuroWordNet* (EWN) (Vossen 1998) and to the *Universal Wordnet* (UWN) (Melo and Weikum 2009), although both of them are general-purpose wordnets with no particular interest in named entities. All three resources distinguish a language-independent and a language-specific layer. Language-

independent entities, i.e. Interlingual Index Records (ILIRs) in EWN and pivots in Prolexbase, provide translation of lexemes in language-specific layers (but ILIRs unlike pivots form an unstructured list of meanings). UWN, conversely, provides direct translation links between terms in different languages. The main specificity of Prolexbase w.r.t. EWN and UWN is that proper names are concepts in Prolexbase while they are instances in EWN and UWN. Thus, adding a new proper name to Prolexbase implies enlarging its conceptual hierarchy, which does not seem possible e.g. with automatic UWN population algorithms.

Prolexbase population from Wikipedia and GeoNames can be seen as an instance of the *ontology learning* problem. According to the taxonomy proposed by Petasis *et al.* (2011), we simultaneously perform *ontology enrichment* (placing new conceptual proper names and relations at the correct positions in an existing ontology) and *ontology population* (adding new instances of existing concepts). The former is based on integrating existing ontologies (as opposed to constructing an ontology from scratch and specializing a generic ontology). The latter is atypical since we use instances of existing ontologies and inflection tools, rather than extraction from text corpora.

Ontology integration corresponds roughly to what Shvaiko and Euzenat (2013) call *ontology matching*. Our position with respect to the state of the art in this domain is twofold. Firstly, we perform a mapping of Wikipedia classes and GeoNames categories on Prolexbase types (cf. Section 3.2). This fully manual mapping produces subsumption relations and results in a particular type of an n:1 alignment. Namely, a Wikipedia infobox class is mapped on one Prolexbase type and on a set of relations (cf. Section 3.2.1). Note also that instance-based ontology matching approaches, mentioned in the same survey, can be seen as opposed to ours. They use instances attached to concepts as evidence of concept equivalence, while we, conversely, rely on the types of proper names (i.e. concepts) from Wikipedia or GeoNames in order to find the equivalent names (i.e. instances), if any, in Prolexbase.

Secondly, we map names from Wikipedia and GeoNames on conceptual proper names (pivots) in Prolexbase (cf. Section 3.4). This mapping is inherently multilingual and subsumption-based. It outputs 1:n alignments, due to e.g. diachronic synonymy as in Example (2). It is supported by a full-fledged matching validation interface and leads

to ontology merging (as opposed to question answering). It uses string equality on the terminological level, is-a similarity on the structural level, object similarity on the extensional level and does not apply any method on the semantic level.

This comparison with ontology matching state of the art is not quite straightforward since no conceptualization of proper names takes place in Wikipedia and GeoNames (but also in other common ontologies, like WordNet). Thus, mapping multilingual sets of instances (names) from Wikipedia and GeoNames on Prolexbase pivots corresponds to an instance-to-concept rather than a concept-to-concept matching. This is why our method can more easily be situated with respect to the problem of the creation and enrichment of lexical and semantic resources, in particular for proper names, and their alignment with free encyclopaedia and thesauri. This problem has a rather rich bibliography most of which was initially dedicated to English and is more recently being applied to other languages. Several approaches are based on aligning WordNet with Wikipedia: (Toral *et al.* 2008), (Toral *et al.* 2012), (Fernando and Stevenson 2012), (Nguyen and Cao 2010), *YAGO* (Suchanek *et al.* 2007) and *YAGO2* (Hoffart *et al.* 2011). Others build new semantic layers over Wikipedia alone: *Freebase* (Bollacker *et al.* 2007), *MENTA* (Melo and Weikum 2010), *DBpedia*[11] (Bizer *et al.* 2009; Mendes *et al.* 2012). DBpedia is the only resource to explicitly provide support for natural language processing tasks (data sets of variants, thematic contexts, and grammatical gender data).

Table 3 shows a contrastive study of these methods[12]. As can be seen, we offer one of the approaches which explicitly focus on modelling proper names instead of all nominal or other entities and concepts. Like YAGO and Freebase authors, but unlike others, we use multiple knowledge sources, and like three other approaches we consider several languages simultaneously rather than English alone. We share with DBpedia the idea of a manual typology mapping from Wikipedia infobox templates to ontology types, but we extend the relative (with respect to categories) reliability of infobox assignment by including articles from categories automatically judged as reliable. Like Universal

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

[12] A more complete survey can be found in (Savary *et al.* 2013).

Wordnet but unlike others we order input entries by popularity (estimated via Wikipedia hits, while the UWN uses corpus frequencies). Like Freebase[13] but unlike others we manually validate all preprocessed data.

Most important, we aim at a limited size but high quality, manually validated resource explicitly dedicated to natural language processing and focused on proper names. Thus, we are the only ones to:

- consider proper names as concepts of our ontology, which results in non-standard instance-to-concept matching,
- describe the full inflection paradigms for the retrieved names (notably for Polish being a highly inflected language),
- associate names not only with their variants but with derivations as well.

We also inherit Prolexbase's novel idea of synonymy in which a (diachronic, diaphasic or diastratic) change in the point of view on an entity yields a different although synonymous entity (note that e.g. in WordNet synonyms belong to the same synset and thus refer to the same entity). This fact enables a higher quality of proper name translation in that a synonym of a certain type is straightforwardly linked to its equivalent of the same type in another language. Last but not least, ProlexFeeder seems to be the only approach in which the problem of a proper integration of previously existing and newly extracted data (notably by avoiding duplicates) is explicitly addressed. Thus, we truly propose an enrichment of a pre-existing proper name model rather than its extraction from scratch.

Wikipedia is one of the main sources of data for ontology creation and enrichment in the methods discussed above. An opposed point of view is represented within the Text Analysis Conference[14] (TAC) by the *Knowledge Base Population*[15] (KBP) task. In particular, its 2011 mono-lingual (English) and cross-lingual (Chinese-English[16])

[13] We have not found any information about the proportion of truly manually validated Freebase data (as opposed to the initial seeding data, whose validation method is unclear).

[14] `http://www.nist.gov/tac/about/index.html`

[15] `http://www.nist.gov/tac/2013/KBP/index.html`

[16] In 2012 the set of languages has been extended with Spanish. Proceedings and results from this edition are not yet available.

Table 3: Contrastive analysis of approaches dedicated to extraction and enrichment of structured data from Wikipedia

Reference	**Scope**	**Data Sources**	**Target Resource**	**Ontology Mapping Method**	**Ontology Mapping Source Unit**	**Ontology Mapping Target Unit**	**Population Method**	**Popularity Estimation Source**	**New Entry's Linguistic Features**	**Entry Validation Method**	**Languages**	**# entries**
(Bollacker *et al.*, 2007)	unres-tricted	Wikipedia, user's knowledge	Freebase	semi-manual collaborative, concept-to-concept	versatile	Freebase non-hierarchical types	semi-manual collabo-rative	none	none	manual, collabo-rative	English	38M entities 1,180M facts
(Suchanek *et al.*, 2007) (Hoffart *et al.*, 2011)	unres-tricted	Wikipedia, WordNet, GeoNames	YAGO	automatic, concept-to-concept	Wikipedia leaf category, GeoNames classes	WordNet synset	extracting entries & facts	none	synonyms, variants	none	English	10M entities 120M facts
(Bizer *et al.*, 2009) (Mendes *at al.*, 2012)	unres-tricted	Wikipedia	DBpedia	manual, collaborative & automatic, concept-to-concept	Wikipedia infobox template & attribute	DBpedia Ontology class & property	extracting & updating entries	none	variants, terms, themes & grammatical gender	none	24 languages	5.4M
(Nguyen & Cao, 2010)	proper names	Wikipedia	KIM	automatic, concept-to-concept	WordNet synset	Wikipedia article	adding features	none	NA	NA	English	unknown
(Melo & Weikum, 2010)	proper names	Wikipedia WordNets	MENTA	automatic ontology induction			extracting entries & relations from scratch	none	synonyms, variants	none	200 languages	unknown
(Toral *et al.*, 2012)	proper names	Wikipedia	LMF lexicon	automatic, concept-to-concept	WordNet synset	Wikipedia category	extracting entries & relations from scratch	none	links with SIMPLE lexicon for Italian	automatic: capitali-zation rules, salient terms, Web search	English, Spanish, Italian	1M EN, 137K SP, 125K IT
(Fernando & Stevenson, 2012)	nouns	Wikipedia	WordNet	automatic, concept-to-instance	WordNet synset	Wikipedia article	adding untyped relations	none	NA	NA	English	156K relations
Our approach	proper names	Wikipedia, GeoNames, Translatica	Prolexbase	manual concept-to-concept, semi-manual instance-to-concept	Wikipedia infobox template, GeoNames category, instances	Prolexbase type, relation and pivot	adding entries, relations & features	Wikipedia hits	inflection, variation, derivation	manual	Polish, English, French	39K PL, 33K EN, 100K FR

Entity Linking track is partly relevant to our work. In this track, the initial knowledge base (KB) consists of over 800,000 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. This task is similar to the pivot selection process in ProlexFeeder except that the typology is very light, the source languages are not concerned with high morphological variability in texts and entity mapping evidence is found in a corpus rather than in an existing, already structured, ontology. Sample TAC KBP results of the 2011 cross-language entity linking evaluation spread from 0.386 to 0.809 in terms of the B-cubed F-score. Another TAC KBP track is *Slot Filling*. Given an entity name (person or organization), its type, a document in which it appears, its identifier in the KB, and a certain number of slots, the task is to fill these slots with data extracted from the document. This partly resembles the process of populating relations in ProlexFeeder. However, unlike relations in Prolexbase, the KBP track slots are flat labels or values rather than virtual relations to other existing KB nodes. We are aware of no experiments with an application of TAC-KBP-population methods to creating an actual mono- or multi-lingual lexical-semantic resource.

The above state of the art mentions only some major initiatives in creation and enrichment of lexical and semantic resources. 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 our paper. *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.

7 CONCLUSIONS

We have described resources, methods and tools used for an automated enrichment of Prolexbase, a fine-grained high-quality multilingual lexical semantic resource of proper names. Three languages,

Polish, English and French, were studied. The initial data contained mainly French names. New data were extracted mainly from Wikipedia and partly from GeoNames, and their integration with Prolexbase was based on a manual mapping of the three corresponding typologies. Attention was paid to establishing the degree of popularity of names, represented by their automatically pre-calculated frequency value, based in particular on Wikipedia hits of the corresponding entries. The morphological description of Polish names was supported by automatic inflection tools. The results of these preprocessing tasks were fed to ProlexFeeder, which contains two main modules: the pivot mapping, which automatically finds the proper insertion point for a new entry, and the graphical lexicographer's interface, which enables a manual correction and validation of data.

Two main challenges in this automated data integration process are: (i) preserving the uniqueness of concepts, which are represented in Prolexbase by pivots, i.e. pairs of objects and points of view on these objects, (ii) offering a user-friendly and efficient lexicographer's workbench. Our experimental study has shown that over 97% of pivots proposed automatically by ProlexFeeder for the new incoming data are correctly identified. The lexicographer needs about 2 minutes to process an entry in the validation interface. The most challenging subtask is the Polish inflection of foreign names.

Table 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. With respect to the initial contents of Prolexbase, ProlexFeeder allowed us to add about 18,000 new pivots and 19,000 relations, as well as 23,000 Polish, 19,000 English and 15,000 French prolexemes. These new data required a manual workload of about 4 person-months.

Table 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 acessibility relation
All	2,457 (3%)	65,768 (90%)	6,312 (9%)
Most frequent types	celebrity 1,325 (17%)	city 48,110 (100%)	city 2,214 (5%)
	country 390 (45%)	celebrity 7,053 (88%)	region 1,696 (40%)
	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

The Prolexbase data are referenced in the META-SHARE infrastructure[17] and available[18] under the CC BY-SA license[19], i.e. the same as for Wikipedia and GeoNames. We are currently working on their LMF exchange format according to Bouchou and Maurel (2008).

8 PERSPECTIVES

Prolexbase is an open-ended project. Many perspectives exist for Prolexbase itself, for the ProlexFeeder functionalities, and for future applications exploiting the rich Prolexbase model.

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

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

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

8.1 *Data and model evolutions*

Currently we have almost finished the processing of the names estimated as commonly used. This estimation was based on Wikipedia frequency data for 2010, and on GeoNames classification. Since both the contents of these two resources and the popularity of some names evolve, the Prolexbase frequency values deserve updates, possibly based on larger time intervals. Moreover, now, that the morphosyntactic variability of many names (in particular in Polish) has been described via instances, additional evidence of a name's popularity might stem from its corpus frequency, provided that some word sense disambiguation techniques are available.

Note also that only a part of the relations modelled in Prolexbase has been actually dealt with in ProlexFeeder. The remaining linguistic-level relations, notably classifying contexts, are still to be described. Pragmonyms and ergonyms are under-represented and should be completed. Instances are awaiting an intentional description, possibly encompassing both inflection and word formation (creating aliases and derivatives from prolexemes) within the same framework. It should, in an ideal case, be integrated with open state-of-the-art Polish inflection resources such as *PoliMorf*[20].

In order to ensure an even better pivot selection process, matching prolexemes and aliases could be enhanced by approximate string matching and other methods used in related work. Moreover the preprocessing methods might extend the scope of the automatically predicted relations by integrating approaches which exploit the internal structure of infoboxes and mine free text contained in Wikipedia pages.

We also plan to develop a more powerful Prolexbase browser within the ProlexFeeder's user interface. Multi-criteria search, as well as efficient visualisation and navigation facilities would greatly enhance the usability of the tool.

New development is also planned for the Prolexbase model itself. Firstly, a better representation of metonymy is needed. Recall (Section 2.1) that systematic metonymy (e.g. the fact that any city can be seen as a toponym, and anthroponym or a pragmonym) is currently expressed at the conceptual level by the secondary typology. How-

[20] `http://zil.ipipan.waw.pl/PoliMorf`

ever, some types are concerned with metonymy on a large but not systematic basis. For instance many names of buildings can refer to institutions they contain (*Muzeum Narodowe* 'The National Museum') but it is not always the case since a building can contain several institutions (*Pałac Kultury* 'The Palace of Culture').

Important challenges also concern the representation of the internal structure of multi-word proper names, seen as particular cases of multi-word expressions (MWEs). Recent development in applications such as coreference resolution, corpus annotation and parsing show that enhancement in lexicon/grammar interface is needed with respect to MWEs. For instance, the multi-level annotated National Corpus of Polish represents both named entities and syntactic groups as trees (Przepiórkowski *et al.* 2012). Human or automatic annotation of such a corpus can greatly benefit from a rich linguistic resource of proper names such as Prolexbase. However, multi-word names contained in such as resource should possibly already be described as trees that could be reproduced over the relevant occurrences in the corpus. At least two kinds of trees are needed: (i) syntactic parse trees, (ii) semantic trees whose nodes are names embedded in the given name (e.g. $[[\textit{Wydział Teologii}]_{orgName}\ [\textit{Instytutu Katolickiego w [Paryżu]}_{settlement}]_{orgName}]_{orgName}$ '$[[\text{Faculty of Theology}]_{orgName}$ of the $[\text{Catholic Institute in } [\text{Paris}]_{settlement}]_{orgName}]_{orgName}$'). An efficient representation of such trees within Prolexbase is one of our major perspectives.

Finally, linking Prolexbase to other knowledge bases such as DBpedia or YAGO would combine the Semantic Web modelling benefits with advanced natural-language processing-oriented features and allow interlinking Prolexbase with many other data sets.

8.2 *Future Applications*

Named entity recognition tools such as Nerf (cf. Section 5) do not yet manage to fully exploit the richness of an advanced annotation schema like that of the National Corpus of Polish. In particular they currently fail to provide lemmas for the recognized NEs and derivational bases (*Wielka Brytania* 'Great Britain') for the relational adjectives (*angielski* 'English') and inhabitant names (*Anglik* 'Englishman'). Prolexbase relations will allow NER tools to bridge this gap, by offering explicit links between different inflectional and derivational variants of proper

names and their base prolexemes. They may also serve as a training material for establishing lemmas and derivational bases for less popular proper names.

Other possible applications of Prolexbase are to be seen in establishing relations between named entities in corpora. Note that the synonymy between pivots, as well as all lexical relations among prolexemes and instances, allow us to straightforwardly link variants of a proper name, thus providing a reliable resource for coreference resolution. Furthermore, the meronymy and accessibility relations constitute a means of finding and labeling bridging (associative) anaphora.

Prolexbase is now also a good candidate to experiment with an advanced version of the Entity Linking process (cf. Section 6). Instead of linking NEs occurrences to Wikipedia entries we might map them on Prolexbase, which offers a pure taxonomy and an elaborate set of manually validated relations. Thus, we would obtain a high quality Word Sense Disambiguation (Fernando and Stevenson 2012) resource for "kernel" NEs.

The most elaborate use of the fine-grained Prolexbase model is expected in the domain of machine translation of proper names (Graliński *et al.* 2009). The original idea of a conceptual proper name being a pair of a referred object and a point of view on this object allows the user application to provide the most appropriate equivalent (rather than just any equivalent) for a name in other languages. For some names, popular in one language but unknown or inexistent in others, relations like the classifying context or the accessibility context enable explanation-based translations (e.g. *Hanna Gronkiewicz-Waltz* ⇒ *Hanna Gronkiewicz-Waltz, the president of Warsaw*, *blésois* ⇒ *an inhabitant of Blois*).

Other potential applications include: (i) multilingual named entity recognition (NER) (Richman and Schone 2008), (ii) text classification (Kumaran and Allan 2004), (iii) uni- or cross-lingual question answering (Ferrández *et al.* 2007), and (iv) proper name normalization (Jijkoun *et al.* 2008).

9 ACKNOWLEDGEMENTS

We are deeply indebted to three anonymous reviewers for their insightful critical remarks and advice on previous versions of our paper.

They allowed us to increase its quality and to discover new perspectives for future work. We are also grateful to Denis Maurel for having shared his expertise on the Prolexbase ontology, as well as to Jakub Waszczuk for the integration of Prolexbase data in the Nerf system and providing experimental results for named entity recognition in Polish.

This work has been carried out within two projects: (i) *Nekst*[21], funded by the European Regional Development Fund and the Polish Ministry of Science and Higher Education, (ii) CESAR[22] – a European project (CIP-ICT-PSP-271022), part of META-NET.

REFERENCES

Claire AGAFONOV, Thierry GRASS, Denis MAUREL, Nathalie ROSSI-GENSANE, and Agata SAVARY (2006), La traduction multilingue des noms propres dans PROLEX, *Μετα*, 51(4):622–636, les Presses de l'Université de Montréal.

Christian BIZER, Jens LEHMANN, Georgi KOBILAROV, Sören AUER, Christian BECKER, Richard CYGANIAK, and Sebastian HELLMANN (2009), DBpedia – A crystallization point for the Web of Data, *J. Web Sem.*, 7(3):154–165.

Kurt BOLLACKER, Patrick TUFTS, Tomi PIERCE, and Robert COOK (2007), A Platform for Scalable, Collaborative, Structured Information Integration, in *Proceeding of the Sixth International Workshop on Information Integration on the Web*.

Béatrice BOUCHOU and Denis MAUREL (2008), Prolexbase et LMF : vers un standard pour les ressources lexicales sur les noms propres, *TAL*, 49(1):61–88.

Samuel FERNANDO and 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.

Sergio FERRÁNDEZ, Antonio TORAL, Óscar FERRÁNDEZ, Antonio FERRÁNDEZ, and Rafael MUÑOZ (2007), Applying Wikipedia's Multilingual Knowledge to Cross-Lingual Question Answering, in *Proceedings of the 12th International Conference on Applications of Natural Language to Information Systems, NLDB 2007*, volume 4592 of *Lecture Notes in Computer Science*, p. 352–363, Springer.

Filip GRALIŃSKI, Krzysztof JASSEM, and Michał MARCIŃCZUK (2009), An Environment for Named Entity Recognition and Translation, in *Proceedings of the 13th Annual Meeting of the European Association for Machine Translation (EAMT'09)*, p. 88–96, Barcelona.

[21] `http://www.ipipan.waw.pl/nekst/`

[22] `http://www.meta-net.eu/projects/cesar`

Johannes HOFFART, Fabian M. SUCHANEK, Klaus BERBERICH, Edwin LEWIS-KELHAM, Gerard DE MELO, and 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)*, p. 229–232, ACM.

Krzysztof JASSEM (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*, p. 98–105.

Valentin JIJKOUN, Mahboob Alam KHALID, Maarten MARX, and Maarten DE RIJKE (2008), Named entity normalization in user generated content, in *Proceedings of the Second Workshop on Analytics for Noisy Unstructured Text Data, AND 2008*, ACM International Conference Proceeding Series, p. 23–30, ACM.

Cvetana KRSTEV, Duško VITAS, Denis MAUREL, and Mickaël TRAN (2005), Multilingual Ontology of Proper Names, in *Proceedings of Language and Technology Conference (LTC'05), Poznań, Poland*, p. 116–119, Wydawnictwo Poznańskie.

Giridhar KUMARAN and James ALLAN (2004), Text classification and named entities for new event detection, in *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, p. 297–304.

Denis MAUREL (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*, p. 334–338.

Denis MAUREL, Nathalie FRIBURGER, Jean-Yves ANTOINE, Iris ESHKOL-TARAVELLA, and Damien NOUVEL (2011), Cascades de transducteurs autour de la reconnaissance des entités nommées, *Traitement Automatiques des Langues*, 52(1):69–96.

Gerard de MELO and Gerhard WEIKUM (2009), Towards a universal wordnet by learning from combined evidence, in *Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009, Hong Kong, China, November 2-6, 2009*, pp. 513–522, ACM.

Gerard de MELO and 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*, p. 1099–1108, ACM.

Pablo MENDES, Max JAKOB, and 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, and Stelios PIPERIDIS, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, European Language Resources Association (ELRA), Istanbul, Turkey.

George A. MILLER (1995), WordNet: A Lexical Database for English, *Commun. ACM*, 38(11):39–41.

Hien Thang NGUYEN and 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.

Georgios PETASIS, Vangelis KARKALETSIS, Georgios PALIOURAS, Anastasia KRITHARA, and Elias ZAVITSANOS (2011), Ontology Population and Enrichment: State of the Art, in *Knowledge-Driven Multimedia Information Extraction and Ontology Evolution*, volume 6050 of *Lecture Notes in Computer Science*, p. 134–166, Springer.

Adam PRZEPIÓRKOWSKI, Mirosław BAŃKO, Rafał L. GÓRSKI, and Barbara LEWANDOWSKA-TOMASZCZYK, editors (2012), *Narodowy Korpus Języka Polskiego [Eng.: National Corpus of Polish]*, Wydawnictwo Naukowe PWN, Warsaw.

Alexander E. RICHMAN and Patrick SCHONE (2008), Mining Wiki Resources for Multilingual Named Entity Recognition, in Kathleen MCKEOWN, Johanna D. MOORE, Simone TEUFEL, James ALLAN, and Sadaoki FURUI, editors, *ACL*, pp. 1–9, The Association for Computer Linguistics, ISBN 978-1-932432-04-6.

K. SARAVANAN, Monojit CHOUDHURY, Raghavendra UDUPA, and 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, and Stelios PIPERIDIS, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, European Language Resources Association (ELRA), Istanbul, Turkey.

Agata SAVARY, Leszek MANICKI, and Małgorzata BARON (2013), ProlexFeeder – Populating a Multilingual Ontology of Proper Names from Open Sources, Technical Report 306, Laboratoire d'informatique, François Rabelais University of Tours, France.

Pavel SHVAIKO and Jérôme EUZENAT (2013), Ontology Matching: State of the Art and Future Challenges, *IEEE Trans. Knowl. Data Eng.*, 25(1):158–176.

Ralf STEINBERGER, Bruno POULIQUEN, Mijail Alexandrov KABADJOV, Jenya BELYAEVA, and 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*, p. 104–110.

Fabian M. SUCHANEK, Gjergji KASNECI, and 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*, p. 697–706, Banff, Canada.

Antonio TORAL, Sergio FERRÁNDEZ, Monica MONACHINI, and 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.

Antonio TORAL, Rafael MUÑOZ, and Monica MONACHINI (2008), Named Entity WordNet, in *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2008)*, European Language Resources Association, Marrakech, Morocco.

Mickaël TRAN and Denis MAUREL (2006), Prolexbase: Un dictionnaire relatonnel multilingue de noms propres, *Traitement Automatiques des Langues*, 47(3):115–139.

Mickaël TRAN, Denis MAUREL, Duško VITAS, and 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*.

Piek VOSSEN (1998), Introduction to EuroWordNet, *Computers and the Humanities*, 32(2-3):73–89.

This work is licensed under the Creative Commons Attribution 3.0 Unported License.
http://creativecommons.org/licenses/by/3.0/

Grammatical typology and frequency analysis: number availability and number use

Dunstan Brown[1], *Greville G. Corbett*[2], *Sebastian Fedden*[2], *Andrew Hippisley*[3], *and Paul Marriott*[4]
[1] University of York, UK
[2] University of Surrey, UK
[3] University of Kentucky, USA
[4] University of Waterloo, Canada

ABSTRACT

The Smith-Stark hierarchy, a version of the Animacy Hierarchy, offers a typology of the cross-linguistic availability of number. The hierarchy predicts that the availability of number is not arbitrary. For any language, if the expression of plural is available to a noun, it is available to any noun of a semantic category further to the left of the hierarchy. In this article we move one step further by showing that the structure of the hierarchy can be observed in a statistical model of number use in Russian. We also investigate three co-variates: plural preference, pluralia tantum and irregularity effects; these account for an item's behaviour being different than that solely expected from its animacy position.

Keywords: animacy hierarchy, frequency, number, Russian

1 INTRODUCTION

The morphosyntactic feature of number is found in many languages; it has the values singular and plural, and often others too, such as dual. Number distinctions and the availability of number have been generally well-studied cross-linguistically. One of the most important contributions in this area was the Smith-Stark hierarchy (Smith-Stark 1974), discussed in Corbett (2000). This hierarchy, often also called

the Animacy Hierarchy, offers a typology of the availability of number in languages. In this article we move one step further by demonstrating that the structure of the Smith-Stark hierarchy can be observed in the use of the number feature in Russian[1]. The hierarchy we use in this paper, which is adapted from Smith-Stark (1974) is given in (1):

(1) Speaker > Addressee > Kin > Non-human rational > Human rational > Human non-rational > Animate > Concrete inanimate > Abstract inanimate

The labels 'speaker' and 'addressee' are used for the first and second person pronouns. The other positions of the Smith-Stark hierarchy in (1) are universally applicable lexical categories. We also refer to them as the animacy category of a noun. Nouns of the non-human rational category denote supernatural beings. Human rationals include humans except children, which belong in the Human non-rational category. Corbett (2000) points out that the rational/non-rational distinction has limited justification. However, given the typological importance of the Smith-Stark hierarchy, we took the decision only to extend distinctions within the hierarchy rather than eliminate any. We therefore maintained the human rational/non-rational distinction, and we also added a distinction of concrete and abstract within inanimates, which meant that the original structure of the hierarchy is recoverable. The hierarchy predicts that the availability of number is not arbitrarily distributed. For any language, if the expression of plural is available to a noun it is likewise available to any noun of a semantic category towards the left of the hierarchy. For example, if a language has a singular-plural contrast in animate nouns, it will also have such a contrast in human non-rational, human rational, and non-human rational nouns, kin nouns and the second and first person pronouns. In other words, there is a cut-off point somewhere along the hierarchy. Left of this point, plural is available; further down the hierarchy to the right of this point, plural is not available.

[1] The research reported here was originally funded by the ESRC (UK) under grant R000222419. For the time for recent updating, Brown and Corbett are indebted to the European Research Council under grant ERC-2008-AdG-230268 MORPHOLOGY. The support of both funding bodies is gratefully acknowledged. We thank Alexander Krasovitsky for helpful discussion of specific Russian examples.

The Smith-Stark hierarchy is a typological generalization and as such should be valid cross-linguistically. Our hypothesis is that the use of the grammatical category *number* can be predicted from a typology which in turn makes predictions about the availability of number. A necessary way of testing this generalization is to apply it to a test language. Russian was selected since number is (generally) available to nominals, and the rich morphology of Russian typically makes the expression of number clear, as can be shown by the items in (2) which exemplify each of the different points on the hierarchy.

				(2)
ja 'I'	vs.	*my* 'we'	[speaker]	
ty 'you (singular)'	vs.	*vy* 'you (plural)'	[addressee]	
otec 'father'	vs.	*otcy* 'fathers'	[kin]	
bog 'god'	vs.	*bogi* 'gods'	[non-human rational]	
podruga 'girlfriend'	vs.	*podrugi* 'girlfriends'	[human rational]	
rebenok 'child'	vs.	*deti* 'children'	[human non-rational]	
lošad' 'horse'	vs.	*lošadi* 'horses'	[animate]	
stol 'table'	vs.	*stoly* 'tables'	[inanimate]	
sistema 'system'	vs.	*sistemy* 'systems'	[abstract inanimate]	

This article has four sections. In section 2 we give a summary of our methods and the statistical model we used in our study. In section 3 we present the results of our study. We show that there is a relationship between the points in the availability hierarchy and number use, but that other co-variates can come into play that result in a much higher plural proportion than expected from the position on the hierarchy. This is for example the case for nouns whose referents typically come in pairs (*glaz* 'eye') or in multitudes (*gramm* 'gramme'), and for pluralia tantum, such as *rebjatiški* 'kids', i.e., nouns which have only plural forms. Finally, we give our conclusions.

