LingInfo: Design and Applications of a Model for the Integration of Linguistic Information in Ontologies

Paul Buitelaar♦, Thierry Declerck♦, Anette Frank♦, Stefania Racioppa♦, Malte Kiesel♦, Michael Sintek♦, Ralf Engel♦, Massimo Romanelli♦, Daniel Sonntag♦, Berenike Loos♣, Vanessa Micelli♣, Robert Porzel♣, Philipp Cimiano∗

♦DFKI GmbH. Kaiserslautern/Saarbrücken, Germany
♣European Media Lab, Heidelberg, Germany
∗AIFB, University of Karlsruhe, Karlsruhe, Germany

paulb@dfki.de (contact address)

Abstract

To allow for a direct connection of this linguistic information for terms with corresponding classes and properties in a domain ontology, we developed a lexicon model (LingInfo) that enables the definition of LingInfo instances (each of which represents a term) for each class or property. The LingInfo model is represented by use of a meta-class, which allows for the representation of LingInfo instances with each class, where each LingInfo instance represents the linguistic features of a term for a particular class. Applications of the LingInfo model are in information extraction, dialogue analysis, and knowledge acquisition from text, i.e. in knowledge base generation and ontology learning.

1. LingInfo: Motivation and Design

To allow for automatic multilingual knowledge markup a richer representation is needed of the features of linguistic expressions (such as domain terms, their synonyms and multilingual variants) for ontology classes and properties. Currently, such information is mostly missing or represented in impoverished ways, leaving the semantic information in an ontology without a grounding to the human cognitive and linguistic domain.

Linguistic information for terms that express ontology classes and/or properties consists of lexical and context features[1], such as:

- *language-ID*: ISO-based unique identifier for the language of each term
- *part-of-speech*: representation of the part of speech of the head of the term
- *morphological and syntactic decomposition*: representation of the morphological and syntactic structure (segments, head, modifiers) of a term
- *statistical and/or grammatical context model*: representation of the linguistic context of a term in the form of N-grams, grammar rules or otherwise

To allow for a direct connection of this linguistic information for terms with corresponding classes and properties in the domain ontology, we developed a lexicon model (LingInfo) that enables the definition of LingInfo instances (each of which represents a term) for each class or property. The LingInfo model is represented by use of a meta-class `(ClassWithLingInfo)` and meta-property `(PropertyWithLingInfo)`, which allow for the representation of LingInfo instances with each class, where each LingInfo instance represents the linguistic features (`feat:lingInfo`) of a term for a particular class.

Figure 1 shows an overview of the model with example domain ontology classes and associated LingInfo instances. The domain ontology consists of the class `o:FootballPlayer` with subclasses `o:Defender` and `o:Midfielder`, each of which are instances of the meta-class `feat:ClassWithLingInfo` with the property `feat:lingInfo`.

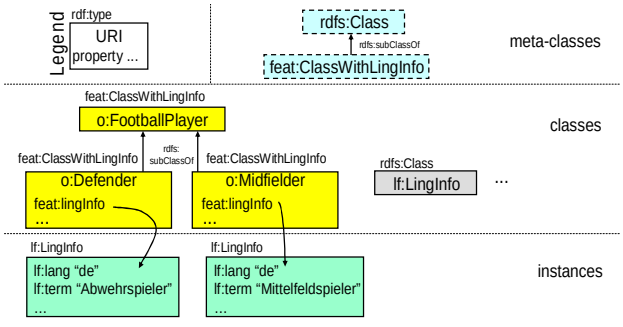

Figure 1: LingInfo model with example domain ontology classes and LingInfo instances (simplified)

Figure 2 shows a sample application of the model with a LingInfo instance (and connected 'stem', 'root' and other instances – for details see the complete LingInfo model in the appendix) that represents the decomposition of the German linguistic expression "Fußballspielers" ("of the football player"). The example shows `inst1` that represents the inflected (genitive) word form with stem "Fußballspieler" (`inst2`, "footballplayer"), which can be decomposed into "Fußball" (`inst3` , "football" with

[1] Morphosyntactic and syntactic features may be based in future versions on the (ISO-TC37/SC4-MAF and ISOTC37/SC4-SynAF) specifications. See also related documentation at the LIRICS project web site: http://lirics.loria.fr/documents.html

`semantics` "o:BallObject") and "Spieler" (`inst8` , "player), recursively continued for "Fußball" with "Fuß" and "Ball" (`inst5` and `inst7`, "foot" and "ball").