2 METHODS AND STATISTICAL MODEL

In this section we outline the methods used for data preparation and data analysis. We also sketch the statistical model used in this research.

2.1 *Data preparation*

To test our hypotheses, we used the corpus of contemporary Russian texts prepared at Uppsala University, Lönngren (1993), which con-

tains about one million tokens. At the time the research was carried out this was the most suitable corpus of Russian as far as scope and design were concerned, as it covered a range of texts within a 25-year time period (1960–1985).[2]

We prepared the data as follows. Nouns were taken from the corpus and marked for semantic, morphosyntactic, and frequency information. The dataset contains 5,450 noun and pronoun lexemes occurring five or more times, with morphosyntactic and frequency information about their 243,466 word forms. This includes first and second person pronouns, but excludes third person pronouns. The third person deserves a separate study; there are around 29,000 examples of third person pronouns in the corpus. We used the concordance tool 'WordSmith' (Oxford University Press) to extract the nouns from the corpus and we indexed them according to position on the Smith-Stark hierarchy, and recorded number information, i.e., the distribution of singulars and plurals. This information was formatted in Microsoft Excel and encoded in such a way so as to facilitate statistical analysis. In particular we noted for each lexeme the proportion of plural forms being used. Numerical values were given for all information on animacy category, i.e., position on the Smith-Stark hierarchy, case and number. The statistical software package used for data analysis was S-PLUS.

The dataset resulting from our study has been made available on our web site.[3]

2.2 *Statistical model*

A number of differing modelling approaches were used for the analysis. The non-parametric bootstrap (Efron and Tibshirani 1993) was

[2] The offline version of the Russian National Corpus is a similar size (see `http://ruscorpora.ru/corpora-usage.html`), while the online version is much bigger. The semantic categories available for searching the online version should map straightforwardly onto the Smith-Stark hierarchy, but currently it is not possible to download the full results of a search. Replicating our results using the RNC would, of course, be a useful future piece of research. For more on the RNC and its history see Grišina and Plungian (2005). See Maier (1994) for more information on the Uppsala corpus.

[3] `http://www.surrey.ac.uk/englishandlanguages/research/smg/files/rusnoms.xls`

used to test if there was a significant difference between the median values of plural usage between groups, while the two sample Kolmogorov-Smirnov test (Conover 1971) was used to test for differences in distributions of the plural usage, again across pairs of groups defined by the hierarchy. The results from non-parametric approaches were checked using a parametric approach using the log-likelihood for inference. The S-PLUS code for this model and explanatory text has been made available at the Surrey Morphology Group website.[4]

Since the results for the parametric method were qualitatively the same as the non-parametric, only the non-parametric results are reported here.

In order to test the differences between the median values of two groups, the bootstrap, a form of randomisation, was used. We extract a subset of lexemes S from the corpus C according to animacy category. We calculate the median frequency of the distribution of the required frequency. Denote this to be $m(S)$ in the subset S and $m(C)$ in the full corpus, C. We need to see if $m(S)$ is significantly different from $m(C)$ assuming the null hypothesis that there is no relationship between the extraction criterion (animacy category) and the measure quantity (frequency). Under this assumption we can evaluate the distribution of $m(S)$ by randomly selecting (with replacement) samples of equal size to S from C, and calculating their median. This procedure is repeated many times and an estimate of the underlying distribution of the median is constructed. This will be the bootstrap distribution of the median under the assumed hypothesis. The actual value of $m(S)$ can then be compared to this bootstrapped distribution to see if it is extreme. A p-value can then be directly calculated from the bootstrap distribution. For details of this procedure see Efron and Tibshirani (1993), Chapter 13.

Initially, informal graphical methods were used to explore the data before any modelling or formal testing was done. The exploratory data analysis showed observed proportions varying continuously in the range from 0 to 1, but also with appreciable finite atoms of probability at exactly 0 or 1. Hence a mixture model was selected using

[4] `http://www.surrey.ac.uk/englishandlanguages/research/smg/files/statisticalmodel.pdf`

a beta distribution as a continuous model for the interval $(0, 1)$ and with the discrete atoms modelled separately. The model was fitted using maximum likelihood and showed very good agreement with the data.

3 RESULTS AND DISCUSSION

In this section we present the details of the results of our investigation into number use in Russian and discuss those cases in which the proportion of plural forms was much higher than we would expect from the position on the hierarchy.

3.1 *The relation between plural marking and hierarchy position*

We analysed 5,450 Russian noun and pronoun lexemes from the Uppsala corpus according to the methodology outlined in Section 2.1, which were represented by 243,466 word forms. We recorded lexemes for their distribution of singular and plural forms, as well as for their animacy category. The sample details are given in Table 1.

The p-value in the second rightmost column in Table 2 represents the probability that the observed median was due to chance variation computed via the bootstrap. The p-value in the last column is from the Kolmogorov-Smirnov test. There is very strong evidence that there is structure in most of the categories. (A value less than 0.05 is

Table 1: Details of the sample of Russian nouns

Animacy category	Lexeme frequency	Word-form frequency	Word-form proportion of sample (%)
Speaker	1	9,610	3.9
Addressee	2	2,805	1.2
Kin	45	4,155	1.7
Non-human rational	5	267	0.1
Human rational	498	17,127	7.0
Human non-rational	28	2,054	0.8
Animate	102	2,826	1.2
Concrete inanimate	2,437	93,442	38.4
Abstract inanimate	2,332	111,180	45.7
TOTALS	5,450	243,466	100

strong evidence that the group is significantly different from the corpus.) From Table 2 we see that the evidence is less strong for Speaker, Addressee, and Non-human rational. The group Kin was significant using the Kolmogorov-Smirnov comparing distributions.

Table 3 gives the p-values for pairwise tests of equality of distribution across the groups in the hierarchy.

These results give more structure to the patterns shown later in Figure 1. Thus, for example, we see that while the Human non-rational and Animate groups are significantly different from the corpus as a whole (Table 2), they are not different from each other (Table 3). On the other hand, groups at the lower end of the hierarchy are both different from the corpus and different from each other. These results show how the structure of the hierarchy is reflected in the observed distribution of number use. It is clear that the position that a lexeme takes in the Smith-Stark hierarchy can have a strong effect on the proportion of one number (plural) being used over another. We can compare the hierarchy for number availability with the broad picture

Table 2: Details of the sample of Russian nouns

Animacy category	Singular forms	Plural forms	Singular + plural forms	Mean plural proportion	Median plural proportion	p-value Bootstrap	p-value K-S test
Speaker	6197	3413	9610	35.5%	35.5%	0.83	0.75
Addressee	2600	205	2805	8.7%	8.7%	0.43	0.71
Kin	3733	422	4155	14.7%	5%	0.07	<0.001
Non-human rational	248	19	267	5.8%	5.5%	0.46	0.12
Human rational	9392	7735	17127	45.1%	45.5%	< 0.001	< 0.001
Human non-rational	854	1200	2054	58.4%	61.8%	< 0.001	< 0.001
Animate	1599	1227	2826	43.4%	48.1%	< 0.001	< 0.001
Concrete inanimate	65427	28015	93442	30%	23.1%	< 0.001	< 0.001
Abstract inanimate	84698	26482	111180	23.8%	0.5%	< 0.001	< 0.001
TOTALS	174,748	68,718	243,466	28.2%	16.7%		

of the results of our investigation into number use. The Smith-Stark hierarchy is given in (3), repeated from (1) above.

(3) Speaker > Addressee > Kin > Non-human rational > Human rational > Human non-rational > Animate > Concrete inanimate > Abstract inanimate

We have made explicit the distinction between human rational and human non-rational (children), and extended the hierarchy to distinguish inanimates that are concrete from inanimates that are abstract. The classes which distinguish singular and plural occupy the upper segments of the hierarchy, and languages make the split between items distinguishing number and those failing to do so at different points of the hierarchy.

Our investigation into number use yielded statistically significant results. We can compare the version of Smith-Stark's hierarchy for number availability in (3) with the picture of number use in Figure 1.

The data are structured with each animacy position having its own median point. The median is represented by the line in the middle of the box; the box itself represents a range of proportions covering the middle 50% of the lexemes in the category; the whiskers cover the

Table 3: Comparison of pairs of groups in the hierarchy

Animacy category	Addressee	Kin	Non-human rational	Human rational	Human non-rational	Animate	Concrete inanimate	Abstract inanimate
Speaker	0.667	0.422	0.375	0.856	0.820	0.858	0.815	0.567
Addressee	–	1.000	0.867	0.196	0.080	0.179	0.519	0.977
Kin		–	0.906	**<0.001**	**<0.001**	**<0.001**	**<0.001**	0.083
Non-human rational			–	0.003	**<0.001**	**0.005**	**< 0.042**	0.538
Human rational				–	0.416	0.960	**<0.001**	**<0.001**
Human non-rational					–	0.258	**<0.001**	**<0.001**
Animate						–	**<0.001**	**<0.001**
Concrete inanimate							–	**<0.001**

remaining 50%, except potential outliers which are indicated separately with circles (Daly *et al.* 1995). This demonstrates that there is a relationship between the positions in the availability hierarchy and number use.

On the one hand, we might have hoped for a correlation between the positions on the hierarchy and number, and clearly this is not found. This means that the hierarchy which accounts well for number availability across languages does not apply straightforwardly to number use, since Russian appears to be a counterexample. On the other hand, when we compare the medians of the proportion of plural forms for the different animacy categories of Smith-Stark, we see that each lexical category has its own median point (Figure 1). This strongly indicates that at a general level, the hierarchy position to which a lexeme belongs has an impact on the way it will distribute its forms. There is a dramatic difference between groups of nominals. Nouns denoting humans and other animates show the highest proportion of plural use, with concrete and abstract inanimates lower. Moreover, for all positions below non-human rationals the *p*-values are highly significant (Table 2 rightmost column). For the kin and non-human rational cate-

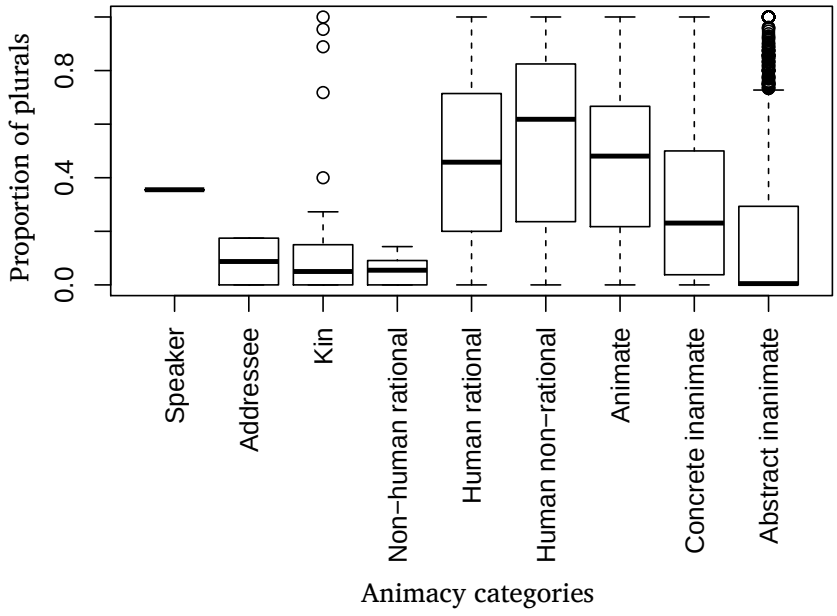


Figure 1: Box plot of proportion of plurals and animacy

gories there are plausible reasons why this might be so. One concerns standard use. As kin terms are often used for addressing individuals, it is reasonable to expect a high proportion of singular forms. Another contributing factor could be the uniqueness of the father and mother kin relations with respect to each individual. For non-human rationals (i.e., god, devil, angel) we expect a higher proportion of singular forms given that Russia's major religion is monotheistic. On the other hand, there is no obvious reason to assume that the pronouns for speech participants would differ in terms of number use.

Another possible explanation for the different structures of availability and use is based on the notion of individuation. When we compare number availability with number use, an interesting picture emerges. If the medians of the proportion of plurals are compared amongst the lexemes belonging to each slot in the hierarchy, as shown in Figure 1, we have a steep hill shape, peaking at the human non-rationals. In other words the left and right edges of the hierarchy have a smaller proportion of plurals, and the middle portion has a significantly higher proportion of plurals. An explanation for the steep hill shape may be based on individuation, running from most individuated (Speaker), to least individuated, to completely non-individuated items (abstract mass nouns). The small proportion of plurals at the bottom of the hierarchy is due to 'individual' plurals being largely unavailable, and only the (rarer) 'sort' and 'container' plurals being available. In this scenario the small proportion of plurals at the top segment of the hierarchy is due to the conceptual difficulty of pluralising highly individuated items. Describing a person using a kin term is individualising him/her further. Pluralising the same person would act to make him/her less individuated. This would explain the lack of plurals in this category.

In sum, the position of a lexeme on the hierarchy has a strong effect on number use. However, further co-variates come into play which account for an item's behaviour being different to that solely expected from its animacy position. We will discuss each of these co-variates, plural preference, pluralia tantum and irregularity effects in turn below.

Table 4: Nouns in the corpus locally unmarked for plural

Example	Example's animacy	Plural proportion	Plural proportion of example's animacy category (median)
roditel' 'parent'	Kin	95%	5%
bliznec 'twin'	Human rational	97%	45.5%
soavtor 'co-author'	Human rational	90%	45.5%
glaz 'eye'	Concrete inanimate	90%	23.1%
botinok 'boot'	Concrete inanimate	88%	23.1%
gramm 'gramme'	Abstract inanimate	81%	0.5%

3.2 *Plural preference*

Some items are naturally 'more plural' regardless of their lexical category. These can be viewed as locally unmarked for plural (Tiersma 1982), for instance items such as *glaz* 'eye' and *bliznec* 'twin' which would be expected to occur in the plural more frequently than the singular because singular contexts are unusual. Table 4 shows how the proportion of plurals for a locally unmarked item was found to be much greater than that expected from its animacy group.[5] Such nouns occur as outliers in our boxplots.

It might be asked why there is no similar section on singular preference. The basic answer is that for a noun to have singular preference is completely normal, as is evident from Table 2 (see column 'Mean plural proportion'), and from cross-linguistic data (see Corbett 2000, p. 281, for data on French, Latin, Sanskrit, Slovene and Upper Sorbian, as well as on Russian). In our count one third of the nouns (almost exactly) occur in the singular only. Note that this does not imply that they are singularia tantum; recall that for inclusion we require that the noun occurs five times or more. It is evident from the list that many nouns which occur five times only, all in the singular, are normal count nouns; they happen not to have occurred in the plural in the corpus.

[5] For further discussion of the semantics of number in Russian, see Ljaševskaja (2004) and references therein.

Table 5: Pluralia tantum in the corpus

Example	Example's animacy	Plural proportion	Plural proportion of example's animacy category (median)
rebjatiški 'kids'	Human non-rational	100%	61.8%
sani 'sledge(s)'	Concrete inanimate	100%	23.1%
brjuki 'trousers'	Concrete inanimate	100%	23.1%
xlopoty 'troubles'	Abstract inanimate	100%	0.5%
sutki '24 hours'	Abstract inanimate	100%	0.5%

3.3 *Pluralia tantum*

Some items lack a means of marking singular; in other words, for them singular is unavailable and they will always appear morphologically plural (even where there is a singular interpretation). Such pluralia tantum are given in Table 5. For example, the noun *sani* 'sledge' is morphologically marked for plural, but can have a singular and a plural reading.

Pluralia tantum are recognizable and are few in number in Russian. On the other hand, genuine singularia tantum are hard to identify; while many nouns normally occur in the singular, there are possibilities for recategorization: that is, they may be recategorized with unit reading or with instance reading (see Corbett 2000, pp 81–82, 84–87, for discussion). To illustrate the instance reading, we may take *mnogo raznyx vin* 'many different wines', where different types of wine are intended. The key point is that while such recategorizations are visible in the plural, the recategorization from mass to count gives a singular form too, hence *odno očen' xorošee vino* 'one very good wine'. This recategorized singular is not distinct from the normal singular.

3.4 *Irregularity effects*

There is a third important co-variate. In certain instances irregularity can affect the distribution of plurals. To appreciate this, it is important to distinguish absolute counting (the straightforward count of items in the corpus) from relative counting (the relation of forms within a lexeme; in our study this is plural versus singular). Irregularity in a

lexeme is correlated with a high occurrence of plurals of that lexeme in the corpus.

Corbett *et al.* (2001) demonstrate for Russian that there is a relation between irregularity in noun lexemes and absolute plural anomaly, i.e., a high absolute number of plural forms in the corpus, and that there is a relation between non-prosodic irregularity (where irregularity is not confined to stress placement), and relative plural anomaly, i.e., a high proportion of plural forms compared to forms in the singular. This means that irregular Russian nouns in general have a high number of plural forms in the corpus. Prosodic irregularity means that there is also a high number of singular forms to match the plural ones (hence no relative plural anomaly), whereas nouns which display segmental irregularity have a higher proportion of plural forms in comparison with singular forms (hence high relative plural anomaly).

In sum, these three types of co-variate (plural preference, pluralia tantum, and irregularity effects) broadly account for the plural outliers in Figure 1.

4 CONCLUSIONS

Typology is typically concerned with the availability of a feature in a language. The special interest of our contribution lies in juxtaposing questions of availability with those of actual use. One hypothesis about the relationship between number use in one language (here Russian) and its relationship with the hierarchy of number availability is that there should be a correlation, a strictly linear relationship where those categories furthest left in the hierarchy show the greatest median plural proportion, with this proportion decreasing as we move rightward along the hierarchy. However, this hypothesis must be rejected. The reality is perhaps more interesting: we have good evidence that the middle part of the hierarchy shows the highest plural proportions of usage, with a consistent decrease in plural proportions as we move rightward from the human rationals to the abstract inanimates. We are in a position to say that this is significant. For the top end of the hierarchy there is less that can be said with certainty, given the lack of significance for certain of the higher positions. If anything our results point to the difference between the pronoun proportion of the

hierarchy (where the results are not significant) and the nominal proportion (where the results are significant). Something that is worthy of further investigation is the question of why the human (rational and non-rational) part of the hierarchy has the highest proportions, compared to animates and concrete inanimates. Further investigation would enable us to decide between two different theories about the way the hierarchy partitions the semantics of plural in use. In one theory, associative readings, 'normal' readings and recategorization effects partition the hierarchy, and the observation of high plural occurrence in the middle of the hierarchy is evidence for the high frequency of 'normal' readings associated with this part of the hierarchy. An alternative theory is that plural usage in the middle of the hierarchy is a reflection of the fact that it can have multiple plural semantics available to it (rather than just the 'normal' readings), and these multiple possibilities are reflected in greater use. While the first of these theories is the more plausible, we have no evidence yet to decide between them. Our research has therefore suggested a new programme of future research to investigate this matter in greater depth.

Our examination of the category of number in a language where nouns typically mark number has shown that the typology proposed by Smith-Stark for number availability has a partial analogue for number use. In other words, we have shown that answers to questions about availability can be reflected in use.

REFERENCES

W.J. CONOVER (1971), *Practical Nonparametric Statistics*, John Wiley & Sons, New York.

Greville G. CORBETT (2000), *Number*, Cambridge University Press, Cambridge.

Greville G. CORBETT, Andrew HIPPISLEY, Dunstan BROWN, and Paul MARRIOTT (2001), Frequency, regularity and the paradigm: a perspective from Russian on a complex relation, in Joan BYBEE and Paul HOPPER, editors, *Frequency and the Emergence of Linguistic Structure*, pp. 201–226, John Benjamins, Amsterdam.

Fergus DALY, David HAND, Chris JONES, Daniel LUNN, and Kevin MCCONWAY (1995), *Elements of statistics*, Addison-Wesley, London.

Bradley EFRON and Robert J. TIBSHIRANI (1993), *An introduction to the bootstrap*, Chapman and Hall, London.

Elena A. GRIŠINA and Vladimir A. PLUNGIAN (2005), Perspektivy razvitija Natsional'nogo korpusa russkogo jazyka [Prospects for developing a national corpus of Russian], *Indrik*, http://ruscorpora.ru/sbornik2005/19grishina.pdf.

Ol'ga N. LJAŠEVSKAJA (2004), *Semantika russkogo čisla [The semantics of Russian number]*, Jazyki Slavjanskoj Kul'tury, Moscow.

Lennart LÖNNGREN (1993), *Častotnyj slovar' sovremennogo russkogo jazyka [A frequency dictionary of contemporary Russian] (Acta Universitatis Upsaliensis, Studia Slavica Upsaliensis 33)*, University of Uppsala, Uppsala.

Ingrid MAIER (1994), Review of Lennart Lönngren (ed.), Častotnyj slovar' sovremennogo russkogo jazyka, *Rusistika Segodnja*, 1:130–136.

T. Cedric SMITH-STARK (1974), The plurality split, in Michael W. LA GALY, Robert A. FOX, and Anthony BRUCK, editors, *Papers from the Tenth Regional Meeting, Chicago Linguistic Society*, pp. 657–671, Chicago: Chicago Linguistic Society.

Peter M. TIERSMA (1982), Local and general markedness, *Language*, 58:832–849.

This work is licensed under the Creative Commons Attribution 3.0 Unported License.
http://creativecommons.org/licenses/by/3.0/


Reactivation of antecedents by overt versus null pronouns: Evidence from Persian

Niloofar Keshtiari[1] *and Shravan Vasishth*[2]
[1] Cluster of Languages of Emotion, Freie Universität Berlin, Germany
[2] Department of Linguistics, University of Potsdam, Germany
and School of Mathematics and Statistics, University of Sheffield, UK

ABSTRACT

In Persian, a construction exists in which a gap can optionally be replaced by an overt pronoun. A self-paced reading study (110 participants) suggests that the overt pronoun results in deeper encoding (higher activation) of the antecedent noun, presumably because of richer retrieval cue specifications during antecedent retrieval at the pronoun; this higher activation has the consequence that the antecedent is easier to retrieve at a subsequent stage. This provides new evidence for reactivation effects of the type assumed in the cue-based retrieval model of parsing (Lewis and Vasishth 2005), and shows that dependency resolution is not simply a matter of connecting two co-dependents; the retrieval cue specification has a differential impact on processing.

Keywords: sentence comprehension, pronoun processing, reactivation, expectation-based processing

1 INTRODUCTION

It is well known that both overt and null pronouns render their antecedents more active (more salient) in memory (MacDonald 1989; Emmorey and Lillo-Martin 1995). One way to characterize the underlying processes in antecedent–pronoun/gap resolution is in terms of the ACT-R based (Anderson *et al.* 2004) architecture of sentence processing discussed in (Lewis and Vasishth 2005; Lewis *et al.* 2006). The computational model developed in these papers has been widely ap-

plied in the study of various phenomena in psycholinguistics (Vasishth *et al.* 2008; Vasishth and Lewis 2006; Patil *et al.* 2012; Reitter *et al.* 2011; Dillon 2011; Boston *et al.* 2011; Patil *et al.* 2013; Dillon *et al.* 2011; Engelmann *et al.* 2013).

A central assumption in the ACT-R architecture is that, in any information-processing task, memory representations must be associated with each other in order to build a mental representation that allows the task to be carried out. In the context of sentence comprehension, the primary events that are modeled are structure building and dependency resolution. All other things being equal, the speed with which a memory representation can be accessed depends on its activation level (this is an abstract, unitless quantity) and on the retrieval cues (these are essentially feature–value matrices) that guide access. Generally speaking, the higher the activation, the faster the retrieval. For example, when the parser encounters a reflexive like *himself*, an antecedent noun may be searched using the fact that the antecedent must c-command the reflexive (one retrieval cue; see Dillon *et al.* 2011) or using an additional cue, here gender (Patil *et al.* 2012). Activation of memory representations is assumed to be undergoing constant decay; this models forgetting over time. An assumption in ACT-R is that decay can be counteracted by a process of reactivation: every retrieval event is assumed to increase the activation of the item retrieved. Such an increase in activation has the obvious consequence of facilitating subsequent retrieval (unless enough time goes by such that decay levels out the activation). Previous work has addressed some of the empirical consequences of this theoretical assumption. For example, in Hindi, processing a verb in a relative clause has been argued to be easier when the relative clause is long vs. short; under the assumption that a long relative clause repeatedly accesses and modifies the head noun and does so more often than a shorter relative clause, we expect a faster reading time when the head noun must be accessed, for example, while processing the verb of the relative clause. This was one of the arguments presented by Vasishth and Lewis (2006) in order to explain faster reading times observed at the verb of the relative clause in long vs. short relative clauses (cf. Levy 2008; Husain *et al.* 2013).

Given such an architecture, it is reasonable to assume that completing a dependency between an antecedent and a pronoun, or be-

tween an antecedent and a gap, will increase activation of the antecedent, making subsequent retrieval easier; this assumes that the decay component has not had enough time to counteract the effect of such an activation increase. Indeed, Bever and McElree (1988) have shown experimentally that such reactivation occurs with gaps and pronouns, and that "gaps access their antecedents during comprehension in the same way as pronouns."

Persian presents an interesting construction in this context. Sentences such as in Example (1) have the property that the first gap can optionally be an overt or null pronoun (Taghvaipour 2004). For example, consider (1a); here, two gaps are present. As shown in (1b), the first one can be replaced with the pronoun *un ro*, 'it DOM').[1]

(1) a. Nazanin [in ketabcha ro]$_i$ [$_{CP}$ ghablaz inke gap_i
Nazanin this booklet DOM before that gap
be-khun-eh] gap_i be man dad, dorost hamoon
prefix-read-3SG gap to me gave.3SG, just that
moghe ke kelas ha tamoom shod.
moment that class PL finish became.3SG.
(Lit.) 'Nazanin gave me this little book before reading (it), when the classes finished.'

b. Nazanin [in ketabcha ro]$_i$ [$_{CP}$ ghablaz inke **unro**
Nazanin this booklet DOM before that it-DOM
be-khun-eh] gap_i be man dad, dorost hamoon
prefix-read-3SG to me gave.3SG, just that moment
moghe ke kelas ha tamoom shod.
that class PL finish became.3SG.
'Nazanin gave me this little book before reading it, when the classes finished.'

This construction is interesting in the context of reactivation effects in parsing because it allows us to investigate whether there is a difference in activation increase due to antecedent–pronoun vs. antecedent–gap dependency resolution. Our study was motivated by the speculation that there might be a difference in the way a anteced-

[1] Abbreviations used are as follows. DOM: direct object marking; 3SG: third singular; PL: plural.

ent–gap and antecedent–pronoun dependency is completed: the antecedent might be activated to a greater extent in the antecedent–pronoun case vs. the antecedent–gap case. This could happen because the pronoun uses a richer set of cues; for example, pronouns provide number information, whereas gaps do not. Another possibility is that the pronoun may focus the antecedent (thereby encoding it more richly) in a way that the gap does not. These two explanations may be related: a richer set of cues would lead to better encoding due to the greater extent of activation increase.

We investigated whether we could find any evidence for differential amounts of activation increase in the above construction. We employed the self-paced reading methodology (Just *et al.* 1982) described below. In (1a), the word *ketabche*,[2] 'booklet', is co-indexed with a gap; this gap presumably activates the antecedent once the dependency is completed. Due to reactivation effects, the activation increase of *ketabche* should increase the speed or rate at which its retrieval is completed subsequently at the verb *dad*, 'gave.' Example (1b) is identical except that instead of the gap we have an overt pronoun *unro*. Our speculation was that this might boost activation of *ketabche* to a greater extent than the gap, facilitating retrieval at the verb.[3]

In order to understand the role of the overt pronoun, we compared sentences with null and overt pronouns, as shown in (2a,b) (the front-slashes in the examples represent the partitioning of the segments in the self-paced reading task; this is described below). We were also interested in exploring, in the same experiment, a related kind of reactivation effect: modification of the noun *ketabche*, 'booklet', by a relative clause. As mentioned above, it has been argued (Vasishth and Lewis 2006; Hofmeister 2011; Vasishth *et al.* 2012) that modification of a noun increases its activation, making subsequent retrieval easier (cf. Levy 2008 for an alternative explanation in terms of expectations).

[2] In (1a,b), the direct object marker *ro* induces a sound change on the word it modifies, changing *ketabche* to *ketabcha*.

[3] Note that the exact location of the gap before the verb *dad*, 'gave' is not important; even assuming that there is a gap there is not necessary. All that we need to assume is that a dependency must be completed between the noun *ketabche* and the verb in order to determine who did what to whom. This assumption of a dependency resolution requirement is well-motivated by previous work; see, e.g., Gibson (2000); Bartek *et al.* (2011); Vasishth *et al.* (2008).

It follows that relative clause modification should also increase activation of the noun, resulting in faster retrieval of the noun at the verb. We were interested in determining whether we find facilitation at the verb due to relative clause interposition (2c,d); if yes, this would provide new evidence for the proposal in the literature that modification increases activation of the modified element.

An alternative possibility is greater processing difficulty at the verb in the relative clause conditions; this effect has been found by Grodner and Gibson (2005) for English (also see Bartek *et al.* 2011), and could be a consequence of increased distance between the verb and its argument(s). Such locality effects can be explained in terms of distance as defined in the Dependency Locality Theory (Gibson 2000) or in terms of decay and interference (Lewis and Vasishth 2005; Van Dyke and McElree 2006).

(2) a. Nazanin / [in ketabcha ro]$_i$ / [$_{CP}$ ghablaz inke /
Nazanin this booklet DOM before that
gap_i be-khun-eh] / gap_i be man dad /,
gap prefix-read-3SG gap to me gave.3SG ,
dorost hamoon moghe ke / kelas ha / tamoom
just that moment that class PL finish
shod /.
became.3SG .

‘Nazanin gave me this little book before reading (it), when the classes finished.’

b. Nazanin / [in ketabcha ro]$_i$ / [$_{CP}$ ghablaz inke /
Nazanin this booklet DOM before that
unro be-khun-eh] / gap_i be man dad /,
it-DOM prefix-read-3SG gap to me gave.3SG ,
dorost hamoon moghe ke / kelas ha / tamoom
just that moment that class PL finish
shod /.
became.3SG .

’Nazanin gave me this little book before reading it, when the classes finished.’

c. Nazanin / [in ketabcha ro /]$_i$ [**ke hafte pish** /
Nazanin this booklet DOM that week last

kharid-eh bood /] ghablaz inke / gap_i
bought-3SG.PC was before this gap
be-khun-eh / gap_i be man dad / , dorost
prefix-read-3SG gap to me gave.3SG , just
hamoon moghe ke / kelas ha / tamoom
that moment that class PL finish
shod / .
became.3SG .

'Nazanin gave me this little book which she has bought last week, before reading (it), when the classes finished.'

d. Nazanin / [in ketabcha ro]$_i$ / [**ke hafte pish** /
Nazanin this booklet DOM that week last
kharid-eh bood] / ghablaz inke / **unro**
bought-3SG.PC was before that it-DOM
be-khun-eh / gap_i be man dad / , dorost
prefix-read-3SG gap to me gave.3SG , just
hamoon moghe ke / kelas ha / tamoom
that moment that class PL finish
shod / .
became.3SG .

'Nazanin gave me this little book which she has bought last week, before reading it, when the classes finished.'

Thus, our predictions are: the presence of overt pronoun interposition should result in a facilitation in processing at the main verb (compared to the conditions where a gap is present); modifying the antecedent with a relative clause could show a facilitation due to reactivation, or increased difficulty due to locality effects. We had no predictions about whether there would be an interaction between the pronoun/gap and relative clause factors. The results of the study are presented next.

2 EXPERIMENT

2.1 *Method: Participants, stimuli and fillers, procedure*

One hundred and ten native speakers of Persian, all living in Tehran, participated in the experiment in August and September 2009. Since

we had no access to a laboratory in Tehran, the first author visited the participants at their homes and carried out the experiment there. Participants were asked to complete a questionnaire on their educational background, language background, and average reading time per day. The questionnaire and items are available from the second author. Participants' ages ranged from 18 to 75 years, with mean age 34.6 years. Each participant was paid the Iranian-Rial equivalent of five Euros.

A total of 161 Persian sentences (5 examples, 60 fillers, 96 stimulus sentences) were prepared by the first author. The 96 stimulus sentences were designed as follows: Following standard experimental methodology for repeated measures (within-subjects) designs, twenty-four stimuli sentences were prepared, and four versions of each sentence were constructed; these correspond to the four conditions in the experiment (see Example 2). Each version of the twenty-four sentences was assigned to one of four lists; that is, each list contains only one of the four versions of a stimulus sentence. Because each participant is shown items from only one list, they read (apart from the fillers and examples) a total of twenty-four target sentences, each representing one condition in the experiment design. This has the consequence that, for example, a subject exposed to List 1 would see Sentences 1a, 2b, 3c, 4d, 5a,…; and a subject exposed to List 2 would see Sentences 1b, 2c, 3d, 4a, 5b,…This partitioning into lists is commonly referred to as counterbalancing and serves to minimize bias introduced by any one stimulus sentence.

Thus, each participant saw $5 + 60 + 24 = 89$ sentences. Because reading time generally increases at the end of a sentence (so-called sentence final wrap-up effects, thought to reflect higher-level integration processes that are triggered after a sentence is read), we added an adverbial phrase to the end of each of the stimuli sentences. In addition, at the end of each sentence a period was presented after pressing the space bar as a separate final segment. This extra material at the end of the sentence makes it less likely that our critical region (the verb of the main clause) is contaminated by higher-level sentence-final processing effects.

The experiment began by the first author explaining the task to each participant verbally; then, the five practice sentences were presented, and these were followed by the actual experiment (fillers and

stimulus sentences, pseudo-randomly ordered). Participants pressed the spacebar (marked with a star) to reveal each successive segment; every time the space bar was pressed, the previous segment would disappear and the next segment would appear in the center of the screen. The time the participants spent reading each segment was recorded as the time between key presses. The segmentation is shown in Example 2 and in the items file provided as supplementary material with this paper.

The experiment was run using Linger version 2.88 by Douglas Rohde on a laptop.[4] Participants were asked to read at a pace that was normal for them. A true/false statement was presented after each sentence; this was meant to ensure that subjects were attending to the sentences and not just pressing the space bar without reading. In order to prevent subjects from developing a strategy for answering the true/false statements without completely parsing the sentence, the statements were directed at every part of the previous sentence, including the noun and the verb of both the main clause and the relative clause. These true/false statements were balanced in their yes–no responses. No feedback was given for correct/incorrect responses. For examples, see the items file provided as supplementary information with this paper.

Participants took approximately 30 minutes to complete the experiment. Reading time at the verb of the main clause (in milliseconds) was taken as a measure of relative momentary processing difficulty. In the following section, the results of the study are reported and discussed.

2.2 *Results*

We fit linear mixed models using the package `lme4` (Bates and Sarkar 2007) in the R programming environment (R Development Core Team 2006); see the appendix for some background on the statistical models used here. The reader unfamiliar with psycholinguistic data analysis methods would benefit from reading the appendix before proceeding with the present section. For the critical analyses at the verb and the region following it, we used JAGS (Plummer 2010) to also fit a hierarchical Bayesian model (a linear mixed model) using non-informative

[4] See `http://tedlab.mit.edu/~dr/Linger/`.

priors. For brevity, the Bayesian analysis is omitted from the main paper, but is included in the supplementary material.

The response variables were response accuracy and response time, and reading times at the verb of the main clause (hereafter, critical region) and the spillover region (hereafter, post-critical region).

Due to an error in the design, nine items (labeled 7, 12, 14, 15, 17, 20, 22, 23, 24 in the supplementary material) were removed from the analysis. In these items, the argument that was co-indexed with the pronoun/gap was either not modified by the relative clause, or the pronoun was not co-indexed with the correct antecedent. This reduced the original data by about 40%.

The response time and reading time data (both in milliseconds) were transformed to a negative reciprocal ($-\frac{1000}{rt}$) in order to stabilize variance; the choice of transform was determined using the Box-Cox procedure (Box and Cox 1964; Venables and Ripley 2002). The reciprocal transform converts speed to rate; see (Kliegl *et al.* 2010) for further discussion on the use of this transform for reading time data.

In the reading times, the transform revealed some extremely fast values (0.7% of the data) that dramatically affected the residuals; these were a few values that were 200–250 ms long. Although such reading times cannot in general be categorized as "too fast", in the context of the present experiment they are not representative of the reading time distributions (based on our experience in our own lab and elsewhere, in languages like Hindi, German, and Japanese, we also see remarkably slow reading times compared to English). These extreme values were removed in the final analysis.

All data and R code associated with the analyses presented here are provided as supplementary material with this paper.

2.2.1 Response accuracy and response time

Response accuracy and response time and their analyses are summarized in Tables 1 and 2. For accuracy, a generalized linear mixed model was fit using the binomial link function to evaluate the effect of pronoun (pron), the effect of relative clause insertion (RC), and the interaction of these two factors (see the appendix for more detail on generalized linear models). A standard ANOVA contrast coding was used: the factor pron was coded -0.5 for the gapped conditions (Ex-

amples 2a,c) and 0.5 for the pronoun conditions (Examples 2b,d); the factor RC was coded −0.5 for the −RC conditions (Examples 2a,b), and 0.5 for the +RC conditions (Examples 2c,d). Items and participants were included as crossed random factors (crossed varying intercepts; see Appendix for discussion). Consistent with the predictions of the locality-based accounts discussed earlier, we found significantly lower accuracies in the conditions where the relative clause was present; these conditions also had longer response times. No other effect reached statistical significance.

Table 1: Mean question accuracy (percentages) and negative reciprocal response time (abbreviated as response rate)

	−RC gap	−RC pronoun	+RC gap	+RC pronoun
accuracy	87	86	78	81
response rate	−0.22	−0.22	−0.21	−0.21

Table 2: Summary of the effects of pronoun (pron), relative clause insertion (RC), and the pron×RC interaction on response accuracy

contrast	coef	se	z	p
pron	0.04	0.14	0.29	n.s.
RC	−0.48	0.15	−3.3*	<0.01
pron×RC	0.12	0.15	0.85	n.s.

Table 3: Summary of the effects of pronoun (pron), relative clause insertion (RC), and the pron×RC interaction on negative reciprocal response time

contrast	coef	se	t
pron	−0.00	0.004	−0.9
RC	0.02	0.004	3.7*
pron×RC	−0.00	0.004	−0.6

2.2.2 Analyses of reading times

The negative mean reciprocal reading times ($-s^{-1}$) with 95% confidence intervals are summarized in Figure 1. The results of the statistical analyses are shown in Table 4. The linear mixed models had varying intercepts and slopes for subject and item, and varying slopes by subject for pron, RC, and the pron×RC interaction.

Analyses at the critical region (the main verb) showed a marginally significant main effect of pron: the overt pronoun resulted in faster

region	contrast	coef	se	t
critical	pron	−0.05	0.03	−1.89
	RC	−0.04	0.03	−1.58
	pron×RC	−0.03	0.03	−1.02
post-critical	pron	−0.07	0.03	−2.62*
	RC	0.008	0.03	0.30
	pron×RC	−0.03	0.03	−1.10

Table 4:
Summary of planned comparisons in the linear mixed models analysis

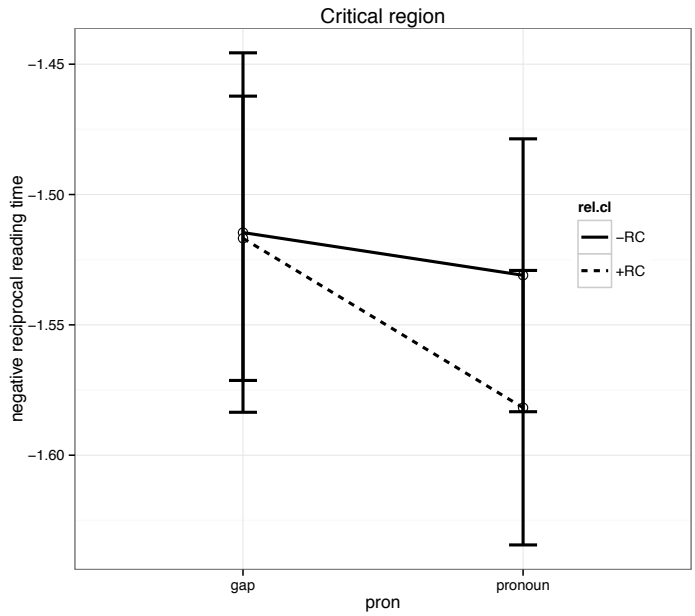
reading times at the verb, as predicted by the reactivation hypothesis. In addition, there was only a marginal effect of relative clause interposition: reading time (rather, reading rate) was marginally faster at the verb in the RC conditions. The interaction between the factors pron and RC did not reach statistical significance either. Figure 1 suggests that the marginally significant facilitation at the critical region due to the overt pronoun is driven by the RC conditions (2c,d). This was confirmed in a post-hoc analysis where the effect of pronoun was investigated within the −RC and +RC conditions. Table 5 summarizes these analyses (the third contrast in Table 5, the effect of RC, is redundant since this was already investigated in our planned comparisons shown in Table 4, but is included because we wanted to use all three degrees of freedom available for parameter estimation of fixed effects).

The post-critical region showed an effect of pron: reading rate was faster when the pronoun was present. The post-hoc analysis showed that the effect of pron was present in both the −RC and the +RC conditions with approximately the same magnitude.[5]

2.2.3 Discussion

To summarize the results for response accuracy and response time, we find lower accuracies and longer response times when the relative

[5] As an aside, note that in psycholinguistics analyses are conventionally carried out on raw reading times. Had we done this conventional analysis for the present data, we would have reported null results. However, in the model based on such untransformed data, the normality assumption for residuals and the homoscedasticity assumption are not met; this would make the model output based on raw reading times meaningless for purposes of statistical inference. The source code provided with this paper gives more detail.



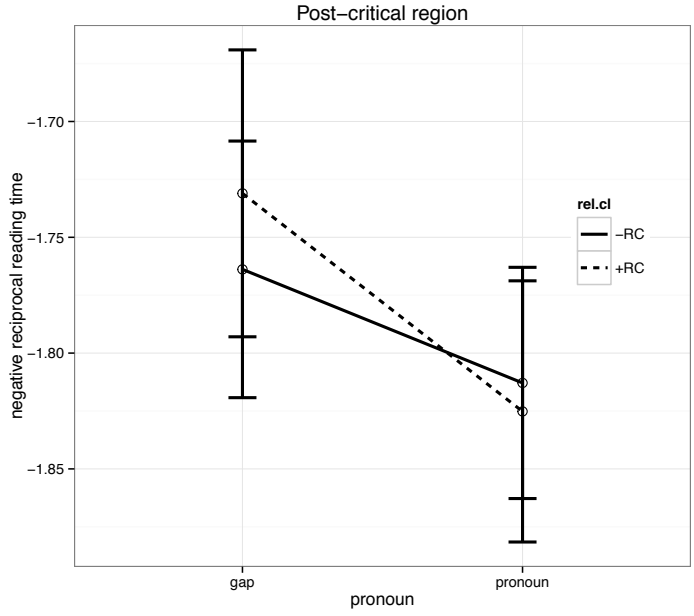


Figure 1: Negative reciprocal reading times at the critical and post-critical regions

region	contrast	coef	se	t
critical	pron (−RC)	−0.02	0.04	−0.63
	pron (+RC)	−0.08	0.04	−2.03*
	RC	−0.04	0.03	−1.58
post-critical	pron (−RC)	−0.04	0.04	−0.94
	pron (+RC)	−0.10	0.04	−2.43*
	RC	0.008	0.03	0.3

Table 5:
Summary of post-hoc nested contrasts at the critical and post-critical regions

clause is present. Response accuracy and response time showed no effect of pronoun and no interaction between the pronoun and relative clause conditions. In the reading time data, at the post-critical region (the word following the verb), reading rate is faster if the pronoun (*unro*) is present. A post-hoc analysis revealed that the facilitation due to the pronoun was driven by the relative clause conditions. At the verb, there is a slight facilitation due to relative clause interposition, but this does not reach statistical significance. No interactions were found.

Although response accuracy and response time are of secondary interest to the research question, they do show that interposing a relative clause renders the sentence more difficult to process in later stages of processing. This finding is partly consistent with locality accounts such as Dependency Locality Theory (Gibson 2000) and the cue-based retrieval architecture (Lewis and Vasishth 2005), both of which predict increased integration cost at the verb when the distance between the subject of the sentence (*Nazanin*) and the main verb (*dad*) is increased. This increased distance (or increased syntactic complexity) could plausibly make it more difficult to retain an accurate representation of the sentence meaning in memory in order to respond to the question.

The Dependency Locality Theory and the cue-based retrieval account, however, also predict a slowdown at the verb in reading times; this prediction turns out to be incorrect because relative clause interposition at the verb results in a marginal facilitation. This tendency towards a facilitation makes sense given prior findings; it can be explained in terms of reactivation due to relative clause modification (Vasishth and Lewis 2006; Hofmeister 2011; Vasishth *et al.* 2012), as

discussed earlier. An alternative explanation for the relative-clause facilitation effect lies in expectation-based processing (Levy 2008). Assuming that treebank-based distributions in Persian turn out to be similar to the distributions in languages that Levy examined, the proposal would be that the expectation for a verb gets stronger and stronger if the appearance of the verb is delayed – this is the situation when the relative clause is interposed. In the relative clause conditions, by the time the verb is encountered, it is highly expected compared to the non-relative clause conditions. This expectation-based account has been proposed as an alternative to the reactivation account. Both explanations are plausible, but the expectation-based account's key prediction has been falsified by Levy *et al.* (2013): they showed in a series of experiments that in Russian relative clauses, if the verb's appearance is delayed inside a relative clause, there is a slow-down at the verb, not a speed-up. The evidence from Russian relative clauses is therefore strongly in favor of locality based explanations. In any case, in our study, the reactivation-based explanation seems more plausible, as discussed in connection with the pronoun results below.

Next, we turn to the main research question in this paper, the effect of the pronoun/gap manipulation. At the verb and the region following it, the pronoun conditions have a faster reading rate than the gap conditions. This suggests that completing the antecedent–pronoun dependency results in higher activation of the antecedent compared to the antecedent–gap case; as a consequence, the antecedent of the pronoun is retrieved faster at the verb. This facilitation probably spills over to the word following the verb. In sum, the data are consistent with our original speculation: replacing a gap with a pronoun appears to increase the activation of the antecedent, making it easier to retrieve at a subsequent stage.

Why was the facilitation effect due to the pronoun driven by the relative clause conditions? In the non-relative clause conditions, even though the pronoun may be activating the antecedent more than the gap, decay of the antecedent noun might be setting in by the time the verb is encountered. By contrast, in the relative clause conditions, the reactivation of the antecedent noun by the relative clause (which modifies this noun) could be providing a counteracting activation boost that reverses the effect of decay. If this is correct, the reactivation account may be the correct explanation for the relative clause facilita-

tion effect discussed above. Naturally, this conclusion does not challenge the expectation-based explanation per se, which probably also plays a role in sentence processing; there is considerable evidence for expectation-based effects (see, e.g., Husain *et al.* (2013) for new evidence from Hindi), and these effects cannot be explained in terms of reactivation.

Returning to our main finding, that the pronoun increases activation of the antecedent, our results raise the question: what is it about antecedent–pronoun vs. antecedent–gap dependencies that results in a more robust encoding or higher activation of the antecedent? One explanation may be that the pronoun may be focusing the antecedent; another may be that a richer set of retrieval cues is used to complete the antecedent–pronoun dependency. We do not have a clear answer for the underlying reason; but given the present data, it seems clear that pronouns activate the antecedent to a greater degree than gaps do.

Why is it that accuracy scores are lower for the relative clause conditions, but in reading times the relative clause conditions result in a (marginal) facilitation at the verb? The former supports the locality account but the latter directly contradicts it. One possible explanation lies in the relative timing of retrieval and expectation effects: locality costs, which reflect retrieval difficulty, may be appearing at a later stage during parsing, whereas expectation-based effects appear earlier. Vasishth and Drenhaus (2011) have proposed this in the context of German. One problem with this account is that locality effects have been found in reading times in English (Grodner and Gibson 2005; Bartek *et al.* 2011) and most recently Hindi (Husain *et al.* 2013) and Hungarian (Kovács and Vasishth 2013). Perhaps a more plausible explanation is that locality effects are longer lasting than expectation effects: the former show effects in online as well as offline measures, but expectation only shows effects in online measures. Under this view, locality effects could have been masked by or are much weaker than reactivation and/or expectation effects while processing the verb; in the offline question–response stage, only locality effects remain visible. All the above explanations are speculative and need to be investigated in new studies pitting locality against expectation.

A broader consequence of the pronoun-driven facilitation we report is that the notion of dependency resolution in parsing needs to

be made more precise. It is widely assumed, implicitly or explicitly, that dependency resolution is simply a matter of connection to elements subject to certain constraints, such as locality (Gibson 2000). But completing an antecedent–gap dependency and an antecedent–pronoun dependency cannot be only a function of locality; it matters which retrieval cues are deployed in retrieval. This has important implications for theories of parsing: an architecture driven by retrieval cues seems to be better motivated than one that ignores the nature of the cue.

3 AUTHOR NOTE

We are grateful to Felix Engelmann for assisting with setting up the software for running the experiments, and to Philip Hofmeister for comments on a draft. Colin Philips and Brian Dillon provided very detailed and useful comments on the paper and have contributed to improving it considerably. We are also grateful to the anonymous reviewers for their careful, high quality reviews. Dr. Jeremy Oakley of the School of Mathematics and Statistics, University of Sheffield, provided very useful guidance on fitting linear mixed models and Bayesian models, but we are of course responsible for any errors in this paper. This paper is based on the master's thesis of Niloofar Keshtiari in the MSc program European Masters in Clinical Linguistics, University of Potsdam. The experiment design is due to Shravan Vasishth; the experiment items were prepared by Niloofar Keshtiari and the experiment was conducted by her in Tehran. The data analysis and the majority of the text in the paper (except the materials and methods section) were written by Shravan Vasishth. The corresponding author for this paper is Shravan Vasishth (Email: vasishth@uni-potsdam.de).

REFERENCES

John R. ANDERSON, Dan BOTHELL, Michael D. BYRNE, Scott DOUGLASS, Christian LEBIERE, and Yulin QIN (2004), An integrated theory of the mind, *Psychological Review*, 111(4):1036–1060.

Brian BARTEK, Richard L. LEWIS, Shravan VASISHTH, and Mason SMITH (2011), In search of on-line locality effects in sentence comprehension, *Journal of Experimental Psychology: Learning, Memory and Cognition*, 37(5):1178–1198.

Douglas BATES and Deepayan SARKAR (2007), *lme4: Linear mixed-effects models using S4 classes*, R package version 0.9975-11.

Thomas G. BEVER and Brian MCELREE (1988), Empty categories access their antecedents during comprehension, *Linguistic Inquiry*, 19(1):35–43.

Marisa F. BOSTON, John T. HALE, Shravan VASISHTH, and Reinhold KLIEGL (2011), Parallel processing and sentence comprehension difficulty, *Language and Cognitive Processes*, 26(3):301–349.

George E.P. BOX and David R. COX (1964), An analysis of transformations, *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2):211–252.

Brian DILLON, Alan MISHLER, Shayne SLOGGETT, and Colin PHILLIPS (2011), A computational cognitive model of syntactic priming, *Journal of Memory and Language*, 69(4):85–103.

Brian W. DILLON (2011), *Structured access in sentence comprehension*, Ph.D. thesis, University of Maryland.

Karen EMMOREY and Diane LILLO-MARTIN (1995), Processing spatial anaphora: Referent reactivation with overt and null pronouns in American Sign Language, *Language and Cognitive Processes*, 10(6):631–653.

Felix ENGELMANN, Shravan VASISHTH, Ralf ENGBERT, and Reinhold KLIEGL (2013), A framework for modeling the interaction of syntactic processing and eye movement control, *Topics in Cognitive Science*, 5(3):452–474.

Edward GIBSON (2000), Dependency Locality Theory: A distance-based theory of linguistic complexity, in Alec MARANTZ, Yasushi MIYASHITA, and Wayne O'NEIL, editors, *Image, Language, Brain: Papers from the First Mind Articulation Project Symposium*, MIT Press, Cambridge, MA.

Daniel GRODNER and Edward GIBSON (2005), Consequences of the serial nature of linguistic input, *Cognitive Science*, 29:261–290.

Philip HOFMEISTER (2011), Representational complexity and memory retrieval in language comprehension, *Language and Cognitive Processes*, 26(3):376–405.

Samar HUSAIN, Shravan VASISHTH, and Narayanan SRINIVASAN (2013), Locality effects depend on processing load and expectation strength, in *Proceedings of the Conference on Architectures and Mechanisms for Language Processing*, p. 96, Aix-Marseille Université, Marseille, France.

Marcel A. JUST, Patricia A. CARPENTER, and Jacqueline D. WOOLLEY (1982), Paradigms and processes in reading comprehension, *Journal of Experimental Psychology: General*, 111(2):228–238.

Reinhold KLIEGL, Michael E.J. MASSON, and Eike M. RICHTER (2010), A linear mixed model analysis of masked repetition priming, *Visual Cognition*, 18(5):655–681.

Nóra KOVÁCS and Shravan VASISHTH (2013), The processing of relative clauses in Hungarian, in Cheryl FRENCK-MESTRE, F-Xavier ALARIO, Noël NGUYEN, Philippe BLACHE, and Christine MEUNIER, editors, *Proceedings of the Conference on Architectures and Mechanisms for Language Processing*, p. 13, Aix-Marseille Université, Marseille.

Roger LEVY (2008), Expectation-based syntactic comprehension, *Cognition*, 106:1126–1177.

Roger LEVY, Evelina FEDORENKO, and Edward GIBSON (2013), The syntactic complexity of Russian relative clauses, *Journal of Memory and Language*, 69(4):461–495.

Richard L. LEWIS and Shravan VASISHTH (2005), An activation-based model of sentence processing as skilled memory retrieval, *Cognitive Science*, 29:1–45.

Richard L. LEWIS, Shravan VASISHTH, and Julie VAN DYKE (2006), Computational principles of working memory in sentence comprehension, *Trends in Cognitive Sciences*, 10(10):447–454.

Maryellen C. MACDONALD (1989), Priming effects from gaps to antecedents, *Language and Cognitive Processes*, 4(1):35–56.

Umesh PATIL, Sandra HANNE, Frank BURCHERT, Ria De BLESER, and Shravan VASISHTH (2013), Sentence comprehension in aphasia: A computational evaluation of representational and processing accounts, manuscript, accepted pending revision, Cognitive Science.

Umesh PATIL, Shravan VASISHTH, and Richard L. LEWIS (2012), Retrieval interference in syntactic processing: The case of reflexive binding in English, manuscript.

Martin PLUMMER (2010), JAGS Version 2.2.0 user manual.

R DEVELOPMENT CORE TEAM (2006), *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, `http://www.R-project.org`, ISBN 3-900051-07-0.

David REITTER, Frank KELLER, and Johanna D. MOORE (2011), A computational cognitive model of syntactic priming, *Cognitive Science*, 35(4):587–637.

Mehran TAGHVAIPOUR (2004), An HPSG analysis of Persian relative clauses, in Stefan MÜLLER, editor, *Proceedings of the HPSG-2004 Conference, Center for Computational Linguistics, Katholieke Universiteit Leuven*, pp. 274–293, CSLI Publications, Stanford.

Julie VAN DYKE and Brian MCELREE (2006), Retrieval interference in sentence comprehension, *Journal of Memory and Language*, 55:157–166.

Shravan VASISHTH, Sven BRUESSOW, Richard L. LEWIS, and Heiner DRENHAUS (2008), Processing polarity: How the ungrammatical intrudes on the grammatical, *Cognitive Science*, 32(4).

Shravan VASISHTH and Heiner DRENHAUS (2011), Locality in German, *Dialogue and Discourse*, 1:59–82, `http://elanguage.net/journals/index.php/dad/article/view/615`.

Shravan VASISHTH and Richard L. LEWIS (2006), Argument–head distance and processing complexity: Explaining both locality and antilocality effects, *Language*, 82(4):767–794.

Shravan VASISHTH, Rukshin SHAHER, and Narayanan SRINIVASAN (2012), The role of clefting, word order and given–new ordering in sentence comprehension: Evidence from Hindi, *Journal of South Asian Linguistics*, 5:35–56.

William N. VENABLES and Brian D. RIPLEY (2002), *Modern Applied Statistics with S-PLUS*, Springer, New York.

APPENDIX: A NOTE ON THE STATISTICAL METHODS USED

Here, we summarize the statistical methods used in this paper.

Experimental designs such as the present one are generally referred to as repeated measures or within-subjects designs; this refers to the fact that each participant is exposed to each level of each factor in the experiment design (in our design, we have four factors). The logic of the experiment in general is that our dependent variable or DV (this could be accuracy, measured for example as a percentage, or reading times in milliseconds; or a transformation of these values – see below) is expected to be a linear function of the predictors, which are the factors of our experiment:

$$DV \propto predictors \tag{1}$$

The central idea is that the observed data (the dependent variable, DV) is generated by an underlying statistical model with unknown parameters θ. Formally, DV is a random variable with a particular probability density/distribution function (PDF) associated with it; the PDF is specified in terms of the parameters θ. The goal of statistical analysis is to obtain estimates of these parameters and to draw inferences from these estimates (as discussed below). For example, our pronoun vs. gap manipulation tests the prediction that the presence of the pronoun will result in shorter reading time at the verb. In the linear model setting, this amounts to the claim that there exists some point value (an unknown parameter) with a particular sign (positive or negative) that represents the mean speed-up due to pronoun insertion. Generalizing this, we define a statistical model as shown below:

$$DV_i = \beta_0 + \beta_1 Pronoun_i + \beta_2 RC_i + \epsilon_i \tag{2}$$

Arranging the data in an arbitrary but fixed ordering, let DV_i represent the i-th dependent variable, for example, reading time at the verb in one of the four conditions from each of the participants. Note that each participant would have seen 24 sentences in our experiment, with six instances each of the four levels of the experiment (Pronouns present vs. absent × Relative Clause (RC) present vs. absent). This is where the term repeated measures comes from: we have multiple measurements from each group (here, subject).