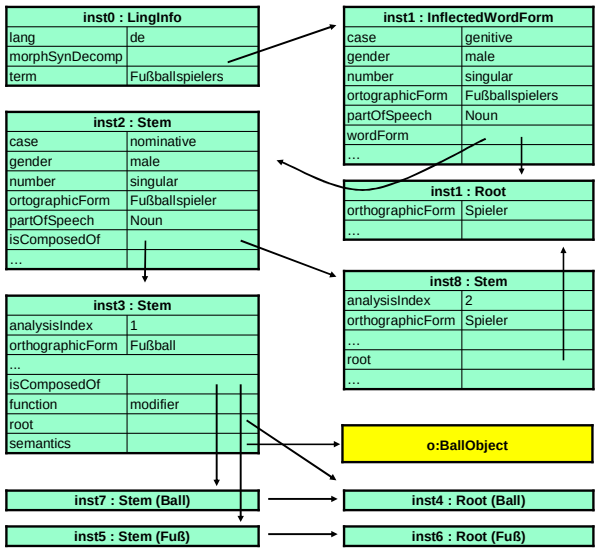

Figure 2: LingInfo instance (partial) for the morphosyntactic decomposition of "Fußballspielers"

2. Comparison with Related Work

2.1 Simple Knowledge Organization Systems

There is some overlap between the LingInfo model and the proposed SKOS[2] (Simple Knowledge Organization Systems) format for the formalized representation of thesauri. However, there is a technical and conceptual reason why SKOS does not fulfill the needs of our scenario[3].

On the technical side, SKOS uses sub-properties (`skos:prefLabel`, `skos:altLabel`) of `rdfs:label` together with `xml:lang` to attach multilingual terms to concepts. Furthermore, the RDFS specification[4] defines the range of `rdfs:label` to be `rdfs:Literal` and from the definition of `rds:subPropertyOf` follows that the range of `skos:prefLabel` and `skos:altLabel` is also (or a specialization of) `rdfs:Literal`. This is not sufficient in our scenario since we want to attach more linguistic information to classes than simple multilingual strings. This led us to the decision to use a meta-class `ClassWithFeats`, which allows us to attach complex information to classes with the properties `lingFeat` and `imgFeat`.

The conceptual problem we see with SKOS for the use in our scenario is that it mixes linguistic and semantic knowledge. SKOS uses `skos:broader` and `skos:narrower` to express "semantic" relations without clearly stating the semantics of these relations intentionally, and defines the sub-properties `skos:broaderGeneric`/`narrowerGeneric` to have class subsumption semantics (i.e., they inherit the `rdfs:subClassOf` semantics from RDFS).

Instead, the LingInfo model clearly keeps the linguistic and semantic, ontology-based knowledge representations separate: the ontology is represented using the semantic relations defined in RDFS or OWL-Full[5] with linguistic knowledge attached to classes and properties.

2.2 Wordnets and OntoWordNet

Our approach in effect integrates a domain-specific multilingual Wordnet into the ontology, although the Wordnet model does not distinguish clearly between linguistic and semantic information (Miller et al., 1995). Alternative lexicon models that are more similar to our approach include (Bateman et al. 1995) and (Alexa et al. 2002), but these concentrate on the definition of a top ontology for lexicons instead of linguistic features for domain ontology classes and properties as in our case. This is also the main difference with the proposed OntoWordNet model (Gangemi et al., 2003), which aims at merging the foundational ontology DOLCE (Gangemi et al., 2002) with WordNet to provide the latter with a formal semantics.

2.3 Lexical Markup Framework

Closest to our work are some recent initiatives of the ISO TC37/SC4[6] working group on the management of language resources, which was established in 2002 and continues the work from previous standardization initiatives, like EAGLES[7] (Expert Advisory Group on Language Engineering Standards) for morphological and syntactic annotation and ISLE[8] (International Standards for Language Engineering) for the representation of lexicon entries.

In the various initiatives of ISO TC37/SC4 the focus is on the more abstract level of meta-annotation and of frameworks supporting the creation and the exchange of annotations, data structures and resources. An important part of this work consists of the definition of procedures for the creation and maintenance of data categories for the various annotation frameworks. Data categories are formalized representations of the most relevant linguistic concepts, such as 'part of speech', 'lemma', etc.

The ISO TC37/SC4 standardization initiative that is most closely related to the LingInfo model is LMF, the Lexical Markup Framework, 'a common standardized framework for the construction of NLP lexicons' (Francopoulo et al. 2006). However, the main difference between LMF and the LingInfo model is again the level of division between linguistic and semantic knowledge. In LMF these are integrated in the same model by way of a lexical semantics slot, whereas in the LingInfo model all lexical semantics is to be found in the domain ontology - that is outside of the lexicon model per se.

As a further consequence of this approach, the LingInfo model allows also for the representation of non-linguistic, i.e. multimedia features (Buitelaar et al., 2005).

---

[2] http://www.w3.org/TR/swbp-skos-core-guide/

[3] In fact, the argumentation applies to all approaches based on `rdfs:label` and `xml:lang` for attaching multilingual labels to classes and properties.

[4] http://www.w3.org/TR/rdf-mt/

[5] OWL-Lite and OLW-DL do not support meta-classes and meta-properties (see http://www.w3.org/TR/owl-features/)

[6] http://www.tc37sc4.org

[7] http://www.ilc.cnr.it/EAGLES96/home.html

[8] http://www.ilc.cnr.it/EAGLES96/isle/

3. LingInfo in Context

3.1. The SmartWeb Project

The LingInfo model is developed and used within the SmartWeb[9] project on intelligent mobile information services for various domains, with a focus on soccer and the World Cup 2006 in particular. SmartWeb integrates question answering and ontology-based information extraction within a multimodal dialog system for a wide range of mobile devices. Information access to topical information available on the web is improved by adding machine-understandable semantics using a variety of techniques that range from semi- to fully automatic linguistic and semantic tagging to data-driven ontology learning.