According to the statistical model above, the i-th DV_i is assumed to be generated by a random component, ϵ, and a systematic component (the rest of the right hand side in the above equation). Here, ϵ_i is assumed to be generated by a normal distribution with mean zero and standard deviation σ (yet another unknown parameter which is estimated from the data). We write this compactly as $\epsilon \sim N(0, \sigma)$.[6] The variables Pronoun and RC are indicator variables; for example, when the pronoun is present, the variable Pronoun could be coded as 1, otherwise 0. Similarly, when the relative clause is present, the indicator variable RC could be coded as 1, otherwise 0. Thus, in the statistical model, β_0 is the mean reading time of the gap no-relative clause condition. Setting up indicator variables in this manner is called contrast coding, and the example of contrast coding given here is called treatment contrasts. Different contrast codings are possible; each reflects the theoretical question to be studied (in this paper, we use an anova-style contrast coding and a nested contrast coding; see main text). In the above example, we expect the parameter β_1 to be negative; this reflects our prediction that pronoun insertion will facilitate processing. We can state this as a hypothesis test; we could write that our null hypothesis H_0 is:

$$H_0 : \beta_1 = 0 \tag{3}$$

Standard statistical theory then attempts to obtain, given the data, the best estimate, for example, a maximum likelihood estimate for β_1, call it $\hat{\beta}_1$, along with an uncertainty measure for the estimate, the standard error. Note that β_1 is the true (unknown) difference between the gap and pronoun conditions; and $\hat{\beta}_1$ is the difference in the mean reading times for the gap vs. pronoun conditions. The goal of the frequentist hypothesis testing procedure is to determine, given the null hypothesis above, the probability of obtaining an absolute value of $\hat{\beta}_1$ or something more extreme. This probability, called a p-value, is the conditional probability of the data given a particular hypothesized parameter value (in the example above, $\beta_1 = 0$): we can write it as P(data | parameter).

[6] One typically defines the normal distribution in terms of variance, σ^2, but we can simply talk about standard deviation here.

A convention widely accepted in psycholinguistics is that a p-value of less than 0.05 gives grounds for rejecting the null hypothesis and accepting the alternative hypothesis. The value 0.05 is related to the fact that we conventionally fix the probability of incorrectly rejecting the null hypothesis (i.e., rejecting it when it is actually true) to 0.05. This quantity is called Type I error probability. Note here that if we fit k separate statistical models, the Type I error probability increases, and a correction is needed in order to retain an overall Type I error probability of 0.05. One popular one is the Bonferroni correction: the corrected Type I error is $0.05/k$.

The above statistical model includes an important assumption, namely that the DV_i are independent and identically distributed. The independence requirement means that each data point i is assumed to be independent from all others – this assumption is implausible when the same participant is delivering 24 data points. The other requirement, identical distribution, is that all DVs are assumed to come from a normal distribution with the same variance. Linear mixed models have been developed to address the fact that the DVs are not independent; these models estimate between-subject variance (or equivalently, within-subject covariance) in addition to the σ mentioned above, thereby taking the dependency within each subject's responses into account. The above linear model can be expanded quite easily to a linear mixed model:

$$DV_i = \beta_0 + b_i + \beta_1 Pronoun_i + \beta_2 RC_i + \epsilon_i \tag{4}$$

Here we have an extra term, b_i, called a varying intercept term, which represents each subject's adjustment to the baseline reading time β_0: subjects who are faster than average would have negative b_i, and those slower than average, positive b_i. These b_i are not estimates; they are best linear unbiased predictors (BLUPs) generated after between-subject variance has been estimated. Thus, the above model assumes that $\epsilon \sim N(0, \sigma)$ but also that $b_i \sim N(0, \sigma_{b1})$, where σ_{b1} is the between-subject variance (or, equivalently, within-subject covariance). For our purposes, subject-level variance is a nuisance variable, and we only need to take it into account in the model; our principal interest remains focused on the estimates of the coefficients β_1 and β_2, and in hypothesis tests associated with these coefficients.

Note that in linear mixed models p-values are difficult to compute for various technical reasons, but an absolute t-value greater than 2 can reasonably be assumed to be statistically significant at Type I error probability 0.05.

Two extensions of the above model are as follows. Apart from between-subject variance, we also want to take into account between-item variance. For this reason, we can introduce another additive term in the model analogous b_i, call it c_i, which comes from a distribution $N(0, \sigma_{b2})$. Thus, it is easy to add a varying intercept for items as well, and this requires the model to estimate a further variance component, that due to between-item variability. A second extension involves an adjustment for subject (and item) in the coefficients β_1 and β_2. This is called a varying slopes model. Such adjustments to the coefficients β_1 and β_2 simply take into account possible variability between subjects (and items) in their response to the pronoun and relative clause manipulation; for example, some subjects may speed up much more than others due to the pronoun vs. gap manipulation, and some items may have a greater impact on the pronoun vs. gap manipulation. Models with varying slopes take this variability into account. This is by no means the whole story regarding linear mixed models, but it does provide the reader with some guidance on how we analyzed the data.

One further twist is the issue of non-normality of residuals. Statistical inference based on the models discussed above crucially depends on normal distribution theory; the residuals ϵ are assumed to be normally distributed, and the BLUPs are too. When residuals are not normally distributed (this is usually the case in analyses of reading studies using raw reading times as dependent measures) the above models may no longer be applicable. One solution to this issue is to use generalized linear mixed models (we will not discuss this solution here); another is to find a transformation to the data such that the variance is stabilized and the normality assumptions are approximately satisfied. Box and Cox (1964) developed a procedure for stabilizing variance, which is now known as the Box-Cox procedure and is available through the `MASS` package (Venables and Ripley 2002) in R as the function `boxcox`. Briefly, the function discovers (using maximum likelihood estimation) the transformation needed in order to stabilize variance.

In our reading time data, the transform suggested by the Box-Cox procedure is the reciprocal. We changed this to a negative reciprocal in order to make it easier to interpret the sign of the estimated coefficients (the reciprocal converts the reading time to rate of processing; so a smaller value on the transformed scale is a slower rate, and a larger value corresponds to a faster rate).

Next, we briefly explain the model fitted for response accuracy, which is a proportion. To analyze these, we used the generalized linear modeling (GLM) framework (more specifically, generalized linear mixed models). The basic idea in GLMs is that we assume that responses are generated by a binomial distribution. Instead of assuming that the dependent variable is a proportion μ, we transform it to log-odds and specify a linear relationship between the predictor x and the log-odds: $\log[\frac{\mu}{1-\mu}] = \beta_0 + \beta_1 x$. The generalized linear mixed model extends this framework to deal with grouped data, as discussed above, with varying intercepts, etc. The essential point here is that again we are testing null hypotheses of the form $H_0 : \beta_1 = 0$, where the coefficient is in the log-odds scale.

This work is licensed under the Creative Commons Attribution 3.0 Unported License.
http://creativecommons.org/licenses/by/3.0/

Syntax-driven semantic frame composition in Lexicalized Tree Adjoining Grammars

Laura Kallmeyer and Rainer Osswald
Heinrich-Heine-Universität Düsseldorf, Düsseldorf, Germany

ABSTRACT

The grammar framework presented in this paper combines Lexicalized Tree Adjoining Grammar (LTAG) with a (de)compositional frame semantics. We introduce elementary constructions as pairs of elementary LTAG trees and decompositional frames. The linking between syntax and semantics can largely be captured by such constructions since in LTAG, elementary trees represent full argument projections. Substitution and adjunction in the syntax then trigger the unification of the associated semantic frames, which are formally defined as base-labelled feature structures. Moreover, the system of elementary constructions is specified in a metagrammar by means of tree and frame descriptions. This metagrammatical factorization gives rise to a fine-grained decomposition of the semantic contributions of syntactic building blocks, and it allows us to separate lexical from constructional contributions and to carve out generalizations across constructions. In the second half of the paper, we apply the framework to the analysis of directed motion expressions and of the dative alternation in English, two well known examples of the interaction between lexical and constructional meaning.

Keywords:
Lexicalized tree adjoining grammars, syntax-semantics interface, decompositional frame semantics, elementary constructions, metagrammar, feature structures, dative alternation, directed motion expressions

1 INTRODUCTION

The meaning of a verb-based construction depends not only on the lexical meaning of the verb but also on its specific syntagmatic en-

vironment. Lexical meaning interacts with constructional meaning in intricate ways and this interaction is crucial for theories of argument linking and the syntax-semantics interface. These insights have led proponents of Construction Grammar to treat every linguistic expression as a construction (Goldberg 1995). But the influence of the syntagmatic context on the constitution of verb meaning has also been taken into account by lexicalist approaches to argument realization (e.g., Van Valin and LaPolla 1997). The crucial question for any theory of the syntax-semantic interface is how the meaning components are distributed over the lexical and morphosyntactic units of a linguistic expression and how these components combine. In this paper, we describe a grammar model that is sufficiently flexible with respect to the factorization and combination of lexical and constructional units both on the syntactic and the semantic level.

The proposed grammar description framework combines *Lexicalized Tree Adjoining Grammar* (LTAG) with *decompositional frame semantics* and makes use of a constraint-based, "metagrammatical" specification of the elementary syntactic and semantic structures. The LTAG formalism has the following two key properties (Joshi and Schabes 1997):

- *Extended domain of locality*: The full argument projection of a lexical item can be represented by a single elementary tree. The domain of locality with respect to dependency is thus larger in LTAG than in grammars based on context-free rules. Elementary trees can have a complex constituent structure.
- *Factoring recursion from the domain of dependencies*: Constructions related to iteration and recursion are modelled by the operation of adjunction. Examples are attributive and adverbial modification. Through adjunction, the local dependencies encoded by elementary trees can become long-distance dependencies in the derived trees.

Bangalore and Joshi (2010) subsume these two properties under the slogan "complicate locally, simplify globally." The idea is that basically all linguistic constraints are specified over the local domains represented by elementary trees and, as a consequence, the composition of elementary trees can be expressed by the two general operations substitution and adjunction. This view of the architecture of grammar,

which underlies LTAG, has direct consequences for semantic representation and computation. Since elementary trees are the basic syntactic building blocks, it is possible to assign complex semantic representations to them without necessarily deriving these representations compositionally from smaller parts of the tree. Hence, there is no need to reproduce the internal structure of an elementary syntactic tree within its associated semantic representation (Kallmeyer and Joshi 2003). In particular, one can employ "flat" semantic representations along the lines of Copestake *et al.* (2005). This approach, which supports the underspecified representation of scope ambiguities, has been taken up in LTAG models of quantifier scope and adjunction phenomena (Kallmeyer and Joshi 2003; Gardent and Kallmeyer 2003; Kallmeyer and Romero 2008).

The fact that elementary trees can be directly combined with semantic representations allows a straightforward treatment of idiomatic expressions and other non-compositional phenomena, much in the way proposed in Construction Grammar. The downside of this "complicate locally" perspective is that it is more or less unconcerned about the nature of the linguistic constraints encoded by elementary trees and about their underlying regularities. In fact, a good part of the linguistic investigations of the syntax-semantics interface are concerned with argument realization, including argument extension and alternation phenomena (e.g. Van Valin 2005; Levin and Rappaport Hovav 2005; Müller 2006). Simply enumerating all possible realization patterns in terms of elementary trees without exploring the underlying universal and language-specific regularities would be rather unsatisfying from a linguistic point of view.

The mere enumeration of basic constructional patterns is also problematic from the practical perspective of grammar engineering (Xia *et al.* 2010): the lack of generalization gives rise to redundancy since the components shared by different elementary trees are not recognized as such. This leads to maintenance issues and increases the danger of inconsistencies. A common strategy to overcome these problems is to introduce a tree description language which allows one to specify sets of elementary trees in a systematic and non-redundant way (e.g. Candito 1999; Xia 2001). The linguistic regularities and generalizations of natural languages are then captured on the level of descriptions. Since LTAG regards elementary trees as the basic components

of grammar, the system of tree descriptions is often referred to as the *metagrammar*. Crabbé (2005) proposes a purely constraint-based approach to metagrammatical specification (see also Crabbé and Duchier 2005), which does not presume a formal distinction between canonical and derived patterns but generates elementary trees uniformly as minimal models of metagrammatical descriptions. We will adopt this approach for our framework because of its clear-cut distinction between the declarative level of grammatical specification and procedural and algorithmic aspects related to the generation of the elementary trees.

Existing metagrammatical approaches in LTAG are primarily concerned with the organization of general valency templates and with syntactic phenomena such as passivization and wh-extraction. The semantic side has not been given much attention to date. However, there are also important semantic regularities and generalizations to be captured within the domain of elementary constructions. In addition to general semantic constraints on the realization of arguments, this also includes the more specific semantic conditions and effects that go along with argument extension and modification constructions such as resultative and applicative constructions, among others. In order to capture phenomena of this type, the metagrammatical descriptions need to include semantic constraints as well. In other words, an analysis of the syntax-semantics interface given by elementary constructions that goes beyond the mere enumeration of form-meaning pairs calls for a (meta)grammatical system of constraints consisting of both syntactic and semantic components. Note that such an approach does not imply a revival of the idea of a direct correspondence between syntactic and semantic (sub)structures – an assumption which LTAG has abandoned for good reasons.

The framework proposed in this paper treats the syntactic and the semantic components of elementary constructions as structured entities, trees on the one hand and frames on the other hand, without requiring that there be any structural isomorphism between them. Frames can be understood as cognitive structures representing situations or states of affairs, and they can be formalized as typed feature structures (see Section 3). The metagrammar specifies the syntactic and semantic properties of constructional fragments and defines how they can combine to form larger constructional fragments. There is

no need for a structural isomorphism between syntax and semantics simply because the relation between the syntactic and semantic components is explicitly specified. Below we illustrate this program of decomposing syntactic trees and semantic frames in the metagrammar by a case study on directed motion expressions and on the dative alternation in English, which are well known to be sensitive to lexical and constructional meaning components. We will show how the constructional aspects that these phenomena have in common can be captured within the metagrammatical decomposition.

A long-term goal of the work described in this paper is the development of a grammar engineering framework that allows a seamless integration of lexical and constructional semantics. More specifically, the approach provides Tree Adjoining Grammars with a decompositional lexical and constructional semantics and thereby complements existing proposals which are focused on standard sentence semantics. From a wider perspective, the framework can be seen as a step towards a formal and computational account of some key ideas of Construction Grammar à la Goldberg since the elementary trees of LTAG combined with semantic frames come close to what is regarded as a construction in such approaches. Frameworks with similar goals are Embodied Construction Grammar (Bergen and Chang 2005) and Sign-Based Construction Grammar (Sag 2012).

The structure of the paper is as follows: Section 2 gives a short introduction to LTAG and the metagrammar approach; Section 3 introduces the idea of a frame-based semantics and provides the formal details of the kind of feature structures and feature logic we use for frame-semantic modelling. In Section 4, we present our model of the syntax-semantics interface, which crucially relies on elementary constructions defined as pairs of LTAG trees and semantic frames. We put the framework to work by modelling the syntax-semantics interface of directed motion and caused motion constructions (Section 5) and the dative alternation in English (Section 6). Section 7 briefly discusses the computational complexity of syntactic and semantic composition.

2 LEXICALIZED TREE ADJOINING GRAMMARS

2.1 *Brief introduction to TAG*

Tree Adjoining Grammar (*TAG*, Joshi and Schabes 1997; Abeillé and Rambow 2000) is a tree-rewriting formalism. A TAG consists of a finite set of *elementary trees*. The nodes of these trees are labelled with non-terminal and terminal symbols, with terminals restricted to leaf nodes. Starting from the elementary trees, larger trees are derived by *substitution* (replacing a leaf with a new tree) and *adjunction* (replacing an internal node with a new tree). Sample elementary trees and a derivation are shown in Figure 1. In this derivation, the elementary trees for *John* and *spaghetti* substitute into the subject and the object slots of the elementary tree for *likes*, and the *obviously* modifier tree adjoins to the VP node.

In case of an adjunction, the tree being adjoined has exactly one leaf that is marked as the *foot node* (marked by an asterisk). Such a tree is called an *auxiliary* tree. To license its adjunction to a node n, the root and foot nodes must have the same label as n. When adjoining it to n, in the resulting tree, the subtree with root n from the original tree is attached to the foot node of the auxiliary tree. Non-auxiliary elementary trees are called *initial* trees. A derivation starts with an initial tree. In a final derived tree, all leaves must have terminal labels. In a TAG, one can specify for each node whether adjunction is mandatory and which trees can be adjoined.

In order to capture syntactic generalizations in a more satisfactory way, the non-terminal node labels in TAG elementary trees are usually enriched with feature structures. The resulting TAG variant is called *Feature-structure based TAG* (FTAG, Vijay-Shanker and Joshi 1988). In an FTAG, each node has a top and a bottom feature structure (except substitution nodes that have only a top structure). Nodes

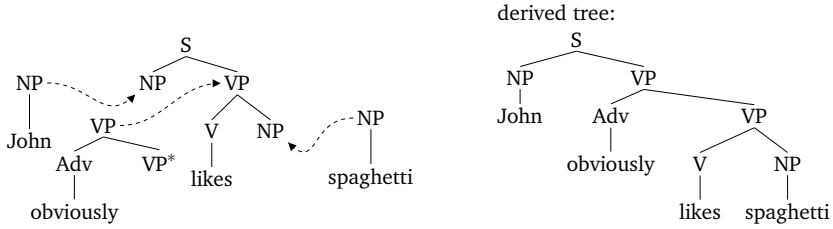


Figure 1: A sample derivation in TAG

in the same elementary tree can share features. Adjunction constraints are expressed via the feature structures. An example is given in Figure 2, where the top feature structure is notated as a superscript and the bottom feature structure as a subscript of the respective node. The features in this example are taken from the XTAG grammar (XTAG Research Group 2001). In the *singing* tree the label 1 is used to express the fact that the agreement features of the VP have to be the same as those of the subject NP. Furthermore, the different MODE values in the top and bottom feature structures of the VP node express an *obligatory* adjunction constraint. Since the values are different, the two feature structures cannot unify. Consequently, one has to adjoin a tree that separates the top structure from the bottom structure.

During substitution and adjunction, the following unifications take place. In a substitution operation, the top of the root of the new

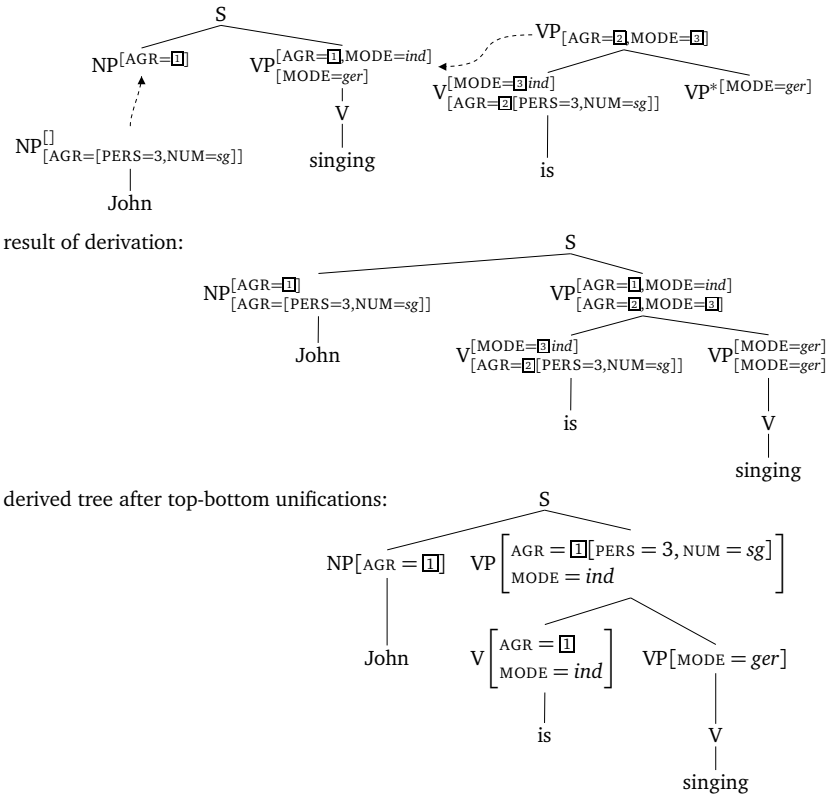


Figure 2: Feature sharing and adjunction constraints in FTAG

initial tree unifies with the top of the substitution node. In an adjunction operation, the top of the root of the new auxiliary tree unifies with the top of the adjunction site and the bottom of the foot of the new tree unifies with the bottom of the adjunction site. Furthermore, in the final derived tree, top and bottom must unify for all nodes. See again Figure 2 for an example. The middle tree shows the result of the derivation, including feature unifications arising from substitutions and adjunction. The lower tree shows the result one obtains after the final top-bottom unification. As illustrated by this example, constraints among dependent nodes can be more easily expressed in FTAG than in the original TAG formalism. Since the set of feature structures allowed in a given TAG is finite, the feature structures do not extend the generative capacity of the formalism and do not increase its parsing complexity.

2.2 *Elementary trees and tree families*

The elementary trees of a TAG for natural languages are subject to certain principles (Frank 2002; Abeillé 2002). Firstly, they are lexicalized in that each elementary tree has at least one lexical item, its *lexical anchor*. A *lexicalized* TAG (LTAG) is a TAG that satisfies this condition for every elementary tree. Secondly, each elementary tree associated with a predicate contains "slots", that is, leaves with non-terminal labels (substitution nodes or foot nodes) for all and only the arguments of the predicate (*θ-criterion for TAG*). Most argument slots are substitution nodes, in particular the nodes for nominal arguments (see the elementary tree for *likes* in Figure 1). Sentential arguments are realised by foot nodes in order to allow long-distance dependency constructions such as *Whom does Paul think that Mary likes?*. Such extractions can be obtained by adjoining the embedding clause into the sentential argument (Kroch 1989; Frank 2002).

LTAG allows for a high degree of factorization inside the lexicon, i.e., inside the set of lexicalized elementary trees. One factorization arises from separating the specification of *unanchored* elementary trees from their lexical anchors. The set of unanchored elementary trees is partitioned into *tree families* where each family represents the different realizations of a single subcategorization frame. For transitive verbs such as *hit*, *kiss*, *admire*, etc. there is a tree family (see Figure 3) containing the patterns for different realizations of the arguments (canon-

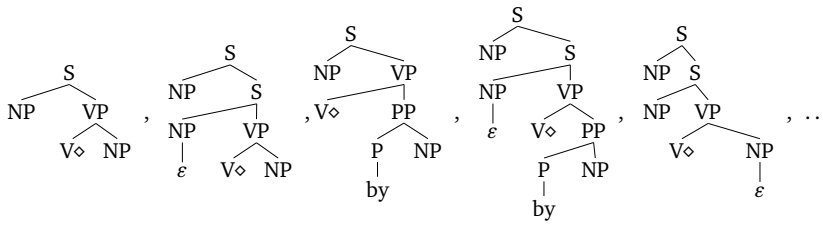


Figure 3: Unanchored tree family for transitive verbs

ical position, extraction, etc.) in combination with active and passive. The node marked with a diamond is the node that gets filled by the lexical anchor.

2.3 *Metagrammatical decomposition of elementary trees*

Unanchored elementary trees are usually specified by means of a *metagrammar* (Candito 1999; Crabbé and Duchier 2005; Crabbé *et al.* 2013) which consists of dominance and precedence constraints and category assignments. The elementary trees of the grammar are then defined as the *minimal models* of this constraint system. The metagrammar formalism allows for a compact grammar definition and for the formulation of linguistic generalizations. In particular, the metagrammatical specification of a subcategorization frame defines the set of all unanchored elementary trees that realize this frame. Moreover, the formalism allows one to define tree fragments that can be used in different elementary trees and tree families, thereby giving rise to an additional factorization and linguistic generalization. Phenomena that are shared between different tree families such as passivization or the extraction of a subject or an object are specified only once in the metagrammar and these descriptions become part of the descriptions of several tree families.

Let us illustrate this with the small metagrammar fragment given in Figure 4 that can be implemented in XMG (*eXtensible MetaGrammar*; cf. Crabbé *et al.* 2013). Each class is represented in a box with the name of the class on top. The box contains a graphical representation of the tree description specified in the class and it contains other classes used in this class. The class *Subj* does not use any other class and it contains a tree description specifying that there are four nodes labelled S, NP, VP and V with the dominance (dashed edge), immediate dominance (solid edge) and immediate linear precedence ($\prec$)

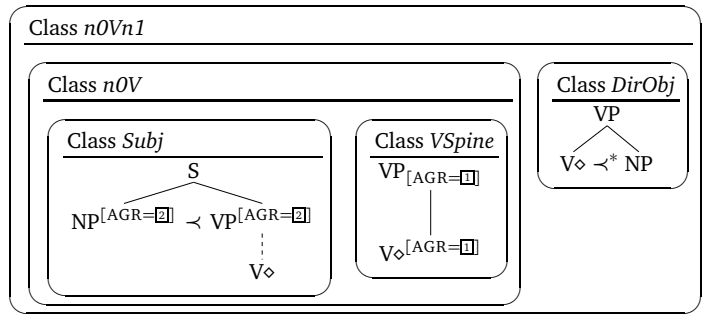


Figure 4: MG classes for transitive verbs (only canonical word order)

relations depicted in Figure 4. The diamond on the V node marks this node as the lexical anchor. Concerning features, the AGR feature values of the NP and VP nodes must be equal. The class *VSpine* specifies that there are nodes with categories VP and V with an immediate dominance such that the AGR features of the two nodes are equal. The class *n0V* uses the two classes for the subject and the verbal spine without adding any further constraints to the resulting tree description. Computing a minimal model for this class amounts to finding a tree that contains only nodes and edges described in the class. Since the lexical anchor is unique, the two anchor V nodes in *n0V* must be equal. This means that the only minimal model of *n0V* is the elementary tree for intransitive verbs with the subject in canonical position. The class *DirObj* in Figure 4 specifies that a direct object can be realized by an NP node that is immediately dominated by a VP node and that is a right sister of the V anchor node ($\prec^*$ stands for linear precedence). Combining this class with the *n0V* class yields the class *n0Vn1* for transitive verbs that leads, in this simple example, to the first tree of Figure 3 as a minimal model. The tree descriptions allow for conjunction and disjunction but not for quantification or negation. As we have seen, each class can use other classes and add new constraints on the minimal models to be computed.

3 DECOMPOSITIONAL FRAME SEMANTICS

3.1 *Frame semantics and lexical decomposition*

The program of Frame Semantics initiated by Fillmore (1982) aims at capturing the meaning of lexical items in terms of *frames*, which

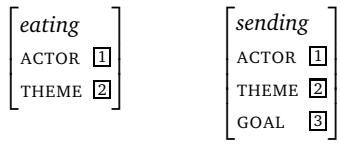


Figure 5:
Simple role frames for eating and sending events

are to be understood as cognitive structures that represent the described situations or state of affairs. In their most basic form, frames represent the type of a situation and the semantic roles of the participants; they correspond to feature structures of the kind shown in Figure 5. Frame semantics as currently implemented in the Berkeley FrameNet project (Fillmore *et al.* 2003) builds basically on role frames of this form, and it is a central goal of FrameNet to record on a broad empirical basis how the semantic roles are expressed in the morphosyntactic environment of the frame-evoking word. The ultimate goal of the FrameNet project is to devise a sufficiently rich collection of frames that allows one to describe all kinds of specific and general situation types. FrameNet frames can be related to each other in various ways. For instance, a frame can be characterized as being more specific than another frame (inheritance), or as putting a different focus on the involved participants (perspectivalization), or as being the cause or the effect component of a complex causation event. It is this interrelatedness of frames which, according to Fillmore (2007), will eventually give rise to generalizations about the morpho-syntactic realization of semantic arguments.

As shown by Osswald and Van Valin (2014), Fillmore's goal of deriving generalizations about the syntax-semantics interface would profit considerably from an analysis that takes into account the internal structure of events and state of affairs in a systemic way. This observation is in line with the fact that, in contrast to pure semantic role approaches to argument realization, many current theories of the syntax-semantics interface are based on predicate decomposition and event structure analysis (cf. Levin and Rappaport Hovav 2005, for an overview). These theories assume that the morphosyntactic realization of an argument depends crucially on the structural position of the argument within the decomposition. A simple example of such a

decomposition for the transitive verb *break* is shown in (1), using the notation of Rappaport Hovav and Levin (1998).[1]

(1) [[x ACT] CAUSE [BECOME [y *BROKEN*]]]

However, the precise status of such terms with respect to formal interpretation and inferencing is mostly neglected in the literature on argument realization.

3.2 *Semantic frames as models of meaning*

It is a central goal of the approach presented in this paper to integrate the template-based event structure decompositions with a fully formalized frame semantics. Such a decompositional semantic representation allows us to associate specific components of the semantic representation with specific syntactic fragments in the metagrammar. Crucially, we take the semantic structures associated with the syntactic structures as *genuine semantic representations*, not as some kind of yet to be interpreted logical expressions. The grammar generated by the metagrammar then consists of pairs of elementary morphosyntactic trees and elementary meaning structures. In Section 4, we will refer to such pairs as *elementary constructions*. The minimal model view is thus adopted for the semantic dimension as well: the semantic structures of elementary constructions are defined as minimal models of the constraints specified in the metagrammar – in much the same way as the syntactic structures of elementary constructions are minimal models of the specifications in the metagrammar.[2]

Let us return to the decomposition given in (1). It says, basically, that an event denoted by transitive *break* consists of an activity of someone or something x which causes a certain change of state of

[1] The idea of using representations of this kind for predicting argument realization patterns can be traced at least back to Foley and Van Valin (1984); see also Van Valin and LaPolla (1997).

[2] The use of minimal models in computational semantics has also been proposed by Blackburn and Bos (2003), among others; see also Konrad (2004). A view closely related to ours is advocated by Hamm *et al.* (2006), who propose that the logical expressions used in semantics are best considered as constraints on possible models, understood as conceptual representations. The main purpose of the logical expressions is then to characterize the minimal models, which play a crucial role in semantic processing.

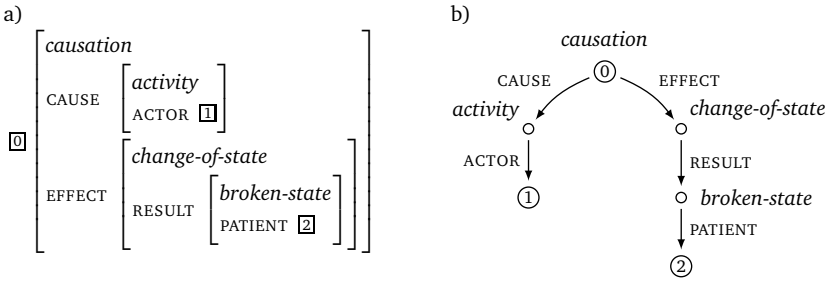


Figure 6: Possible frame representation corresponding to template (1)

something y, namely y becoming broken. Put differently, the events in question are of type *causation* and have as their CAUSE component an (unspecified) activity of x and as their EFFECT component a change of state that results into y's state of being broken. There are various ways of explicating (1) in logical terms. For instance, if we take the paraphrase just given as a blueprint, a possible logical formulation could look like (2).

(2) $\textit{causation}(e) \wedge \text{CAUSE}(e, e') \wedge \text{EFFECT}(e, e'') \wedge \textit{activity}(e') \wedge \text{ACTOR}(e', x)$
$\wedge\ \textit{change-of-state}(e'') \wedge \text{RESULT}(e'', s) \wedge \textit{broken-state}(s) \wedge \text{PATIENT}(s, y)$

Let us keep aside for the moment the question of how to treat the arguments of the predicates in (2), that is, whether to treat them as constants or as variables, and if as variables, whether and how they are bound by quantification or lambda abstraction. Representations of verb meaning like (2) are closely related to Neo-Davidsonian approaches to event semantics as proposed in Parsons (1990), among others (cf. Maienborn 2011).

It is important to notice that CAUSE is used differently in (1) and (2). In (1), CAUSE is to be interpreted as a relation between an activity and a change of state, that is, as the causation relation between events. In (2), by contrast, CAUSE denotes a functional relation that relates a causation event e to its cause component e'. In fact, it is an essential property of (2) that *all* binary relations involved are *functional*. This allows us to associate with (2) the frame shown in Figure 6, with frames understood as potentially nested typed feature structures, possibly extended with additional constraints. The graph on the right of the figure can be regarded either as an equivalent presentation of the frame, or as a minimal model of the structure on the left if the latter is seen as a frame description. Frame representations combine two central as-

pects of template-based decompositions and logical representations: like decompositional schemas, frames are concept-centered and have inherent structural properties; and like logical representations they are flexible and can be easily extended by additional subcomponents and constraints. Using frames for semantic representation is also in line with Löbner's (2014) hypothesis about the structure of representations in the human cognitive system. And, last but not least, due to their functional backbone, frames have good computational properties (see Section 3.3 and Section 7).

3.3 *Formal specification of frames*

In the following, we define frames of the type depicted in Figure 6b in a more formal way as graph-like structures, and we define the notions of subsumption and unification of frames. Moreover, we introduce a specification language for frames, which we employ for extending the description of elementary syntactic trees in the metagrammar by a frame-semantic dimension. In analogy to the syntactic dimension, semantic frames are considered as minimal models of metagrammatical specifications. A good part of the following formal framework builds on existing work on feature logics as summarized in Rounds (1997).

3.3.1 Base-labelled feature structures with types and relations

Ordinary feature structures as defined, e.g., in Carpenter (1992) come with a distinguished node, the *root*, from which each other node of the structure is reachable along the (directed) edges of the graph. In the example in Figure 6, the root node is given by the node labelled by 0. The standard unification of feature structures requires the respective roots to be identified. Semantic composition associated with TAG operations, however, typically calls for unifying a certain semantic structure with a *sub*structure of another structure; this is even the case for plain argument insertion. Moreover, in later sections we will see examples of semantic structures for which the assumption of more than one root node seems appropriate. We therefore employ typed feature structures with *multiple* base nodes.[3] Furthermore, we also allow *relations* between nodes (see Section 5.2 for an application where

[3] Our approach builds partly on Hegner (1994). The need for multi-rooted feature structures in the context of language modelling has also been noted by Sikkel (1997).

the relation in question is the mereological part-of relation between regions of space).

It is useful to first define the structures in question without explicitly mentioning the multiple base nodes involved. The following definition presumes a *signature* $\langle A, T, R\rangle$ consisting of a finite set A of *attributes*, a finite set T of *types*, and a finite set R of *relation symbols*. Each relation symbol $r \in R$ has an *arity* $\alpha(r) \in \{2, 3, 4, \dots\}$ and we write R_n for the set of n-ary relation symbols.[4] A *typed feature structure with relations* over the signature $\langle A, T, R\rangle$ is a quadruple $\langle V, \delta, \tau, \rho\rangle$ in which V is a finite set of *nodes*; δ is a partial function from $V \times A$ to V, the *node transition function*; τ is a function from V to $\wp(T)$, the *typing function*; and ρ is a function defined on $\bigcup\{V^n \mid n \in \alpha(R)\}$ which takes elements from V^n to subsets of R_n. Note that our definition comes without a type hierarchy and that the typing function τ assigns *sets* of types to each node. The reason is that we prefer to handle type hierarchies as generated by type constraints (cf. Section 3.3.4 below). If $\tau(v) = \varnothing$, this means that v has the most general type.[5]

The standard definition of *subsumption*[6] can be adapted to our framework in a straightforward way. Given two feature structures $F_1 = \langle V_1, \delta_1, \tau_1, \rho_1\rangle$ and $F_2 = \langle V_2, \delta_2, \tau_2, \rho_2\rangle$ over $\langle A, T, R\rangle$, then F_1 *subsumes* F_2, in symbols, $F_1 \sqsubseteq F_2$, if there is a function h from V_1 to V_2, called a *morphism*, which has the following properties:

- If $\delta_1(v, f)$ is defined for $v \in V$ and $f \in A$,
  then $\delta_2(h(v), f)$ is defined and $\delta_2(h(v), f) = h(\delta_1(v, f))$.
- For every $v \in V$, $\tau_1(v) \subseteq \tau_2(h(v))$.
- For every $n \in \alpha(R)$ and $v_1, \dots, v_n \in V$,
  $\rho_1(v_1, \dots, v_n) \subseteq \rho_2(h(v_1), \dots, h(v_n))$.

[4] Strictly speaking, the arity function α is part of the signature, but we keep this aside for ease of exposition.

[5] Note that types could have also been introduced as unary relation symbols; but for the task of semantic modelling it seems appropriate to concede a special status to sortal information. Conversely, we could get rid of the relations by reifying tuples of nodes by separate nodes which are related to the elements of the tuple by special "argument" attributes.

[6] Cf., e.g., Rounds (1997).

Figure 7: Non-isomorphic feature structures which subsume each other

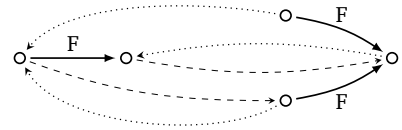
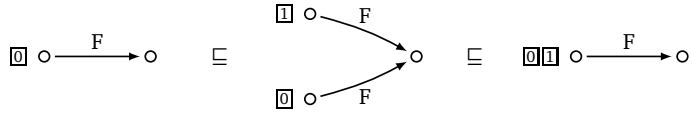


Figure 8: Strict subsumption between base-labelled feature structures



The subsumption relation defined this way is a preorder, but notice that mutual subsumption does not imply isomorphism, as illustrated by the example in Figure 7.

The above definition of feature structures does not yet capture one of the most crucial aspects of frame-based modelling, namely the property that every component is accessible via attributes from a distinguished set of base nodes. In order to formalize this requirement, let B be a countably infinite set of *base labels*. Without loss of generality, we may assume that $B = \{\boxed{0}, \boxed{1}, \boxed{2}, \dots\}$. A *base-labelled* feature structure over $\langle A, T, R, B\rangle$ is now defined as feature structure $\langle V, \delta, \tau, \rho\rangle$ over $\langle A, T, R\rangle$ together with a partial function β from B to V, the *base-labelling function*, such that every node is reachable from some *base node*, i.e., from some element of $\beta(B) \subseteq V$ via node transitions; that is, with δ extended to a partial function from $V \times A^*$ to V in the usual way, for every $v \in V$ there is a $v' \in \beta(B)$ and an *attribute path* $p \in A^*$ such that $v = \delta(v', p)$.

Morphisms between base-labelled feature structures are required to respect the base labelling. That is, a *morphism* between two base-labelled feature structures $G_1 = \langle F_1, \beta_1\rangle$ and $G_2 = \langle F_2, \beta_2\rangle$ is a morphism h from F_1 to F_2 such that if $\beta_1(l)$ is defined for $l \in B$, then $\beta_2(l) = h(\beta_1(l))$. In particular, this implies that if we add additional base labels to a given feature structure by extending the domain of the base-labelling function, we get a more specific feature structure with respect to subsumption; see Figure 8 for an example. It is not difficult to see that there exists at most one morphism between two base-labelled feature structures over a given signature; hence mutual subsumption now implies isomorphism. It follows that we can speak of "the" *least upper bound* $G_1 \sqcup G_2$ of two base-labelled feature structures G_1 and G_2 with respect to $\sqsubseteq$, which is uniquely determined up to

isomorphism, if existent at all. Hegner (1994) shows that there are efficient *unification algorithms* for computing $G_1 \sqcup G_2$ (see also Section 7).

The usual scenario for the unification of base-labelled feature structures in the applications described in the following sections presumes that they come with disjoint sets of labels. This means that feature structures need to be *relabelled*, if required. Formally, $\langle F, \beta' \rangle$ is a *relabelling* of $\langle F, \beta \rangle$ if there is a function σ on B such that $\beta'(\sigma(B)) = \beta(B)$, i.e., if the same nodes of F are base-labelled as before.

Let $G_1 = \langle F_1, \beta_1 \rangle$ and $G_2 = \langle F_2, \beta_2 \rangle$ be two base-labelled feature structures with disjoint labellings, that is, β_1 and β_2 have disjoint domains. Suppose $\boxed{0}$ is a base label of G_1 and $\boxed{1}$ is a base label of G_2. Then, when we speak of the unification of G_1 and G_2 under "identification of $\boxed{0}$ and $\boxed{1}$", we mean the unification, as defined above, of G_1' and G_2, where G_1' is the relabelling of G_1 resulting from adding the label $\boxed{1}$ to the node labelled by $\boxed{0}$. Note that we can also define G_1' without resorting to relabelling by unifying G_1 with a single-node feature structure without attributes, type, and relations, where the single node carries the labels $\boxed{0}$ and $\boxed{1}$.

3.3.2 Attribute-value descriptions

In order to specify semantic frames in the metagrammar, we need a declarative language for describing the structures introduced in the last section. The crucial point about the base labelling is that a feature structure can be characterized completely by restricting explicit reference to base-labelled nodes only. The reason is that every node of a base-labelled feature structure is accessible from one of the base nodes via successive attribute transitions.

The following language of *attribute-value descriptions* builds on the versions summarized in Rounds (1997), extended by notations taken from Hegner (1994) and Osswald (1999). First, we introduce the language of *general* attribute-value descriptions, which allows us to talk about arbitrary nodes of a feature structure. The *primitive* general attribute-value descriptions over a signature $\langle A, T, R \rangle$ are expressions of the form t, r, $p : t$, $p \doteq q$, $p \triangleq q$, $(p_1, \ldots, p_n) : r$ and $\langle p_1, \ldots, p_n \rangle : r$, with $p, p_i, q \in A^*$, $t \in T$, and $r \in R$. Let F be a feature structure $\langle V, \delta, \tau, \rho \rangle$ of signature $\langle A, T, R \rangle$ with $v, w, v_i \in V$. The *satisfaction relation* $\models$ between nodes (and node tuples) of F and attribute-value descriptions is defined as follows:

(3)

a.	$v \vDash t$	iff	$v \in \tau(t)$
b.	$\langle v_1, \ldots, v_n \rangle \vDash r$	iff	$\langle v_1, \ldots, v_n \rangle \in \rho(r)$
c.	$v \vDash p : t$	iff	$\delta(v,p) \vDash t$
d.	$v \vDash p \doteq q$	iff	$\delta(v,p) = \delta(v,q)$
e.	$\langle v, w \rangle \vDash p \triangleq q$	iff	$\delta(v,p) = \delta(w,q)$
f.	$v \vDash (p_1, \ldots, p_n) : r$	iff	$\langle \delta(v,p_1), \ldots, \delta(v,p_n) \rangle \vDash r$
g.	$\langle v_1, \ldots, v_n \rangle \vDash \langle p_1, \ldots, p_n \rangle : r$	iff	$\langle \delta(v_1,p_1), \ldots, \delta(v_n,p_n) \rangle \vDash r$

We allow general attribute-value descriptions (of the same arity) to be combined by all Boolean connectives plus $\top$ (true) and $\bot$ (false), with the usual Boolean semantics. Moreover, it is convenient to allow attribute prefixing for arbitrary (one-place) attribute-value descriptions. For instance, if ϕ is a general one-place description and $p \in A^*$, then $p : \phi$ is also a general attribute-value descriptions, with $v \vDash p : \phi$ iff $\delta(v,p) \vDash \phi$.

Now let us add base labels to the description language. *Labelled* attribute-value descriptions are of the form $l \cdot \phi$, $l \cdot p \triangleq k \cdot q$, and $\langle l_1 \cdot p_1, \ldots, l_n \cdot p_n \rangle : r$, with $k, l, l_i \in B$. In contrast to general descriptions, which are satisfied by nodes of a feature structure, labelled descriptions are satisfied by base-labelled feature structures.

(4)

a.	$\langle F, \beta \rangle \vDash l \cdot \phi$	iff	$\beta(l) \vDash \phi$
b.	$\langle F, \beta \rangle \vDash l \cdot p \triangleq k \cdot q$	iff	$\langle \beta(l), \beta(k) \rangle \vDash p \triangleq q$
c.	$\langle F, \beta \rangle \vDash \langle l_1 \cdot p_1, \ldots, l_n \cdot p_n \rangle : r$	iff	$\langle \delta(\beta(l_1), p_1) \ldots, \delta(\beta(l_n), p_n) \rangle \vDash r$

In the case of the empty attribute path ε, we write l instead of $l \cdot \varepsilon$. Again, we allow Boolean combinations of labelled descriptions.

Every base-labelled feature structure can be characterized by a finite conjunction of primitive labelled attribute-value descriptions. For instance, the frame structure of Figure 6 is specified by the following conjunction:

(5) $\boxed{0} : \textit{causation} \;\wedge\; \boxed{0} \cdot \textsc{cause} : \textit{activity} \;\wedge\; \boxed{0} \cdot \textsc{cause actor} \triangleq \boxed{1} \;\wedge$
$\boxed{0} \cdot \textsc{effect} : \textit{change-of-state} \;\wedge\; \boxed{0} \cdot \textsc{effect result} : \textit{broken-state} \;\wedge$
$\boxed{0} \cdot \textsc{effect result patient} \triangleq \boxed{2}$

Note that the attribute-value matrix shown in Figure 6a can be regarded as a normal form of the attribute-value description in (5), with conjunction symbols left implicit.

3.3.3 Reformulation in first-order predicate logic

It is not difficult to reformulate the attribute-value descriptions introduced above as expressions in first-order predicate logic, thereby regarding feature structures as standard set-theoretic models. This viewpoint is useful because predicate logic is the most conceptually basic logical language at hand and, moreover, it gives us a better connection to standard approaches in linguistic semantics, such as Neo-Davidsonian approaches (cf. Section 3.2).

First we need to become clear about what to make of the signature in the context of a first-order interpretation. For the elements of A, T, and R this is fairly obvious: attributes denote functional relations and are hence to be seen as two-place predicates; types are one-place predicates; n-ary relation symbols are n-place predicates. The treatment of the elements of B is slightly more intricate. Since base labels serve as *names*, it seems appropriate to regard them as constants. However, the standard way of interpreting constants in first-order logic requires each of them to correspond to an element of the underlying domain, which is not the case for the base labels since only some of them are used in a given structure. The solution is to treat them as one-place predicates, with the additional requirement that they are satisfied by at most one element of the domain.[7]

An *interpretation* of A, T, R, and B in the usual sense of first-order logic is a pair $\langle D, M \rangle$, consisting of a set D, the *domain* (or *universe*) and an *interpretation function* M which takes the elements of A to binary relations on D, the elements of T and B to subsets of D, and elements of R of arity n to n-ary relations on D. Since we require attribute relations to be functional and base labels to denote at most one element, we are only interested in interpretations that satisfy the following axioms for all $f \in A$, $l \in B$:

[7] Our use of base labels is similar to using *nominals* in modal logic reformulations of attribute-value logic as proposed by Blackburn (1993). While Blackburn introduces nominals to replace path-value identities, we keep the latter expressions and reserve base labels for node identification "visible from the outside". In particular, base labels matter for subsumption and unification. Another approach worth mentioning is that of Reape (1994), who introduces a polymodal language with nominals and relations.

(6) a. $\forall x \forall y \forall z (f(x,y) \wedge f(x,z) \rightarrow y = z)$
b. $\forall x \forall y (l(x) \wedge l(y) \rightarrow x = y)$

In other words, we are interested in *models* of the *theory* given by the axioms in (6).

Next, we need to rephrase the attribute-value descriptions in (3) and (4) as first-order expressions. One way to do this is to explicate the intended meaning of these descriptions in terms of predicate logic. Consider, for instance, attribute-value descriptions of the form $p : t$ (3-c). Descriptions of this sort can be phrased as 'x such that the p of x is a t'. For example, EFFECT : *change-of-state* is short for 'e such that the effect of e is a change-of-state'. Formally, this formulation can be rendered as $\lambda x(t(\iota y(p(x,y))))$. Elimination of the definite description gives $\lambda x(\exists y(p(x,y) \wedge t(y)))$ plus a uniqueness constraint that is already captured by (6-a). Hence, $p : t$ can be explicated as in (7-a), given (6-a).

(7) a. $p : t$ $\quad\quad$ $\lambda x \exists y(p(x,y) \wedge t(y))$
b. $p \doteq q$ $\quad\quad$ $\lambda x \exists y(p(x,y) \wedge q(x,y))$
c. $f\,p$ $\quad\quad$ $\lambda x \lambda z \exists y(f(x,y) \wedge p(y,z))$

By a similar argument, $p \doteq q$ can be translated as in (7-b). (7-c) simply says that attribute concatenation means relational composition. (The empty attribute path ε is interpreted by the identity relation on D.) A possible explication of (4-a) and (4-b) is shown in (8-a) and (8-b), respectively.

(8) a. $l \cdot p : \phi$ $\quad\quad$ $\exists x(l(x) \wedge \exists y(p(x,y) \wedge \phi(y)))$
b. $l \cdot p \doteq k \cdot q$ $\quad\quad$ $\exists x \exists y(l(x) \wedge k(y) \wedge \exists z(p(x,z) \wedge q(y,z)))$

The labelled description $l \cdot p : \phi$ says that the element labelled by l satisfies $p : \phi$, in symbols, $\exists y(p(\iota x(l(x)), y) \wedge \phi(y))$. Another elimination of the definite description, using (6-b), then gives rise to $\exists x(l(x) \wedge \exists y(p(x,y) \wedge t(y)))$, as desired. (8-b) can be derived in a similar vein, and the same is true of the translations of the remaining descriptions of (3) and (4).

If we apply the described reformulation technique to the description in (5), then the resulting first-order expression is equivalent, under the axioms in (6), to the following expression:

(9) $\exists e \exists e' \exists e'' \exists s \exists x \exists y \, (\boxed{0}(e) \wedge \textit{causation}(e) \wedge \textsc{cause}(e, e') \wedge \textsc{effect}(e, e'')$
$\wedge \textit{activity}(e') \wedge \textsc{actor}(e', x) \wedge \boxed{1}(x) \wedge \textit{change-of-state}(e'')$
$\wedge \textsc{result}(e'', s) \wedge \textit{broken-state}(s) \wedge \textsc{patient}(s, y) \wedge \boxed{2}(y))$

We can rewrite (9) more succinctly by the following slight abuse of notation. Since base labels are unique identifiers (if they refer at all), we can introduce them via the back door as constants by replacing $\boxed{0}$ by $\lambda x(x = \boxed{0})$, etc. Then (9) simplifies to (10).

(10) $\exists e' \exists e'' \exists s \, (\textit{causation}(\boxed{0}) \wedge \textsc{cause}(\boxed{0}, e') \wedge \textsc{effect}(\boxed{0}, e'') \wedge \textit{activity}(e')$
$\wedge \textsc{actor}(e', \boxed{1}) \wedge \textit{change-of-state}(e'') \wedge \textsc{result}(e'', s)$
$\wedge \textit{broken-state}(s) \wedge \textsc{patient}(s, \boxed{2}))$

Let us now look at the *(minimal) generic model* $\langle D, M \rangle$ of formula (9) under the "background" theory (6). "Generic" means that in this model no attribute-value description holds which is not derivable from (9) and (6) by logical inference. The domain D consists of six elements, one for each variable in (9), and the interpretation of the attributes, types, and base labels can be directly read off from (9). The resulting model is basically the structure depicted by Figure 6b, now viewed as a first-order model. This observation can be generalized as follows: There is a one-to-one correspondence between the most general base-labelled feature structures that satisfy conjunctive labelled attribute-value descriptions and the minimal generic first-order models of these descriptions rephrased as first-order expressions, presuming the axioms in (6).

3.3.4 Attribute-value constraints

We define *attribute-value constraints* as universally quantified (one-place) general attribute-value descriptions. That is, if ϕ is a one-place attribute-value description then $\forall\phi$ (in first-order notation, $\forall x(\phi(x))$) is a constraint, which is satisfied by a feature structure if and only if ϕ is satisfied by every node of the structure. We write $\phi \preceq \psi$ for $\forall(\phi \rightarrow \psi)$.

Since boolean expressions have an equivalent *conjunctive normal form*, every constraint $\forall\phi$ can be transformed into a conjunction of constraints of the form listed in (11), in which the ϕ_i's and ψ_j's are primitive attribute-value descriptions.

(11) a. $\phi_1 \wedge \ldots \wedge \phi_n \preceq \psi_1 \vee \ldots \vee \psi_m$
b. $\top \preceq \psi_1 \vee \ldots \vee \psi_m$
c. $\phi_1 \wedge \ldots \wedge \phi_n \preceq \bot$

If $m = 1$, then $\forall\phi$ is called a *Horn constraint*. In the following we are concerned with Horn constraints, if not otherwise indicated. Here are some examples of Horn constraints:

(12) a. *activity* $\preceq$ *event*
b. *causation* $\preceq \neg$*activity*
(equivalently, in normal form: *causation* $\wedge$ *activity* $\preceq \bot$)
c. AGENT : $\top \preceq$ AGENT $\doteq$ ACTOR
d. *activity* $\preceq$ ACTOR : $\top$
e. *activity* $\wedge$ *motion* $\preceq$ ACTOR $\doteq$ MOVER

Note that the first-order translation of the constraint in (12-c) gives rise to $\forall x(\exists y(\text{AGENT}(x,y)) \rightarrow \exists z(\text{AGENT}(x,z) \wedge \text{ACTOR}(x,z)))$, which is logically equivalent under (6-a) to the formula $\forall x \forall y(\text{AGENT}(x,y) \rightarrow \text{ACTOR}(x,y))$. Constraints of the form (12-c) thus express *attribute inclusions*.

In order to apply constraints to labelled descriptions for inferencing, the former need to be turned into labelled descriptions themselves. Recall that constraints hold at each node of a frame, and each node can be accessed from a base node along some attribute path. A constraint $\forall\phi$ thus gives rise to infinitely many labelled descriptions $l \cdot p : \phi$, with $l \in B$ and $p \in A^*$. In fact, $l \cdot p : \phi$ is a logical consequence of $\forall\phi$ in terms of first-order logic, under the axioms in (6). The constraint (12-a), for instance, implies the labelled description $\boxed{0} \cdot \text{CAUSE} : \textit{activity} \rightarrow \boxed{0} \cdot \text{CAUSE} : \textit{event}$, which can be applied to the description in (5) for type inference.

An important application scenario of the constraints is unification. The crucial question while unifying is: under which conditions do we need to consider only a finite number of descriptions? Feature structure unification under a finite set of labelled Horn descriptions is well-defined and computationally tractable (Hegner 1994); cf. Section 7. A simple sufficient condition is that none of the constraints enforces the introduction of additional nodes. If this is the case then the number of nodes of the unified structure is finite, and hence also

the number of relevant paths.[8] It follows that the number of labelled descriptions to be considered for inferencing is finite.

Let us have a look at the constraints in (12) from this perspective. (12-a) and (12-b) are unproblematic since they have no attribute expressions in their consequents. (12-c) also poses no problem because the constraint just adds an attribute leading to a node already given in the antecedent. (12-d) and (12-e), by contrast, both do imply the existence of additional nodes in their consequent. Note that the issue with (12-e) can be remedied by conjoining ACTOR : $\top$ (or MOVER : $\top$) to the antecedent. Note also that a constraint like (12-d) does not necessarily imply that an infinite number of base constraints has to be taken into account. In fact, in this case it wouldn't. But a more careful analysis is required in such cases in general.[9]

The constraints (12-a) and (12-b) express *type inclusion* and *exclusion*, respectively. Recall that our definition of feature structures in Section 3.3.1 does not make use of a type hierarchy. Instead, we explicitly specify the possible combinations of atomic types by type inclusion and exclusion constraints. The elements of the type hierarchy in the usual sense are then defined as the sets of atomic types that are closed and consistent with respect to type inclusion and exclusion constraints. It is well known that in the finite case, Horn constraints give rise to bounded-complete ordered sets (ordered by set inclusion) by this construction, and that every bounded-complete ordered set can be constructed this way.[10] Whether to precompile the type hierarchy or to do type inference on the fly is an issue of implementation.

Relational descriptions can also be used in constraints. For instance, the transitivity of a binary relation is expressed by the following Horn constraint:

(13) $\quad (p_1, p_2) : r \wedge (p_2, p_3) : r \rightarrow (p_1, p_3) : r$

[8] Infinite paths that might arise through cyclic structures can be avoided by limiting the maximal length of a path to the number of nodes.

[9] See also Carpenter (1992, pp. 95ff).

[10] An ordered set is *bounded-complete* if every subset that has an upper bound has a least upper bound. Note that bounded-complete ordered sets come with a least element, namely the least upper bound of the empty set.

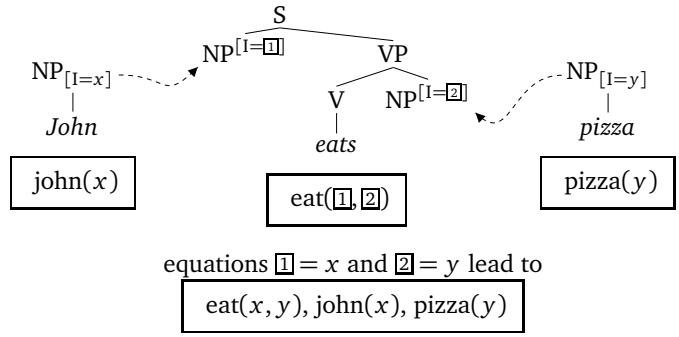


Figure 9: Syntactic and semantic composition as in Gardent and Kallmeyer (2003) and Kallmeyer and Romero (2008)

Since the consequent of (13) does not instantiate new nodes, this constraint is unproblematic when being processed during unification.[11]

4 LTAG WITH FRAME SEMANTICS

4.1 *Elementary constructions and the syntax-semantics interface*

As to the syntax-semantics interface, we basically build on approaches which link a semantic representation to an entire elementary tree and which model composition by unifications triggered by substitution and adjunction. For example, in Gardent and Kallmeyer (2003), every elementary tree is paired with a set of typed predicate logical formulas containing meta-variables linked to features in FTAG structures (see also Kallmeyer and Joshi 2003, Kallmeyer and Romero 2008). The syntactic composition then triggers unifications that lead to equations between semantic components. A (simplified) example is given in Figure 9. The feature I on the nodes is a syntax-semantics interface feature which stands for "individual". Linking, i.e., the assignment of semantic roles to syntactic arguments, is done via these interface features. In Figure 9, the syntactic unifications lead to equations $\boxed{1} = x$ and $\boxed{2} = y$. Therefore, in the semantic formulas, we have replacements of the variables $\boxed{1}$ and $\boxed{2}$ with x and y respectively. The formulas we obtain after having applied these assignments are collected in a set that is then interpreted conjunctively.

[11] Inference closure under (13) corresponds to calculating the transitive closure of the denoted relation on the given domain; its time complexity is known to be better than $\mathcal{O}(n \cdot e)$, where e is the number of pairs initially falling under r.

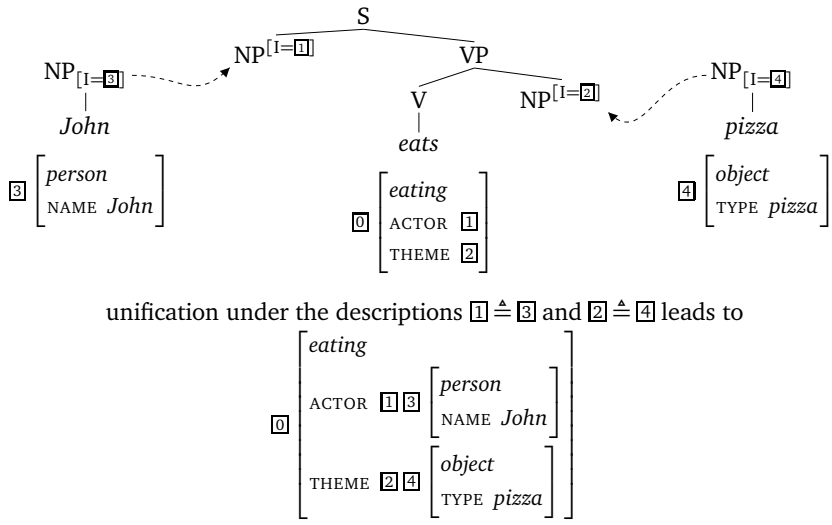


Figure 10: Syntactic and semantic composition for *John eats pizza*

The focus of this paper is on a decompositional semantics for elementary LTAG trees using frames. Figure 10 shows how the example from Figure 9 can be translated into a frame-based semantic representation in a fairly straightforward way. Each elementary tree is paired with a frame, that is, with a base-labelled feature structure as defined in Section 3.3, and base labels are used as values of interface features on the tree. Syntactic unification then triggers label equations. Here, the substitutions give rise to $\boxed{1} \triangleq \boxed{3}$ and $\boxed{2} \triangleq \boxed{4}$. Unification of the semantic frames is then performed under the additional constraints triggered by syntactic composition. This leads to an insertion of the corresponding argument frames into the frame of *eats*. Note that when using an elementary tree with its frame in a derivation, in order to avoid unintended identifications of feature structure, we always use a relabelling with fresh base labels.

A key advantage of syntax-driven approaches to semantic composition like those of Gardent and Kallmeyer (2003) and Kallmeyer and Romero (2008) is that they overcome the limitations of approaches which adhere solely to logical mechanisms such as functional application. In particular, the order of semantic argument filling is not specified by successive lambda abstraction or the like. Instead, semantic argument slots can be filled in any order (in particular, independently

of surface word order) via unifications triggered by syntactic composition.

The frame-semantic representations introduced in this paper retain this crucial property and they show a number of additional advantages. A first point is that even plain Fillmorean role frames of the kind employed in Figure 10 provide a natural way for representing semantic arguments that are not necessarily realized in the syntax (cf. Fillmore 1986). For instance, we can assume that an *eating* frame always contains a role THEME even if the theme is not overtly expressed. More importantly, using decompositional frames for semantic modelling comes with the assumption that all subcomponents of a semantic structure (i.e., participants, subevents, paths, etc.) can be accessed via *functional* relations (attributes, features) from a controlled set of base elements (cf. Section 3). As a consequence, the semantic unifications triggered by substitution and adjunction come down, to a large extent, to feature structure unifications. In particular, if a semantic structure combines values of a feature coming from different constituents then the feature values are necessarily unified as well. (We will see examples in Sections 5 and 6 below.) Moreover, feature structure unification under constraints is computationally tractable, given that certain general conditions are respected (cf. Section 3.3 and Section 7).

Notice that the use of frames does not preclude an approach as in Gardent and Kallmeyer (2003) and Kallmeyer and Romero (2008) for modelling semantic composition beyond the level of elementary trees, including the effect of logical operators such as quantifiers and other scope taking elements. But the technical details of the integration of quantifiers into frames remain to be worked out and are beyond the scope of this article.

Another approach to the syntax-semantics interface worth mentioning in this context is the synchronous TAG model of Nesson and Shieber (2006). That model employs the TAG formalism not only for the syntax of the object language but also for representing the structure of type-logical formulas on the semantic side. In our approach, by comparison, the semantic structures associated with the syntactic trees are not regarded as expressions of a formal language but as semantic models. Since feature structures are not necessarily trees, the use of synchronous TAG is not an option in our case.

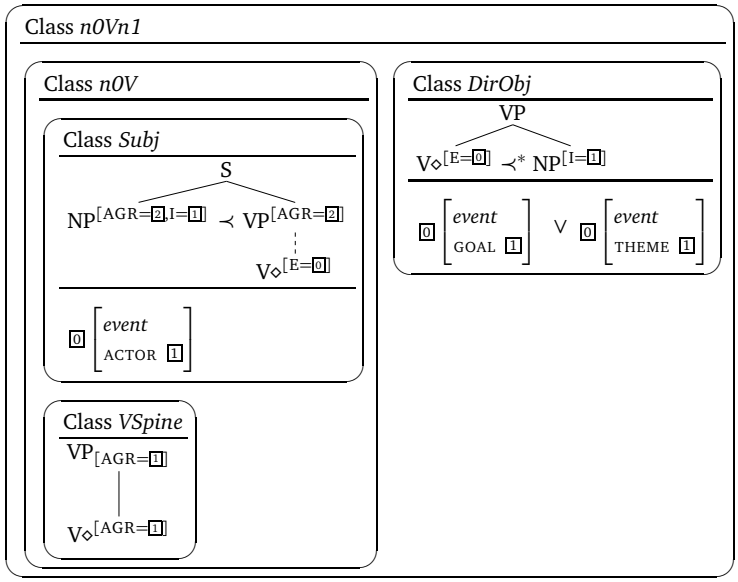


Figure 11:
MG classes with
semantic frames

4.2 *Metagrammatical decomposition of elementary constructions*

Similar to the metagrammar factorization in the syntax, a decomposition of the semantic frames paired with unanchored elementary trees is possible as well. Firstly, the semantic contribution of unanchored elementary trees, i.e., constructions, can be separated from their lexicalization, and, secondly, the meaning of a construction can be decomposed further into the meaning of fragments of the construction. Due to this factorization, relations between the different parts of a syntactic construction and the components of a semantic representation can be expressed.

As an example consider Figure 11 that repeats the MG classes from Figure 4, equipped with frame-semantic descriptions. The *Subj* class now tells us that the subject can contribute the actor of an event. (This is of course not the only possible contribution of a subject; the example is highly incomplete.) According to the *DirObj* class, the object NP can contribute the goal or the theme of an event. When compiling the description of *n0Vn1*, i.e., when computing its minimal models, both the actor-theme and the actor-goal combination are generated.

Several remarks are in order concerning this example. The interface feature E ("event") is the label of the event frame of the verb. By

Figure 12:
Tree and frame descriptions in *DirObj*

Class *DirObj*
$n_1[cat : \mathrm{VP}] \wedge$ $n_2(mark : anchor)[cat : \mathrm{V}[top : [e : \boxed{0}]]] \wedge$ $n_3[cat : \mathrm{NP}[top : [i : \boxed{1}]]] \wedge$ $n_1 \rightarrow n_2 \wedge n_1 \rightarrow n_3 \wedge n_2 \prec^* n_3$
$\boxed{0} : event \wedge (\boxed{0} \cdot \textsc{goal} \triangleq \boxed{1} \vee \boxed{0} \cdot \textsc{theme} \triangleq \boxed{1})$

equating different E values on the V nodes, the corresponding event frames are unified. Concerning the status of the semantic elements in the metagrammar classes, we take them to be feature structure descriptions. This is in parallel to the syntactic parts that are tree descriptions. Incorporating a class C into a higher class C' (e.g., *Subj* into *n0V*) amounts to adding the descriptions of C as conjuncts to the syntactic and semantic descriptions of C', using fresh base labels in the descriptions if necessary.

The syntax of the tree descriptions is the one from XMG (Crabbé *et al.* 2013), a quantifier- and negation-free first order logic while the syntax of the feature structure descriptions leans on the attribute-value language introduced in Section 3.[12] To see how the tree descriptions and feature structure descriptions in the metagrammar could look, consider Figure 12 that gives the tree and frame descriptions for the class *DirObj*. In the syntax, we have free variables $n_1, n_2, \ldots$ for nodes. The conjuncts can describe the categories of nodes, special markings (for instance, anchor or foot node), and the feature values defined in the top and bottom feature structures of the nodes. The binary relations on nodes can specify dominance ($\rightarrow^*$), immediate dominance ($\rightarrow$), linear precedence ($\prec^*$) and immediate linear precedence ($\prec$). The node variables are taken to be existentially bound. In the semantics, we have base labels $\boxed{0}, \boxed{1}, \ldots$ and we allow for conjunctions and disjunctions of labelled attribute-value descriptions (cf. Section 3.3.2). In addition there are constraints on frames (cf. Section 3.3.4) that are relevant both for metagrammar compilation and for frame unification during parsing. In any minimal model computed for a metagrammar class, the frame has to satisfy these constraints.

The pairs of elementary trees and frames resulting from the compilation of the grammar are called unanchored *elementary construc-*

[12] Concerning the integration of frame descriptions into XMG, first proposals can be found in Lichte *et al.* (2013).

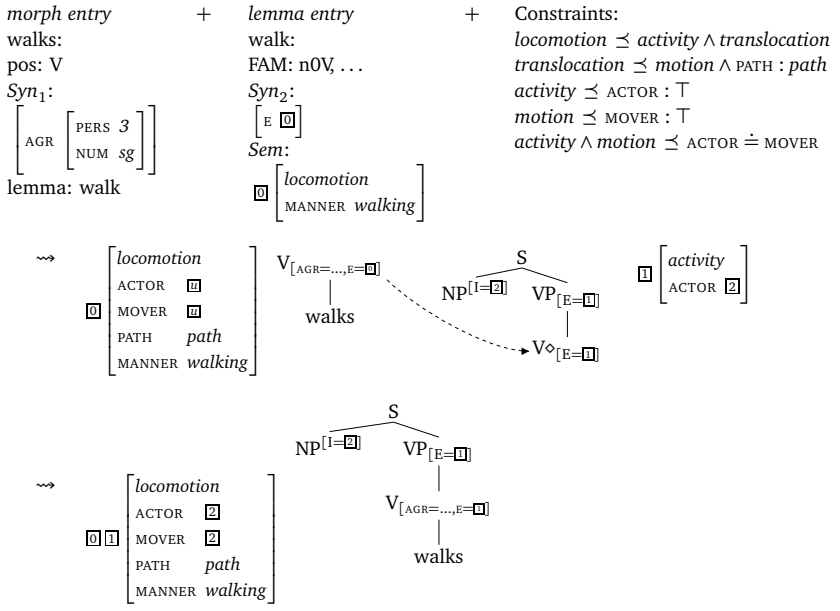


Figure 13: *Morph* and *Lemma* lexicons and lexical anchoring

tions. The step from the description in the metagrammar to the object in the grammar amounts to computing a minimal model. In the syntax, this model is such that all of its nodes and edges have to be present in the description. In the semantics, this minimal model is the smallest feature structure (with respect to subsumption) that satisfies the descriptions given by the metagrammar class and the constraints. Only those metagrammar classes that are marked as characterizing a tree family are compiled in this way (in our example only *n0V* and *n0Vn1*). Each of these classes then yields a set of unanchored elementary constructions which is an unanchored construction family of the LTAG in question.[13]

4.3 *Lexical anchoring*

In order to obtain lexicalized elementary trees, we have to fill the anchor nodes with lexical items. An example is shown in Figure 13. Lexical information is stored in a lemma lexicon and a morphological lexicon containing inflected forms. The latter gives for each form

[13] Note that in the simplified examples in this section, the class *n0V* yields only a single minimal model while *n0Vn1* leads to two minimal models since only the canonical realizations of arguments are taken into account.

the category (part-of-speech), a syntactic feature structure Syn_1 contributed by this form, and its lemma. The lemma lexicon specifies for each lemma the selected tree family (or families), again a syntactic feature structure Syn_2 and a semantic frame description *Sem*. This frame description is combined with the general constraints on frames and a minimal model is computed that is the semantics of the lexical element. In our example, the minimal model has one base label ($\boxed{0}$) and it also satisfies a path identity ($\boxed{0} \cdot \text{ACTOR} \triangleq \boxed{0} \cdot \text{MOVER}$).[14] In the corresponding attribute-value matrix, we express the latter using boxed letters, here $\boxed{u}$, instead of numbers, in order to make clear that this is not a base label but just the common shorthand notation for path identity in the matrix notation of attribute-value descriptions. In parallel, the syntactic feature structures are unified and a node of the category specified in the morphological lexicon is created, decorated with the resulting syntactic feature structure. The lexical item is a daughter of this node. Lexical anchoring can then be considered as a substitution step.

Note that according to the distribution of semantic and syntactic information among the different components in Figure 13, valency information is provided by the unanchored tree (e.g., the information that the actor is contributed by the subject NP). The lexical anchor specifies its semantics, in particular its semantic arguments, but does not determine the syntactic realizations of these arguments. In other words, linking is specified in the metagrammar.

In the following sections, we apply our syntax-semantics architecture to directed motion expressions and to the dative alternation. In the metagrammar decomposition, we will be able to share several metagrammar classes in the specifications of the elementary constructions of the two phenomena.

[14] Note that the description $\boxed{0} \cdot \text{ACTOR} \triangleq \boxed{0} \cdot \text{MOVER}$ is equivalent to $\boxed{0} \cdot (\text{ACTOR} \doteq \text{MOVER})$.

5 APPLICATION I: DIRECTED MOTION EXPRESSIONS

5.1 *The expression of directed motion in English*

Modelling the syntax-semantics interface of directed motion expressions requires us to be explicit about a number of issues concerning the syntactic and semantic structure of such expressions, many of which have been discussed extensively in the literature. In the following, we are concerned with directional expressions in English that are constructed from verbs of motion and directional prepositional phrases (PPs). The relevant constructions include intransitive verbs of motion (14) as well as transitive verbs of caused motion and transport (15).

(14) a. Mary walked to the house.
b. The ball rolled into the goal.

(15) a. John threw/kicked the ball into the goal.
b. John pushed/pulled the cart to the station.
c. John rolled the ball into the hole.

Directional specifications are not restricted to goal expressions as in (14) and (15) but can also describe the source or the course of the path in more detail. Moreover, path descriptions can be iterated to some extent (16).

(16) a. John walked through the gate along the fence to the house.
b. John threw the ball over the fence into the yard.

Below we will use this property as an indicator for distinguishing between arguments and adjuncts.

5.1.1 Verbs of motion

It is common to distinguish between manner-encoding and path-encoding verbs of motion. The first kind of verbs (*run*, *roll*) lexically encode the manner of the motion but no path-related information, while the second kind of verbs (*enter*, *leave*) do not encode the manner but specify the direction of motion. Manner-encoding motion verbs lexically characterize activities or processes. Directional information about the goal or path can be added by appropriate adverbials, i.e.,

by "satellite framing" constructions in the sense of Talmy (2000b). In the following, we focus on manner-encoding verbs since our goal is to model the syntactic and semantic processes of combining directional specifications with motion expressions.

There are also motion verbs for which the actor differs from the entity that undergoes the motion. This class includes verbs of transport and caused motion (*carry*, *drag*, *push*, *throw*). As with manner-of-motion verbs, transport and caused motion verbs do not lexically specify a direction or goal. Again, directional information can be added by adverbials. The verbs of transport and caused motion are basically transitive verbs whose direct object refers to the moving entity. They can be sub-divided into different classes depending on (i) how the motion of the object is enforced by the actor and (ii) the extent to which the activity of the actor and the manner of motion are lexically specified (cf. Ehrich 1996). Concerning (i), we can distinguish between *onset causation* (*throw*, *kick*) and *extended causation* (*pull*, *drag*), following the terminology of Talmy (2000a). Verbs of the first type describe the punctual initiation of a motion event; verbs of the second type describe the continuous enforcement of the motion. As to (ii), some of the verbs in question specify the manner of motion of the moved object but say nothing about the activity of the actor (*roll*, *slide*), while for other verbs the converse is true (*pull*, *drag*).[15]

5.1.2 Syntactic issues

In the context of the LTAG analysis presented in the following sections, a crucial issue is whether to treat directional expressions such as those in (14)–(16) as complements or as adjuncts. Moreover, an argument can be determined by the base lexeme or it can be introduced by a construction or a lexical rule. For instance, sentences of type (15-c) are often characterized as *caused motion constructions* or *causative path resultatives* (Goldberg and Jackendoff 2004). That is, the directional argument is constructionally introduced. Within the LTAG approach both the basic argument structure construction and the extended construction are represented by elementary trees. The relation between these trees, and the fact that one of them builds on the other, is captured in the class structure of the metagrammar (cf. Section 5.3).

[15] Cf. Ehrich (1996) for further distinctions.

Dowty (2003) counts directional PPs as adjuncts of motion verbs since their presence is not obligatory and they do not "complete" but "modify" the meaning of the head verb. Dowty distinguishes adjuncts from elliptical complements by characterizing the latter as cases where a semantically required element must be inferred contextually. Van Valin and LaPolla (1997) classify directional PPs as "argument-adjuncts". Like adjuncts, argument-adjuncts are predicative, but they introduce an argument into the syntactic core of the head verb and they typically share an argument with the predicate encoded by the verb. A well known distinction observed by Jackendoff, Verkuyl and Zwarts, among others, is the distinction between *bounded* and *unbounded* directional PPs, which give rise respectively to telic (17-a) and atelic (17-b) event descriptions (Jackendoff 1991; Verkuyl and Zwarts 1992; Zwarts 2005).

(17) a. She walked to the brook (in half an hour/*for half an hour).
b. She walked along the brook (*in half an hour/for half an hour).

With reference to this distinction, and based on data from Dutch and other languages, Gehrke (2008) argues that bounded directional PPs are complements of the verb while unbounded PPs are adjuncts. For verbs of motion and transport, which are lexically atelic, this means that a directional expression is regarded as a complement just in case it changes the aspectual type of the expression. This assumption is compatible with the formal criterion that expressions that can be added iteratively (as, e.g., prenominal adjectives) need to be analyzed as adjuncts. In the following, we take this criterion as a preliminary working definition of adjuncthood.

5.1.3 Translocation, paths, and directions

Since the general notion of motion covers also motion in place (e.g., shaking), we use the more technical term *translocation* when we refer to the continuous change of an object's position in space (cf. Zlatev *et al.* 2010). A translocation event is by definition associated with some trajectory, trace or path of the moving entity. The approaches found in the literature differ with respect to the explicit representation of the

path in the lexical semantics of the respective verbs. While in Dowty (1979) and Kaufmann (1995), paths are not part of the semantic representations of translocation verbs, Zwarts (2005) proposes a thematic function TRACE that maps motion events to the path traversed by the moving entity and in Mani and Pustejovsky (2012), it is assumed that "manner-of-motion predicates leave a trail of the motion along an implicit path, as measured over time." Similarly, Eschenbach *et al.* (2000) take paths as part of the semantics of verbs of motion. The paths referenced by verbs are here again understood as trajectories, that is, as the collection of "all points the object occupies during its course."

Paths, traces or trajectories provide a straightforward semantic link between motion verbs and directional specifications. Directionals (in English) often occur morphologically combined with locatives. For example, the directional preposition *into* specifies a path whose end point is in the interior of the goal expressed by the nominal complement of the preposition. The interior region associated with an object, as well as other regions specified by locatives, can be regarded as functional attributes of that object. We will employ this view below for the frame representations of directional prepositions.

5.2 *Analysis of directional expressions*

5.2.1 Frame representation of directed motion

The semantics of directional expressions has often been analyzed in terms of logical expressions of one kind or another (cf. Eschenbach *et al.* 2000). Our approach employs frames for semantic representation, with frames understood to be typed feature structures with relations. As explained in Section 3, this does not preclude a logical perspective but puts emphasis on the role of minimal models for semantic representation. Moreover, our frame-semantic approach takes into account semantic composition and thus goes beyond flat role frame approaches à la FrameNet. For example, the verb *throw* expresses a caused motion, that is, the described event can be analyzed as a complex causation event whose cause component consists of the activity of the thrower and whose effect is the ballistic motion of the thrown object. A possible frame-semantic representation of this decompositional analysis is shown on the right side of Figure 14, which also shows the frame for *walk*.

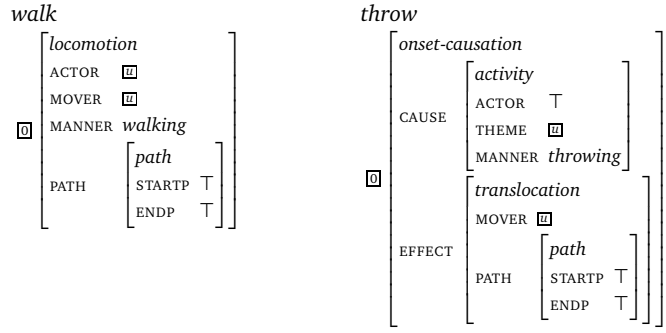
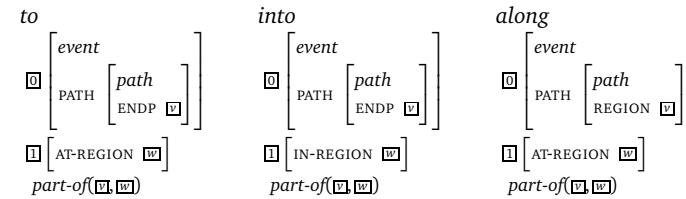


Figure 14:
Possible frame-semantic representations of some verbs of (caused) motion

In the given representations, a good part of the lexical meaning is condensed in the types or left implicit. For instance, the precise way of how the actor induces the (ballistic) motion of the object in throwing events is simply encoded by an atomic value of the attribute MANNER. Similarly, the causation type of throwing events is encoded by the type *onset-causation* of the main event. A more explicit representation would include the temporal characteristics of an onset causation, i.e., punctuality and temporal precedence of the causing event. Notice that the path or trace of the moving entity is made explicit by the frames in Figure 14. As argued above, the path of the moving object is an inherent semantic component of translocation events; the path provides the anchor for directional specifications. It is important to keep in mind that the presence of the PATH attribute in the frame representation of, say, *walk* does not imply by any means that *walk* lexically encodes information about the path of the movement.[16] Concerning semantic roles, we allow multiple descriptions of an event participant in a single frame. Each *motion* event has a participant that moves, the

[16] It is instructive to compare our use of the attribute PATH with the corresponding semantic role (frame element) of the frames 'Motion', 'Motion_directional', and 'Self_motion' in the Berkeley FrameNet database. In our decompositional approach, the path (or trace, or trajectory) is an inherent component of translocation events. In FrameNet, the 'Path' element is directly related to path descriptions such as *down the stairs*, *along the brook*, etc.; see Section 5.2.3 for our analysis of such expressions. Moreover, the relevant FrameNet frames come with core elements 'Goal', 'Direction', 'Source', etc., which is not the case for the representations shown in Figure 14. The underlying intuition is that while the path of a translocation has an end point, it is not part of the concept *per se* to have a goal.

Figure 15: Frame examples for directional prepositions



MOVER. If the event is an activity, this participant becomes at the same time the ACTOR.

The frame representation of directional prepositions follows the outline described in the previous section. The basic idea is that frames associated with directional prepositions can unify with frames of translocation, which gives rise to frames that express directed motion. For example, the frame for the preposition *into* shown in the middle of Figure 15 represents (directed) motion to the interior region of an object $\boxed{1}$ which is denoted by the nominal complement of the preposition. The frame description in the last line of the figure encodes the condition that the end point of the path or trajectory generated by the motion is in fact a mereological part of the region in question. In the matrix notation, now extended by relational descriptions, boxed letters serve again as a shorthand notation for (labelled) attribute paths. That is, *part-of*($\boxed{v}$, $\boxed{w}$) is short for the labelled description in (18).

(18) $\langle \boxed{0} \cdot \text{PATH REGION}, \boxed{1} \cdot \text{AT-REGION} \rangle : \textit{part-of}$

Note that the intended meaning of *part-of* has to be spelled out by appropriate constraints. In the case at hand, this includes transitivity (cf. (13)), reflexivity, and anti-symmetry, as well as type constraints on the domain and range of the relation. So far, our impression is that all (non-functional) relations we need are of this sort. In Section 7, we will say a few words about applying such constraints during unification.

The semantic representations described so far allow us to introduce the basic ideas of a syntax-driven semantic frame composition in the following sections. In a fully developed theory of frame-semantic representations for events, the types and features used in the frames are systematically related to each other by constraints. For instance, the inheritance hierarchy of the event types introduced so far would look like the one depicted in Figure 16. Each type comes with feature declarations that formulate constraints on the frames of this type.

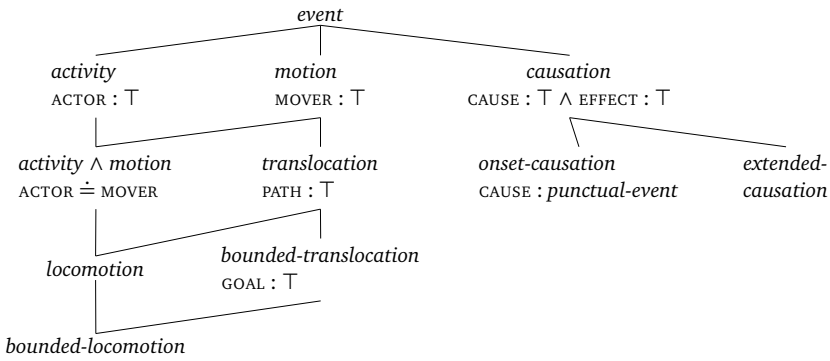


Figure 16: Partial sketch of constraints on event types and attributes

The constraints in Figure 16 specify for instance that frames of type *causation* have a CAUSE and an EFFECT attribute, and that the value of the CAUSE attribute of *onset-causation* events is of type *punctual-event*.

5.2.2 Intransitive directed motion constructions

This section deals with the combination of motion verbs and directional PPs as shown in (19).

(19) a. Mary walked/ran to/into the house.
b. Mary walked/ran along the river.
c. Mary walked/ran over the bridge along the fence through the meadows.

Recall that our criterion for treating a constituent as an argument or an adjunct is iterability. Constituents that cannot be iterated and that add a semantic role (no matter whether the role is already present in the frame contributed by the verb) are taken to be complements in the sense that their integration into the unanchored tree for the verb is part of the metagrammatical specification of elementary trees. For this reason, the examples in (19-a) are treated as PP complements while the PP in (19-b) is considered an adjunct. PPs of the type in (19-b) can be iterated as can be seen in (19-c).

In the PP complement case, the preposition is not part of the elementary tree of the verb since it is not determined by the verb. This is in contrast to constructions where a specific preposition is treated as a coanchor of the elementary tree. An example is the elementary tree for phrasal verbs such as *subscribe to*, as in *Mary subscribes to a*

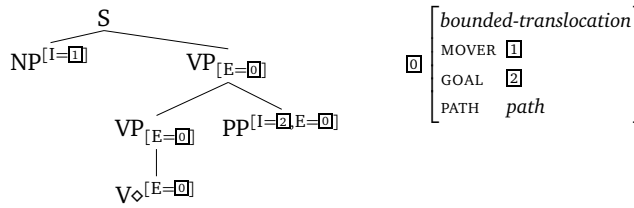


Figure 17: Unanchored tree and semantics of *n0Vpp(dir)* construction

linguistics journal, where the preposition *to* is taken to be a coanchor of the elementary tree.

As explained in Section 5.2.1, we assume that the motion verbs in (19) define a *locomotion* that has a certain path (trace, trajectory) associated with it. This path has a start and an end point. In the directed motion construction, the additional PP adds a further argument with the semantic role GOAL. The way this goal combines with the path, i.e., whether it is its end point, whether it adds a direction to the path, etc., depends on the preposition.

The unanchored elementary tree for an intransitive verb with an additional directional PP is given in Figure 17. Note that we assume a binary left-branching structure for the VP, i.e., every argument inside the VP is the right sister of a VP node and the lowest VP node immediately dominates the verbal anchor. This allows for the adjunction of modifiers between the verb and the directional PP as in (20).

(20) He ran quickly to the river.

The decoration of the elementary tree with the interface features I and E ensures that the substitutions of the subject NP and the object PP fill the corresponding argument roles and, furthermore, that adjunctions of modifiers to the VP node extend the event frame $\boxed{0}$.

The preposition determines the relation between the path of the motion and the goal. Figure 18 shows the elementary trees of different directional prepositions (cf. Figure 15). We assume that objects such as *the house* have a certain topological structure. They come with different types of regions, an at-region that contains all points that can be said to be *at* the object, an in-region that determines the space that constitutes the inner part of the object, etc. The preposition *to* makes reference to the at-region of an object; it expresses that the endpoint of the path must be contained in the at-region of the entity denoted by the NP complement of the preposition. Similarly, *into* expresses that

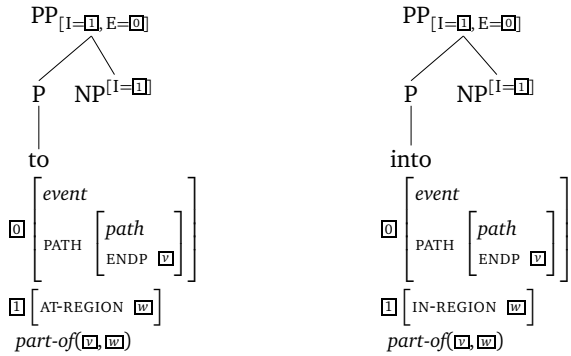


Figure 18: Elementary trees for prepositions

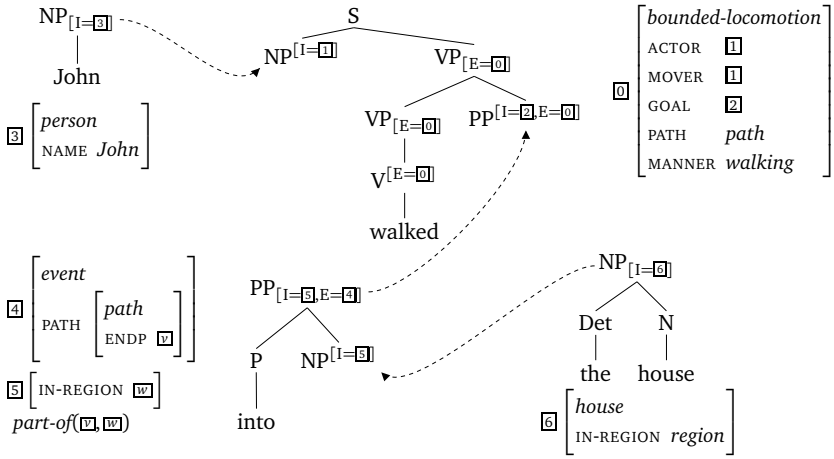


Figure 19: Derivation of (21)

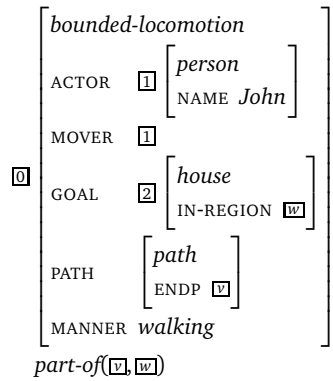
the endpoint must be contained in the in-region of the entity.

As an example, let us consider the derivation of (21). Figure 19 shows the elementary constructions involved and how they are combined.

(21) John walked into the house.

The representation for *the house* comes with an in-region (among others). (The composition of the determiner and the noun into the NP *the house* is left aside in this example.) The preposition *into* links the in-region to the end point of the path traversed throughout the walking activity. The various substitutions give rise to the following identities: $1 \triangleq 3, 2 \triangleq 5 \triangleq 6$ and $0 \triangleq 4$. With the corresponding unifications, the resulting frame is the one given in Figure 20.

Figure 20:
Resulting frame for (21)

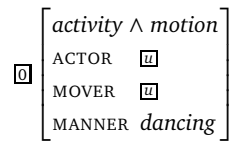


Motion verbs that are turned into a directed motion by adding a goal and a path (such as in (22)) differ from verbs of locomotion (as in (21)) with respect to their lexical semantics.

(22) Mary danced into the room.

Walk comes with a path while *dance* does not. The lexical frame for *dance* is shown in Figure 21. When combining it with the unanchored construction tree, the path attribute is added and the goal argument is linked to the PP.

Figure 21:
Frame for *dance*



5.2.3 Path modification

Now let us turn to the case where the directional PP is an adjunct that gives an additional specification of the path of the event as in (19-b). In these cases, the verb of locomotion anchors an intransitive activity tree. As an example, consider the derivation of (23). Figure 22 shows the adjunction of the *along* elementary tree into the elementary intransitive construction of *walked* (see Figure 13 for the anchoring step for this construction). The frame linked to *along* expresses that the entity denoted by the NP within the PP has an at-region that must contain the entire region of the path. Note that the frame contributed by the preposition does not have a unique root. The reason

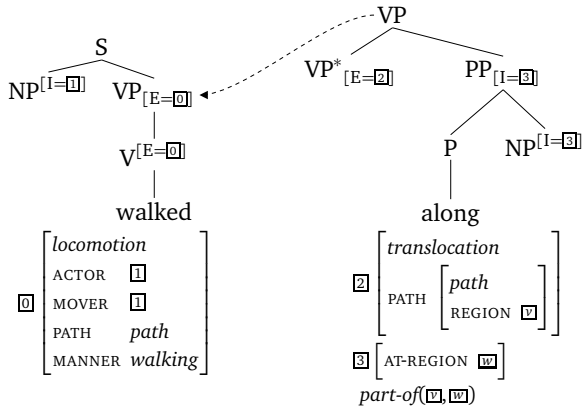


Figure 22: Derivation of (23)

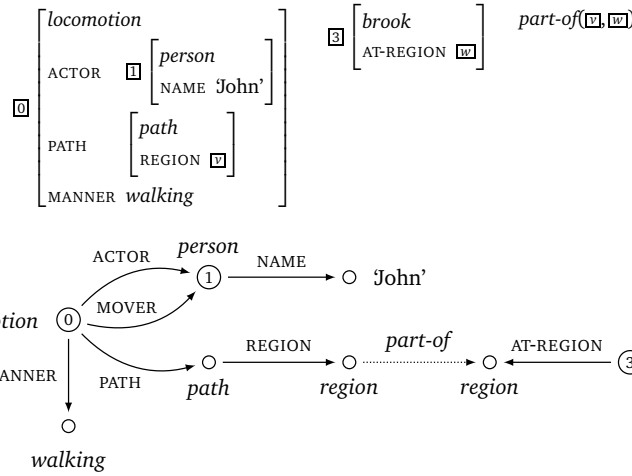


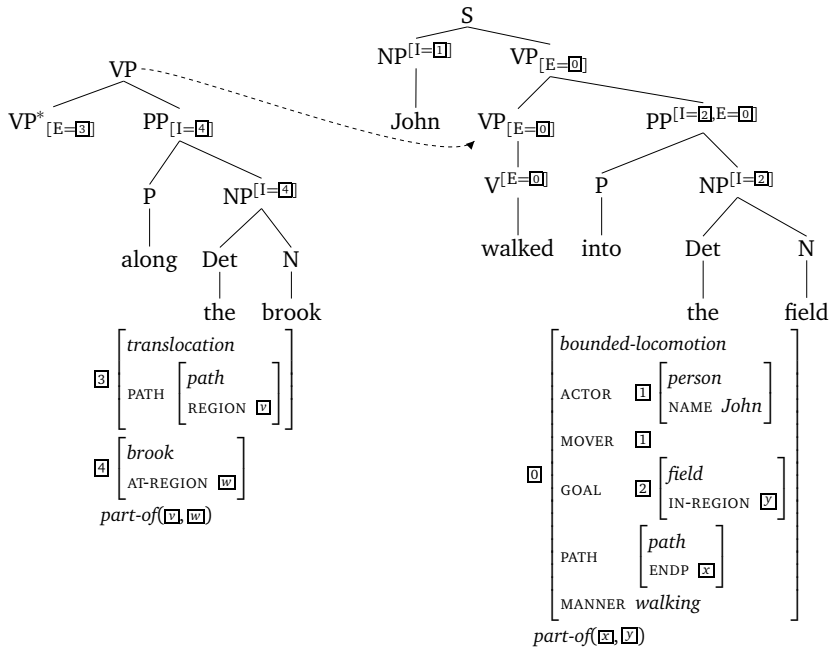
Figure 23: Resulting frame for (23)

is that the NP does not contribute an argument and therefore it does not fill a semantic role slot. The link between the object denoted by the NP and the walking activity concerns only the at-region of the former.

(23) John walked along the brook.

As a result, when combining further with the elementary trees for *John* and *the brook*, we obtain the frame in Figure 23. In addition to the attribute-value matrix, the figure also shows the corresponding feature structure depicted as a graph. The graph shows more clearly that we have more than one root node in this frame and, furthermore,

Figure 24: Derivation of (25)



if we disregard the non-functional relation *part-of*, the frame is not even a connected graph.

Obviously, examples with motion verbs such as *dance* which do not lexically specify translocation work as well, cf. (24). In these cases, the preposition introduces the path.

(24) Mary danced along the fence.

As a last example, let us consider a combination of argument directional PPs and adjoining directional PPs.

(25) John walked along the brook into the field.

Figure 24 illustrates the derivation step that combines the PP *along the brook* with the rest of the sentence. The unification of 3 and 0 triggered by the adjunction gives rise to the frame shown in Figure 25, which combines the two constraints on the path contributed by the two PPs: The entire path (i.e., its REGION) must be contained in the AT-REGION of the brook and the ENDP (endpoint) of the path must be contained in the IN-REGION of the field.

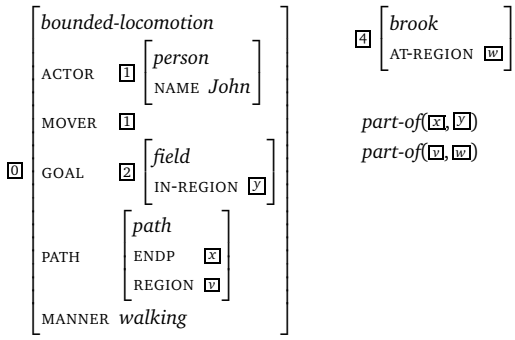


Figure 25:
Resulting frame for (25)

5.2.4 Caused motion constructions

We now turn to verbs of transport and caused motion as exemplified in (26).

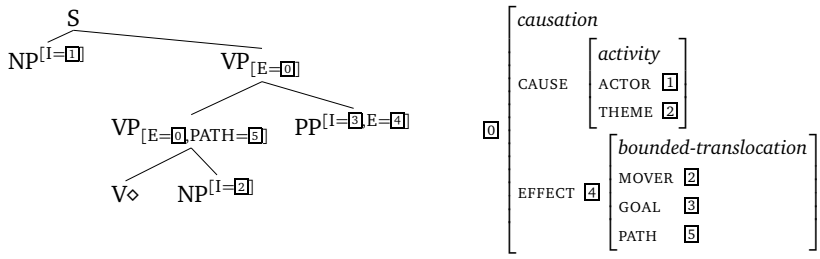
(26) a. Mary threw the ball into the hole.
b. Mary pulled the cart along the river.
c. Mary kicked the ball along the line into the goal.

Our proposal for the unanchored construction and its semantics is shown in Figure 26. The difference relative to the intransitive directed motion construction *n0Vpp(dir)* discussed above is that now the theme, i.e., the entity denoted by the direct object is moving. This movement is the effect of an action performed by the actor that affects the theme. Therefore the directed motion of the moving entity is represented as the effect of a causation whose cause is an action performed by the subject.

A difficulty with this construction is that the PP argument and directional PP modifiers need to access the embedded translocation event while other modifiers might want to access the main event. As a solution that makes both accessible and that distinguishes them, we propose to use the feature E on the PP argument slot for the embedded translocation event (here 4) and the E feature on the VP node for the highest event (here 0). This allows for the insertion of modifiers between the verb and the PP that modify the higher event, as in (27). Such modifiers adjoin to the lower VP node.

(27) Paul threw the ball immediately into his opponent's goal.

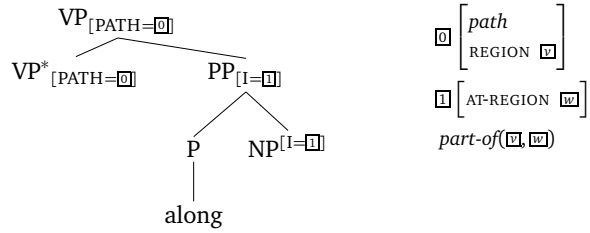
Figure 26: Unanchored tree and semantics of *n0Vn1pp(dir)* construction



However, we also want to allow path modifiers between the verb and the PP that modify the embedded event as in (26-c). A modifier such as *along the line* does not contribute a participant to the motion event, in contrast to the case of the directional argument PP. It only adds some further specification about the path of this event. Therefore, it is actually enough for it to have access to the path and not to the event this path belongs to. For this reason, we propose to add a new interface feature PATH on the syntactic trees. This feature is accessible at nodes where path modifiers could adjoin, in particular on the VP node preceding the PP argument in Figure 26. The feature PATH has to appear as well on the VP nodes in *n0Vpp(dir)* trees, except that here the path is part of the main event.

With the additional interface feature PATH, we have to revise the directional PP modifier trees; they now access the path they refer to via this feature. The associated frame relates PATH to the AT-REGION of the NP; see Figure 27.

Figure 27: Revised elementary tree for *along*



5.3 *MG decomposition of directed motion and caused motion constructions*

We have already introduced metagrammar classes for elementary intransitive and transitive constructions in Section 4.2. We will now ex-

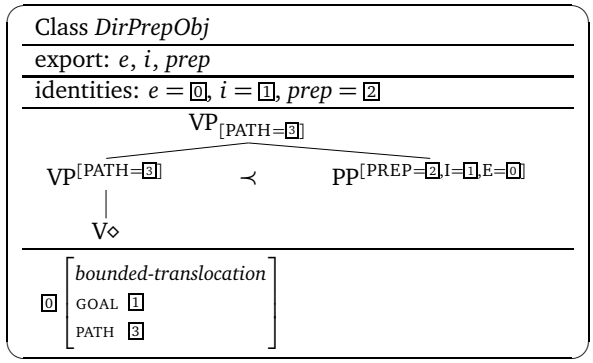


Figure 28:
MG class for a directional PP object

tend these classes in order to cover the directed motion constructions presented in the previous section.

The MG classes for these constructions are given in Figures 28 and 29. In addition to what we have seen in Section 4.2, we now allow the definition of export variables within a class. These export variables are visible to other classes using this class and can then be used to identify nodes between classes. The class for the directional prepositional object, *DirPrepObj* contributes the goal of some directed motion event. The export variable *prep* is not relevant for the directed motion case; it will serve to constrain the preposition in the prepositional object case treated in Section 6 below. Crucially, in the *DirPrepObj* class, the event described here need not be the event denoted by the verb. Therefore, the event identifier 0 is not linked via an E feature to the verb. Depending on the context in which we use this class, this event is either the main event of the verb (28-a) or an embedded event (28-b).

(28) a. Mary walked into the house
b. Mary threw the ball into the hole.

The solution is to make the event in question accessible via the declaration of export variables. For the combination of *DirPrepObj* with the intransitive or with the transitive class, we assume that the two classes *n0V* and *n0Vn1* in turn have an export variable *e* that is the event frame variable linked to the V node. The class *n0Vpp(dir)* (Figure 29, left side) for constructions without a direct object (as in (28-a)) is rather simple since the directional PP adds a participant to the event denoted by the verb. Obviously, this yields the unanchored tree for

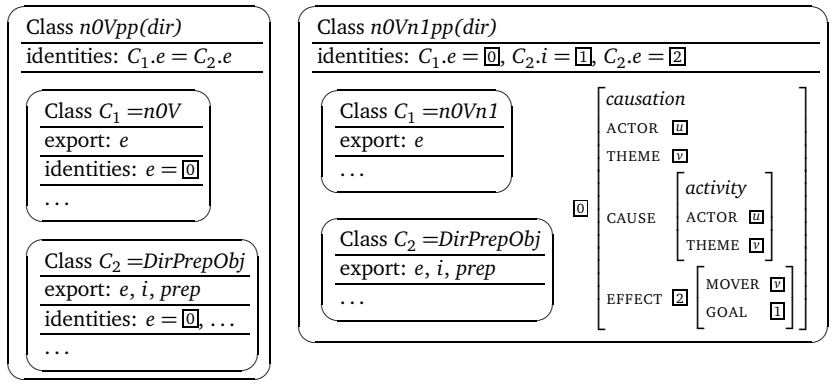


Figure 29: MG classes for directed motion constructions

walked as used in (28-a). Cases such as (28-b) are more complex since they involve an embedding of the event in which the directional PP participates. The class *n0Vn1pp(dir)* (Figure 29, right side) is for constructions as in (28-b) which have a direct object and a directional PP. It identifies the directed motion event of the PP with the event embedded under EFFECT, via the e export variable of the *DirPrepObj* class (identity $C_2.e = \boxed{2}$).

5.4 *Summary*

In this section, we have presented an analysis of verbs and constructions of directed motion and caused motion using LTAG and frame semantics. We have shown how to decompose elementary constructions into, first, the unanchored tree and its semantics and the lexical entry and, second, into smaller syntactic and semantic fragments of which the unanchored elementary construction is built.

We have seen that the metagrammar architecture of LTAG allows us to capture components which several elementary constructions have in common, for instance the class *DirPrepObj*, which contributes the syntactic slot of the goal of a *bounded-translocation*. This small piece of syntactic structure and related meaning can be used in different ways in larger classes, depending on the embedding of the *bounded-translocation* event.

The decoration of the syntactic trees with interface features allows us to access different nodes in a semantic frame, making them accessible for semantic composition. Summarizing, this first case study has demonstrated the flexibility with respect to semantic composition

and the capability of factorization and generalization offered by our LTAG syntax-semantics interface architecture.

6 APPLICATION II: THE DATIVE ALTERNATION

6.1 *Caused possession vs. caused motion construction*

The English dative alternation is concerned with verbs like *give*, *send*, and *throw* which can occur in both the double object (DO) and the prepositional object (PO) construction as exemplified by (29-a) and (29-b), respectively. The PO construction is closely related to the caused motion construction discussed in the previous section, except that the preposition in the PO construction is always *to*.

(29) a. John sent Mary the book.
b. John sent the book to Mary.

The two constructions are traditionally associated with a 'caused possession' (29-a) and 'caused motion' (29-b) interpretation, respectively (see, e.g., Goldberg (1995)). These two interpretations have often been analyzed by decompositional schemas of the type shown in (30-a) and (30-b).

(30) a. [[x ACT] CAUSE [y HAVE z]]
b. [[x ACT] CAUSE [z GO TO y]]

In a similar vein, Krifka (2004) uses event logical expressions of the sort shown in (31) for distinguishing the two interpretations. Note that (31-b) is very close to the semantic frame used in the preceding sections for caused motion.[17]

(31) a. $\exists e \exists s [\text{AGENT}(e, x) \wedge \text{CAUSE}(e, s) \wedge s : \text{HAVE}(y, z)]$
b. $\exists e \exists e' [\text{AGENT}(e, x) \wedge \text{CAUSE}(e, e') \wedge \text{MOVE}(e') \wedge \text{THEME}(e', y) \wedge \text{GOAL}(e', z)]$

The differences between the DO and the PO constructions and their respective interpretations span a wider range of options than those

[17] Recall the difference between the relational uses of CAUSE in (30) and (31) and our use of CAUSE as an event attribute that singles out the cause component of a causation event; cf. Section 3.2.

described so far. Rappaport Hovav and Levin (2008) distinguish three types of alternating verbs based on differences in the meaning components they lexicalize: *give*-type (*lend*, *pass*, etc.), *send*-type (*mail*, *ship*, etc.), and *throw*-type verbs (*kick*, *toss*, etc.).[18] They provide evidence that verbs like *give* have a caused possession meaning in both kinds of constructions. The *send* and *throw* verbs, by comparison, lexically entail a change of location and allow both interpretations depending on the construction in which they occur. The *send* and *throw* verbs differ in the meaning components they lexicalize: *send* lexicalizes caused motion towards a goal, whereas *throw* encodes the caused initiation of motion and the manner in which this is done. A goal is not lexicalized by *throw* verbs, which accounts for the larger range of directional PPs allowed for these verbs (cf. Section 5.2.4).

Beavers (2011) proposes a formally more explicit explanation of these observations based on a detailed analysis of the different types of results that determine the aspectual behavior of the verbs in question. He identifies four main types of results for ditransitive verbs: loss of possession, possession, leaving, and arrival. These results are associated with two different dimensions or "scales": the first two results belong to the "possession scale", while the latter two results are associated with a location or path scale. Only *give* verbs lexicalize actual possession as a result. *Send* verbs and *throw* verbs, by contrast, do not encode actual possession nor do they encode prospective possession when combined with the PO construction. The result condition that makes these verbs telic even if the theme does not arrive at the goal or recipient is the leaving of the theme from the actor. That is, the aspectually relevant result consists in leaving the initial point of the underlying path scale.

With respect to the goals of this article, the main question is how the constructional meaning interacts with the lexical meaning. The DO construction encodes only *prospective* possession. Actual possession must be contributed by the lexical semantics of the verb. This is the case for *give* verbs, which explains why there is no difference between the DO and the PO constructions for these verbs as far as caused

[18] For simplicity, we do not consider verbs of communication (*tell*, *show*, etc.) nor do we take into account differences in modality as between *give* and *offer* (cf. Koenig and Davis 2001).

	lexical meaning					PO pattern	DO pattern
	#args	result	punctual	manner	motion	(◊arrive)	(◊receive)
give	3	receive	yes	no	no	receive (arrive)	receive
hand	3	receive	yes	yes	yes	receive (arrive)	receive
send	3	leave ◊arrive	yes	no	yes	◊arrive	◊receive
throw	2	leave	yes	yes	yes	◊arrive	◊receive
bring	3	arrive	no	no	yes	arrive	receive

Table 1: Semantic classes of verbs in interaction with the DO and PO patterns

possession is concerned. All other alternating ditransitive verbs show such a difference since only the DO pattern implies prospective possession.[19] Beavers (2011) draws a distinction between different types of caused possession verbs. Verbs such as *give* encode pure caused possession without motion necessarily being involved. Verbs like *hand* and *pass*, by comparison, imply actual possession but also arrival of the theme via motion. The possession scale is "two-point" or "simplex" in that its only values are non-possession and possession. It follows that verbs which lexicalize caused possession are necessarily punctual since there are no intermediate "points" on this scale. In contrast to *send* and *throw*, verbs like *bring* and *take* do encode arrival of the theme at the goal (Beavers 2011). That is, for these verbs of accompanied motion, the arrival is actual and not only prospective, and this property can be regarded as lexicalized since the verbs in question are basically three-place predicates. Verbs like *carry* and *pull*, which lexicalize a "continuous imparting of force", behave differently (Krifka 2004). They are basically two-argument verbs, i.e., they do not lexicalize a goal, and they are usually regarded as being incompatible with the DO pattern.[20]

[19] The story is a bit more complicated: if the goal of the PO construction is human or human-like (e.g., an institution), there seems to be a conventional implicature that the (prospective) goal is also a (prospective) recipient, that is, (prospective) possession seems to be entailed in cases like *send the package to London*.

[20] Krifka (2004) explains this fact by pointing out that the continuous imparting of force is a "manner" component that is not compatible with a caused possession interpretation. The strict exclusion of the DO pattern for verbs indicat-

In sum, the DO and PO constructions strongly interact with the lexical semantics of the verb.[21] Table 1, which builds on Beavers' analysis, gives an overview of the contribution of the lexicon and the constructions. Prospectivity is indicated by '◊'.

6.2 *Analysis of DO and PO*

6.2.1 Frame representations

For some of the verbs listed in Table 1, possible frame semantic representations are given in Figure 30. We have added a further event type *undergoing* which comes with a participant role THEME (32-a) and which is incompatible with *activity* (32-b). The main purpose of this extension in the current context is to characterize the MOVER of a non-active motion event as a THEME (32-c), in much the same way as the MOVER of active motion has been co-classified as ACTOR (cf., e.g., Figure 16).

(32) a. $\mathit{undergoing} \preceq \mathit{event} \wedge \textsc{theme} : \top$
b. $\mathit{undergoing} \wedge \mathit{activity} \preceq \bot$
c. $\mathit{undergoing} \wedge \mathit{motion} \preceq \textsc{theme} \doteq \textsc{mover}$

Consider the frame for *send*. The bounded translocation subframe encodes motion towards the goal without necessarily implying arrival. The motion is non-active, i.e., of type *undergoing*, which means that the mover is the theme of the event. The representation for *throw* differs from that for *send* in that *throw* lexicalizes a certain manner of activity.

---

ing accompanied motion like *carry* has been called into question by Bresnan and Nikitina (2010) on the basis of corpus evidence. Building on Krifka's approach, Beavers (2011, pp. 46f) explains the low frequency of the DO pattern by distinguishing between the different kinds of 'have' relations involved: the 'have' of control by the actor during the imparting of force and the final 'have' of possession by the recipient. He proposes a "naturalness constraint" which largely, but not totally, excludes caused possession in cases where the actor has control of the theme at the final point of the event. Conditions of this type would naturally go into a more detailed frame-semantic analysis elaborating on the ones given in this paper.

[21] The DO construction with the caused possession interpretation also occurs for creation verbs with benefactive extension as in *bake her a cake*. The corresponding PO pattern requires a *for*-PP, which will not be taken into account in this paper.

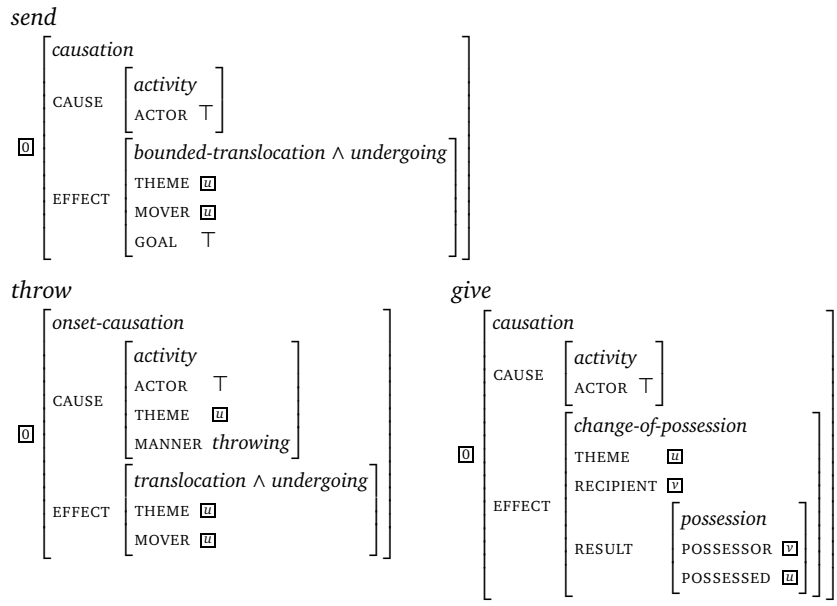


Figure 30: Possible frame representations for some of the lexical items in Table 1.

Moreover, it is inherent in the given representation that the destination of the entity thrown is not part of the lexical meaning of *throw*. Concerning the semantics of *give*, we have a caused change of possession that results in an actual *possession* state. The embedded event type *change-of-possession* introduces a participant role RECIPIENT (33-a).

(33) a. *change-of-possession* $\preceq$ *undergoing* $\wedge$ RECIPIENT : $\top$
b. RECIPIENT : $\top \preceq$ RECIPIENT $\doteq$ GOAL

We furthermore assume that a RECIPIENT can be described as a kind of GOAL (33-b).

6.2.2 Constructions

The PO construction is analyzed as a caused motion construction with a *to*-PP. Some verbs allowing the DO-PO alternation can also be used in a general caused motion construction (tree family *n0Vn1PP(dir)*); see (34).

(34) a. He sends the boy into the house.
b. He throws the ball into the basket/at the boy.

The base trees of the DO and PO families involved in the alternation are depicted on the left side of Figure 31. The fact that the preposition is required to be *to* is encoded in the PREP feature on the PP. The DO tree is flatter since in this construction, modifiers between the verb and the first NP object or between the first NP and the second are not possible.

The semantics of the DO construction is a (prospective) caused possession meaning which gets further constrained when being linked to a specific lexical anchor. More concretely, a RESULT feature is possible but not obligatory for events of type *change-of-possession*. Figure 31a shows how the unanchored tree is linked to its semantic frame. Again, the identities between the interface features I in the syntactic tree and the thematic roles in the semantic frame provide the correct argument linking. The semantics of the PO construction differs in that it triggers a caused motion instead of a caused possession interpretation; see Figure 31b.

6.2.3 Lexical anchoring

Anchoring the trees from Figure 31 means that the lexical anchor is substituted into the anchor node and thereby contributes parts of a semantic frame. The example in Figure 32 shows the lexical anchoring of the PO construction with the anchor *throws*. The resulting anchored elementary tree has a semantic frame that is the unification of the frames 4 and 0. In a similar way, caused possession verbs like *give* can anchor the DO construction.

Now, what happens if *throw* or *send* try to anchor the DO construction? That is, how can, e.g., the frame of *send* (cf. Figure 30) that represents a caused directed motion be unified with the frame of the DO construction which represents a caused change of possession (Figure 31a). The meaning of the combined frame (i.e., of the DO construction anchored with *sends*) is, roughly, a causation with effects along two dimensions: there is a directed motion of the theme and at the same time the theme undergoes a change of possession. In the model presented here, this double perspective can be captured by assuming that the event types *bounded-translocation* and *change-of-possession* do not exclude each other.[22] Hence, the effected event can be charac-

[22] An alternative solution that keeps these types disjoint would be to use set-

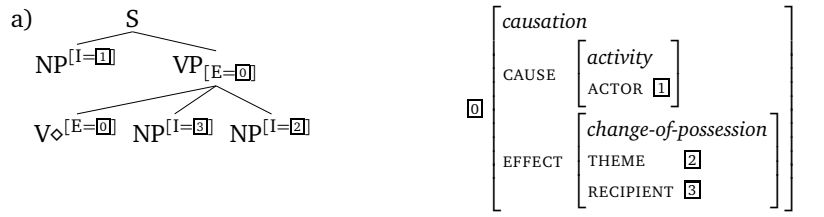
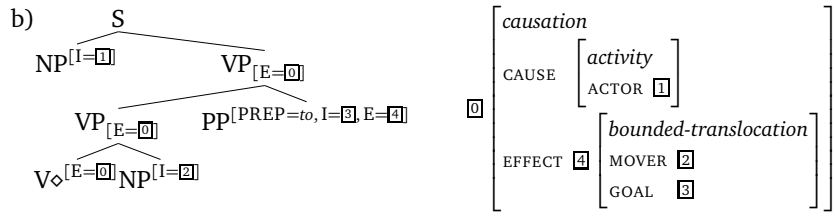


Figure 31: Unanchored elementary trees and semantics of the DO and PO constructions



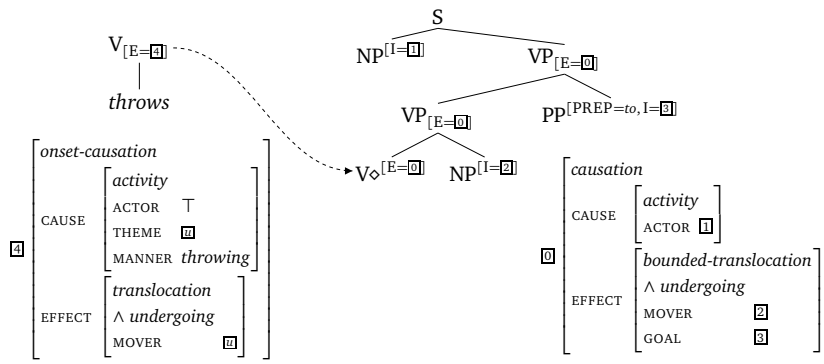


Figure 32: Lexical selection of the elementary tree for *throws* in the PO construction

terized by a conjunction of these types. The appropriate matching of the semantic roles is enforced by the constraints (32-c) and (33-b). The result of the unification is given in Figure 33. A participant can thus have different semantic roles that reflect the ways in which it is involved in the different characterizations of the event.

6.3 *MG decomposition*

We will now consider the metagrammar classes needed for the dative alternation, i.e., for the DO and PO constructions. The factorization

valued attributes, which requires however a specific definition of subsumption for these sets. Another option could be to use different attributes for the different dimensions along which an event is described, for instance LOC-ASPECT and POSS-ASPECT in our case.

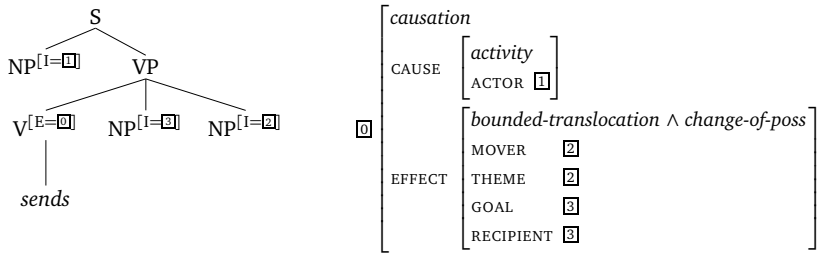


Figure 33: Anchored tree for *sends* with the DO construction

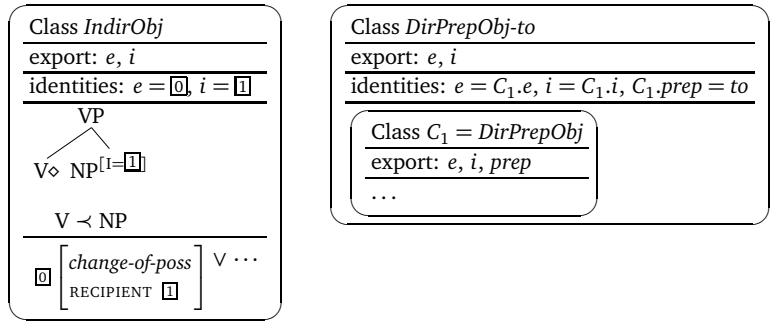


Figure 34: MG classes for indirect object and directional prepositional object

of grammatical information in the metagrammar enables us to generalize from the two phenomena that we deal with in this paper and to use the class for directional PP arguments given in Section 5.3 in both the prepositional object construction of the dative alternation and constructions with verbs of directed motion.

The classes for the indirect object and the prepositional object are given in Figure 34. A dative object (class *IndirObj*) can contribute the recipient of a change of possession event. This is not the only way an indirect object can contribute a participant to an event. Note that, according to the syntactic tree description, the NP node must immediately follow the verb node, in contrast to the NP node of a direct object that stands only in a (not necessarily immediate) linear precedence relation to the verb. The class *DirPrepObj-to* simply uses the *DirPrepObj* given above and specifies in addition that the preposition has to be *to*.

Crucially, in these classes, as in the directional PP class, the event described in the class need not be the event denoted by the verb. Therefore, again, the event identifier 0 is not linked via an E feature to the verb. Depending on the context in which we use the classes, this event is either the event of the verb or an embedded event.

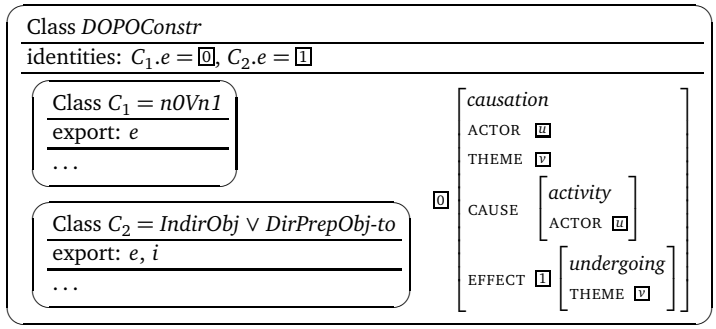


Figure 35: MG class for the alternation between DO and PO constructions

Now let us inspect the way these classes combine with the transitive verb class *n0Vn1* in order to build our unanchored DO and PO trees. The class for the DO-PO alternation is given in Figure 35. We can capture both constructions in a single class that tells us that we combine the transitive class with either *IndirObj* or *DirPrepObj-to*. The result is a causation involving an action performed by the actor of the transitive class. This causation has an effect on the theme of the transitive class. The nature of this effect depends on the class used for the third argument. In the DO case (*IndirObj*) it is a change-of-poss while in the PO case (*DirPrepObj-to*) it is a directed motion.

Note that the PO construction is actually slightly more restricted than the caused motion construction with a directional PP, not only with respect to the prepositions allowed. The NP of the directional PP, even if it is a location, receives a kind of institutional reading. Therefore, purely locational specifications such as *the house* are odd here:

(35) a. She sent the package to London.
b. ?She sent the letter to the house.

Such constraints could be modelled via restrictions on the possible goals. Either the type could be restricted or certain features could be required for the GOAL value in the *DOPOConstr* class. For this paper, we leave the detailed modelling of these constraints aside.

6.4 *Further issues*

It goes without saying that a full account of the dative alternation has to cope with many more phenomena than the distinction between

caused motion and caused possession interpretations and their sensitivity to the lexical semantics of the head verb. The distribution of the DO and PO variants of the alternation is known to be influenced by various other factors, including discourse structure effects, heaviness constraints, and the definiteness, pronominality, and animacy of recipient and theme (cf. Bresnan and Ford 2010). Correspondingly, a full grammar model would have an information structure component, ordering constraints which are sensitive to constituent length, and so on, and, in addition, would allow for defeasible and probabilistic constraints. While our grammar framework seems to be well-suited for implementing requirements of this sort, they are beyond the scope of our study here, which is primarily concerned with modelling the influence of narrow verb classes on constructional form and meaning.

7 COMPLEXITY CONSIDERATIONS

Concerning computational complexity, we have to consider the two main processing components of our architecture: metagrammar compilation and parsing. During metagrammar compilation, we compute a finite set of minimal models. For the syntactic tree, the search for minimal models means that all nodes and edges in the models have to be present in the descriptions given in the metagrammar. The current XMG implementation (Crabbé *et al.* 2013) already provides this model-building step. For the frame descriptions, a frame must first be built containing all nodes and edges described in the MG classes. In a second step, the set of constraints on frames has to be checked on this particular frame, which might lead to additional type assignments and even additional edges and nodes. In other words, we have to compute the closure with respect to the constraints. The tractability of this step depends heavily on the nature of these constraints. For instance, in order to ensure the existence of finite minimal models, we need to avoid constraint loops (see also Carpenter 1992, pp. 95ff). Tying down the exact conditions on the constraint system that make it well-behaved with respect to model construction is part of current research. Note that the metagrammar compilation can be preprocessed and is independent from the size of the input string. Therefore its complexity matters less than the complexity of unification during parsing.

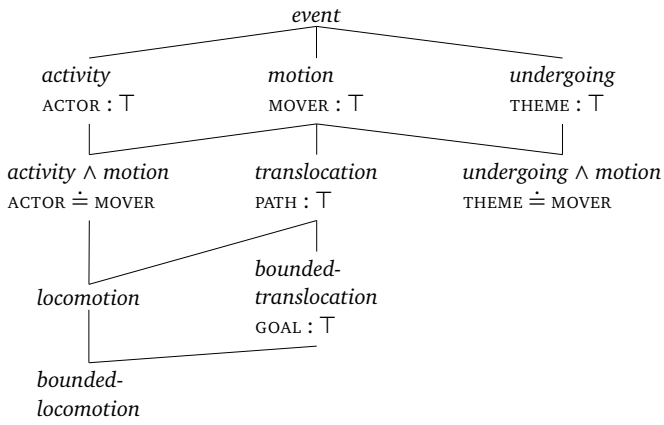


Figure 36:
Partial sketch of constraints on event types

During parsing, we have to build larger trees via substitution and adjunction. For this, algorithms of complexity $\mathcal{O}(n^6)$ with n the length of the input string are known (Vijay-Shanker and Joshi 1985; Joshi and Schabes 1997). On the semantic side, substitution and adjunction go along with the unification of frames (base-labelled feature structures) as defined in Section 3.3.1. As pointed out by Hegner (1994, pp. 136ff), the complexity of the unification of base-labelled feature structures is close to linear in the number of nodes. In fact, Hegner (1994, ibid.) shows that the complexity increases only slightly if the resulting feature structure is moreover required to satisfy a finite number of Horn descriptions. Recall that this is what we need in our approach since the unification of two frames may activate additional Horn constraints. But there is a caveat here. Hegner's result presumes a *finite* number of Horn descriptions while (universal) Horn constraints can give rise to an infinite number of them (cf. Section 3.3.4).

One way to keep this tractable is to make sure that the constraints under consideration do not introduce new nodes to the structure. Then the number of generated descriptions to be taken into account is finite. A new edge and a new node (and possibly further new edges and nodes) would for instance be added if we had a constraint in our system saying that if a frame is of type t_1 and of type t_2, then some additional attribute f has to be added, i.e., $t_1 \wedge t_2 \preceq f : \top$. So far, we do not have such constraints in our system. (The typical situation is illustrated by the partial sketch in Figure 36.) None of the conjunctive types introduces a new feature. Therefore, we make the assumption

that constraints of this type are not allowed, and that, consequently, the frames obtained during parsing do not contain more nodes and edges than the union of the frames involved in the derivation. Constraints on relations, such as the transitivity of *part-of*, must of course be taken with care, too. But again, as long as no new nodes are added by the constraints involved, the complexity remains polynomial.

8 CONCLUSION

In this paper, we introduced an LTAG-based syntax-semantics interface with a fine-grained frame-based semantics. We have shown that this architecture provides the means to associate a detailed decomposition and composition of syntactic building blocks with a parallel decomposition and composition of meaning components. Due to its various possibilities for decomposing elementary trees and because of its extended domain of locality, LTAG allows one to pair not only lexical items with lexical meaning but also constructions with their meaning contributions. Furthermore, due to the metagrammatical specification of TAG elementary trees, the meaning contributions of single argument realizations and of their combinations can be described in a principled way, in parallel to a similar decomposition of the syntactic elementary trees.

We applied the framework to the case of directed motion expressions, and we have shown how to capture the various ways a directional PP adds information about the path of the motion event. Furthermore, we have demonstrated how to model syntax and semantics of the dative alternation, separating constructional aspects of meaning from lexical ones. Finally, we have presented a metagrammatical decomposition of our constructions that allows for an elegant meaning factorization which brings the two phenomena together by characterizing the parts that they have in common.

Besides giving a detailed frame-based analysis of lexical and constructional meaning aspects, our approach integrates this into a syntax-semantics interface. Via substitution and adjunction, the frame-based characterization of the events described by entire sentences can be compositionally derived.

The frames we use for semantics are typed feature structures that do not necessarily have a unique root, that allow to access any node

in the feature structure via designated base nodes, and that allow for relations between nodes, besides the usual functional attributes. Such structures can be formalized as base-labelled feature structures. We have presented a feature logic that allows us to specify these frames and to express general constraints on them. We also described criteria for these constraints such that semantic composition (i.e., unification) under constraints is tractable.

ACKNOWLEDGEMENTS

The research presented here has been supported by the Collaborative Research Center 991 funded by the German Research Foundation (DFG). We would like to thank Timm Lichte, Maribel Romero, and three anonymous reviewers for their valuable comments on earlier drafts of this article.

REFERENCES

Anne Abeillé (2002), *Une Grammaire Électronique du Français*, CNRS Editions, Paris.

Anne Abeillé and Owen Rambow (2000), Tree Adjoining Grammar: An Overview, in Anne Abeillé and Owen Rambow, editors, *Tree Adjoining Grammars: Formalisms, Linguistic Analyses and Processing*, pp. 1–68, CSLI Publications, Stanford, CA.

Srinivas Bangalore and Aravind K. Joshi (2010), Introduction, in S. Bangalore and A. K. Joshi, editors, *Supertagging: Using Complex Lexical Descriptions in Natural Language Processing*, pp. 1–31, MIT Press, Cambridge, MA.

John Beavers (2011), An Aspectual Analysis of Ditransitive Verbs of Caused Possession in English, *Journal of Semantics*, 28:1–54.

Benjamin K. Bergen and Nancy Chang (2005), Embodied Construction Grammar in simulation-based language understanding, in Jan-Ola Östman and Mirjam Fried, editors, *Construction Grammars. Cognitive Grounding and Theoretical Extensions*, pp. 147–190, John Benjamins, Amsterdam.

Patrick Blackburn (1993), Modal Logic and Attribute Value Structures, in Maarten de Rijke, editor, *Diamonds and Defaults*, pp. 19–65, Kluwer, Dordrecht.

Patrick Blackburn and Johan Bos (2003), Computational Semantics, *Theoria*, 18(1):27–45.

Joan BRESNAN and Marilyn FORD (2010), Predicting Syntax: Processing Dative Constructions in American and Australian Varieties of English., *Language*, 86(1):186–213.

Joan BRESNAN and Tatiana NIKITINA (2010), The Gradience of the Dative Alternation, in Linda UYECHI and Lian Hee WEE, editors, *Reality Exploration and Discovery: Pattern Interaction in Language and Life*, pp. 161–184, CSLI Publications, Stanford, CA.

Marie-Hélène CANDITO (1999), *Organisation modulaire et paramétrable de grammaires électroniques lexicalisées. Application au français et à l'italien*, Ph.D. thesis, Université Paris 7, Paris.

Bob CARPENTER (1992), *The Logic of Typed Feature Structures*, Cambridge University Press, Cambridge.

Ann COPESTAKE, Dan FLICKINGER, Carl POLLARD, and Ivan A. SAG (2005), Minimal Recursion Semantics: An Introduction, *Research on Language and Computation*, 3:281–332.

Benoit CRABBÉ (2005), Grammatical Development with XMG, in Philippe BLACHE, Edward STABLER, Joan BUSQUETS, and Richard MOOT, editors, *Proceedings of LACL 2005*, Lecture Notes in Artificial Intelligence 3492, pp. 84–100, Springer, Berlin.

Benoit CRABBÉ and Denys DUCHIER (2005), Metagrammar Redux, in Henning CHRISTIANSEN, Peter Rossen SKADHAUGE, and Jørgen VILLADSEN, editors, *Constraint Solving and Language Processing*, Lecture Notes in Computer Science 3438, pp. 32–47, Springer, Berlin.

Benoit CRABBÉ, Denys DUCHIER, Claire GARDENT, Joseph LE ROUX, and Yannick PARMENTIER (2013), XMG: eXtensible MetaGrammar, *Computational Linguistics*, 39(3):591–629.

David DOWTY (1979), *Word Meaning and Montague Grammar*, D. Reidel, Dordrecht.

David DOWTY (2003), The dual analysis of adjuncts/complements in Categorial Grammar, in Ewald LANG, Claudia MAIENBORN, and Cathrine FABRICIUS-HANSEN, editors, *Modifying Adjuncts*, Interface Explorations 4, pp. 33–66, Mouton de Gruyter, Berlin.

Veronika EHRICH (1996), Verbbedeutung und Verbgrammatik: Transportverben im Deutschen, in Ewald LANG and Gisela ZIFONUN, editors, *Deutsch - typologisch*, pp. 229–260, de Gruyter, Berlin.

Carola ESCHENBACH, Ladina TSCHANDER, Christopher HABEL, and Lars KULIK (2000), Lexical Specifications of Paths, in Christian FREKSA, Wilfried BRAUER, Christopher HABEL, and Karl Friedrich WENDER, editors, *Spatial Cognition II*, Lecture Notes in Computer Science 1849, pp. 127–144, Springer, Berlin.

Charles J. FILLMORE (1982), Frame Semantics, in *Linguistics in the Morning Calm*, pp. 111–137, Hanshin Publishing Co., Seoul.

Charles J. FILLMORE (1986), Pragmatically controlled zero anaphora, in *Berkeley Linguistics Society 12*, pp. 95–107.

Charles J. FILLMORE (2007), Valency Issues in FrameNet, in Thomas HERBST and Katrin GÖTZ-VOTTELER, editors, *Valency: Theoretical, Descriptive and Cognitive Issues*, pp. 129–160, Mouton de Gruyter, Berlin.

Charles J. FILLMORE, Christopher R. JOHNSON, and Miriam R. L. PETRUCK (2003), Background to FrameNet, *International Journal of Lexicography*, 16(3):235–250.

William A. FOLEY and Robert D. VAN VALIN (1984), *Functional Syntax and Universal Grammar*, Cambridge University Press, Cambridge.

Robert FRANK (2002), *Phrase Structure Composition and Syntactic Dependencies*, MIT Press, Cambridge, MA.

Claire GARDENT and Laura KALLMEYER (2003), Semantic Construction in FTAG, in *Proceedings of EACL 2003*, pp. 123–130.

Berit GEHRKE (2008), *Ps in Motion. On the semantics and syntax of P elements and motion events*, LOT, Utrecht.

Adele E. GOLDBERG (1995), *Constructions: A Construction Grammar Approach to Argument Structure*, University of Chicago Press, Chicago, IL.

Adele E. GOLDBERG and Ray JACKENDOFF (2004), The English resultative as a family of constructions, *Language*, 80:532–568.

Fritz HAMM, Hans KAMP, and Michiel VAN LAMBALGEN (2006), There is no opposition between Formal and Cognitive Semantics, *Theoretical Linguistics*, 32(1):1–40.

Stephen J. HEGNER (1994), Properties of Horn Clauses in Feature-Structure Logic, in C. J. RUPP, Michael A. ROSNER, and Rod L. JOHNSON, editors, *Constraints, Language and Computation*, pp. 111–147, Academic Press, San Diego, CA.

Ray JACKENDOFF (1991), Parts and Boundaries, *Cognition*, 41:9–45.

Aravind K. JOSHI and Yves SCHABES (1997), Tree-Adjoining Grammars, in Grzegorz ROZENBERG and Arto SALOMAA, editors, *Handbook of Formal Languages. Vol. 3: Beyond Words*, pp. 69–123, Springer, Berlin.

Laura KALLMEYER and Aravind K. JOSHI (2003), Factoring Predicate Argument and Scope Semantics: Underspecified Semantics with LTAG, *Research on Language and Computation*, 1(1/2):3–58.

Laura KALLMEYER and Maribel ROMERO (2008), Scope and Situation Binding in LTAG using Semantic Unification, *Research on Language and Computation*, 6(1):3–52.

Ingrid KAUFMANN (1995), *Konzeptuelle Grundlagen semantischer Dekompositionsstrukturen. Die Kombinatorik lokaler Verben und prädikativer Argumente*, Niemeyer, Tübingen.

Jean-Pierre KOENIG and Anthony R. DAVIS (2001), Sublexical Modality and the Structure of Lexical Semantic Representations, *Linguistics and Philosophy*, 24:71–124.

Karsten KONRAD (2004), *Model Generation for Natural Language Interpretation and Analysis*, Springer, Berlin.

Manfred KRIFKA (2004), Semantic and pragmatic conditions for the Dative Alternation, *Korean Journal of English Language and Linguistics*, 4:1–32.

Anthony S. KROCH (1989), Asymmetries in long-distance extraction in a Tree Adjoining Grammar, in Mark R. BALTIN and Anthony S. KROCH, editors, *Alternative Conceptions of Phrase Structure*, pp. 66–98, University of Chicago Press, Chicago, IL.

Beth LEVIN and Malka RAPPAPORT HOVAV (2005), *Argument Realization*, Cambridge University Press, Cambridge.

Timm LICHTE, Alexander DIEZ, and Simon PETITJEAN (2013), Coupling Trees and Frames through XMG, in Denys DUCHIER and Yannick PARMENTIER, editors, *ESSLLI 2013 Workshop on High-level Methodologies for Grammar Engineering*, pp. 37–48.

Sebastian LÖBNER (2014), Evidence for Frames from Human Language, in Thomas GAMERSCHLAG, Doris GERLAND, Rainer OSSWALD, and Wiebke PETERSEN, editors, *Frames and Concept Types*, pp. 23–67, Springer, Dordrecht.

Claudia MAIENBORN (2011), Event Semantics, in Claudia MAIENBORN, Klaus VON HEUSINGER, and Paul PORTNER, editors, *Semantics. An International Handbook of Natural Language Meaning. Volume 1*, pp. 802–829, Mouton de Gruyter, Berlin.

Inderjeet MANI and James PUSTEJOVSKY (2012), *Interpreting Motion. Grounded Representations for Spatial Language*, Oxford University Press, Oxford.

Stefan MÜLLER (2006), Phrasal or Lexical Constructions, *Language*, 82(4):850–883.

Rebecca NESSON and Stuart M. SHIEBER (2006), Simpler TAG Semantics Through Synchronization, in Shuly WINTNER, editor, *Proceedings of the 11th Conference on Formal Grammar*, pp. 129–142.

Rainer OSSWALD (1999), Semantics for Attribute-Value Theories, in Paul DEKKER, editor, *Proceedings of the Twelfth Amsterdam Colloquium*, pp. 199–204, University of Amsterdam, ILLC, Amsterdam.

Rainer OSSWALD and Robert D. VAN VALIN (2014), FrameNet, Frame Structure, and the Syntax-Semantics Interface, in Thomas GAMERSCHLAG,

Doris GERLAND, Rainer OSSWALD, and Wiebke PETERSEN, editors, *Frames and Concept Types*, pp. 125–156, Springer, Dordrecht.

Terence PARSONS (1990), *Events in the Semantics of English*, MIT Press, Cambridge, MA.

Malka RAPPAPORT HOVAV and Beth LEVIN (1998), Building Verb Meanings, in Miriam BUTT and Wilhelm GEUDER, editors, *The Projection of Arguments: Lexical and Compositional Factors*, pp. 97–134, CSLI Publications, Stanford, CA.

Malka RAPPAPORT HOVAV and Beth LEVIN (2008), The English dative alternation: A case for verb sensitivity, *Journal of Linguistics*, 44:129–167.

Mike REAPE (1994), A Feature Value Logic with Intensionality, Nonwellfoundedness and Functional and Relational Dependencies, in C. J. RUPP, Michael A. ROSNER, and Rod L. JOHNSON, editors, *Constraints, Language and Computation*, pp. 77–110, Academic Press, San Diego, CA.

William C. ROUNDS (1997), Feature Logics, in Johan VAN BENTHEM and Alice TER MEULEN, editors, *Handbook of Logic and Language*, pp. 475–533, North-Holland, Amsterdam.

Ivan SAG (2012), Sign-Based Construction Grammar: An informal synopsis, in Hans BOAS and Ivan SAG, editors, *Sign-Based Construction Grammar*, pp. 61–188, CSLI Publications, Stanford.

Klaas SIKKEL (1997), *Parsing Schemata*, Springer, Berlin.

Leonard TALMY (2000a), *Toward a Cognitive Semantics. Volume I: Concept Structuring Systems*, MIT Press, Cambridge, MA.

Leonard TALMY (2000b), *Toward a Cognitive Semantics. Volume II: Typology and Process in Concept Structuring*, MIT Press, Cambridge, MA.

Robert D. VAN VALIN and Randy J. LAPOLLA (1997), *Syntax*, Cambridge University Press, Cambridge.

Robert D. VAN VALIN, Jr. (2005), *Exploring the Syntax-Semantics Interface*, Cambridge University Press, Cambridge.

Henk VERKUYL and Joost ZWARTS (1992), Time and Space in Conceptual and Logical Semantics: The Notion of Path, *Linguistics*, 30:483–511.

K. VIJAY-SHANKER and Aravind K. JOSHI (1985), Some computational properties of Tree Adjoining Grammars, in *Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics*, pp. 82–93.

K. VIJAY-SHANKER and Aravind K. JOSHI (1988), Feature Structures Based Tree Adjoining Grammar, in *Proceedings of the 12th International Conference on Computational Linguistics (COLING-88)*, pp. 714–719.

Fei XIA (2001), *Automatic grammar generation from two different perspectives*, Ph.D. thesis, University of Pennsylvania, Philadelphia, PA.

Fei XIA, Martha PALMER, and VIJAY-SHANKER (2010), Developing Tree-Adjoining Grammars with Lexical Descriptions, in S. BANGALORE and A. K. JOSHI, editors, *Supertagging: Using Complex Lexical Descriptions in Natural Language Processing*, pp. 73–110, MIT Press, Cambridge, MA.

XTAG RESEARCH GROUP (2001), A Lexicalized Tree Adjoining Grammar for English, Technical report, Institute for Research in Cognitive Science, University of Pennsylvania, Philadelphia, PA.

Jordan ZLATEV, Johan BLOMBERG, and Caroline DAVID (2010), Translocation, language and the categorization of experience, in Vyvyan EVANS and Paul CHILTON, editors, *Language, Cognition and Space*, pp. 389–418, Equinox, London.

Joost ZWARTS (2005), Prepositional Aspect and the Algebra of Paths, *Linguistics and Philosophy*, 28(6):739–779.

This work is licensed under the Creative Commons Attribution 3.0 Unported License.
http://creativecommons.org/licenses/by/3.0/

EXTERNAL REVIEWERS 2012–2013

The mainstay of any peer-reviewed journal are its reviewers, and JLM is no exception here. Each paper is reviewed by at least 3 carefully selected reviewers, including at least one representing the JLM Editorial Board. To increase reviewing anonimity, we do not give the names of the 25 JLM EB reviewers, but we would like to heartily thank them for their hard and timely work. We also express our sincere gratitude to the following 38 external reviewers for papers reviewed by the end of 2013:

Avery Andrews
Doug Arnold
Frédéric Béchet
Steven Bethard
Francis Bond
Johan Bos
Caroline Brun
Ann Copestake
Benoît Crabbé
Östen Dahl
Elisabet Engdahl
Martin Forst
Radovan Garabík
Włodzimierz Gruszczyński
Krzysztof Jassem
Wojciech Jaworski
Simin Karimi
Lauri Karttunen
Tracy Holloway King
Lothar Lemnitzer
Jadwiga Linde-Usiekniewicz
Marek Maziarz
Diana McCarthy
Kyoko Ohara
Karel Pala
Vladimír Petkevič
Miriam Petruck
Alan Prince
Frank Richter
Jason Riggle
Whitney Tabor
Reut Tsarfaty
Naushad UzZaman
Gertjan van Noord
Ruprecht von Waldenfels
Aleksander Wawer
Shûichi Yatabe
Bartosz Ziółko