LingInfo constitutes an ontology and linguistic knowledge base that provides for all other ontologies used in SmartWeb linguistic information (orthographic realizations, grammatical gender, stem and inflection) on ontology classes and properties for languages that are relevant to the SmartWeb scenario, i.e. German and English (and into some respect also French).

3.2. The SWIntO Ontology

A central component of the SmartWeb system is the integrated SWIntO ontology (Oberle et al., to appear), which consists of three layers: the upper model DOLCE (Gangemi et al., 2002), the domain-independent model SUMO (Niles and Pease, 2001) and several domain ontologies:

- **SportEvents** – As the soccer world cup 2006 will be the main application scenario, corresponding knowledge is modeled in the SportEvents ontology.
- **Navigation** – The SmartWeb user interfaces is based on mobile applications, e.g., by means of PDAs or by integration in cars or motorcycles. Navigation modeling is therefore a core requirement.
- **Discourse** - Multimodal web access is one of the core features of the SmartWeb sytem. It is therefore necessary to model user interaction in a generic way.
- **Multimedia** - The SmartWeb system will be able to display multimedia data such as live video streams. This data is described by means of an MPEG-based multimedia ontology.
- **LingInfo** – as described above

4. LingInfo Applications

The LingInfo model and instances are used in several components of the SmartWeb system, specifically of course in those components that are concerned with text analysis, i.e. in information extraction (IE) and dialogue analysis, and knowledge acquisition from text, i.e. in knowledge base generation and ontology learning.

4.1. Information Extraction from Text

The LingInfo model allows for the definition of *flexible interfaces* to linguistic processing components that ensure *consistency*. The SWIntO ontology, e.g., is interfaced with the IE system SProUT (Drozdzynski et al., 2004). Based on the information encoded in LingInfo, we automatically extract gazetteer entries for named entities, with back-references to the ontology. For terms associated with concepts, we recompile the relevant parts of the ontology, including LingInfo, into a type hierarchy used in the IE system. Thus, LingInfo information can be used to *consistently* identify and mark up (inflected) occurrences of domain-relevant terms.

The following example may illustrate this. It displays an excerpt of the SWIntO ontology that has been compiled into a type hierarchy defined in TDL[10], the representation language used by SProUT:

```
PlayerAction :< SportMatchAction.
SingleFootballPlayerAction :< PlayerAction.
FootballTeamAction :< PlayerAction.
GoalKeeperAction :< SingleFootballPlayerAction.
AnyPlayerAction :< SingleFootballPlayerAction.
```

Properties associated with these concepts are translated to TDL *attributes* of the corresponding *types*, e.g. the property `inMatch` of the SWIntO class `SportMatchAction` translates to the TDL attribute `INMATCH` that is inherited by all subtypes of the TDL type `SportMatchAction`. The SWIntO property `CommittedBy` that is defined for the SWIntO class `SingleFootballPlayerAction` translates to a corresponding TDL attribute `COMMITTEDBY` of the TDL type `SingleFootballPlayerAction`, and is again inherited by all its subtypes:

```
SportMatchAction := swinto_out &
     [INMATCH Football].
SingleFootballPlayerAction := swinto_out &
     [COMMITTEDBY FootballPlayer].
```

Multilingual (e.g. German) terms that are encoded as LingInfo instances are compiled into TDL lexical types:

```
"Teamaktion" :< FootballTeamAction.
"Spieleraktion" :< PlayerAction.
"Torwartaktion" :< GoalkeeperAction.
"Gesperrt" :< Banned.
```

SProUT extraction patterns can thus be triggered by lexical types, and define output structures that correspond directly to the classes and properties of the SWIntO ontology. For instance, the 'banned_player' rule below matches an extraction pattern for the SWIntO (SportEvents) class `BanEvent` with attributes `CommittedBy` and `InMatch` that is triggered for instance by the German LingInfo term "gesperrt".

Example sentences from the SmartWeb development corpus[11] to which this rule applies are as follows:

"… ist Petrow für die Partie gegen Schweden gesperrt."
("… has Petrow been banned for the match against Sweden")

"… ist David Trezeguet von der FIFA für zwei Spiele gesperrt worden."
("… has David Tezeguet been banned by FIFA for two matches")

[9] http://www.smartweb-projekt.de

[10] Type Description Language – see (Krieger and Schäfer 1994) for details

[11] See also http://www.dfki.de/sw-lt/olp2_dataset/

```
banned_player :>

@seek(player) & [IMPERSONATEDBY #player, INMATCHTEAM #team1]

(@seek(weekday_only) & [DOFW #dofw])? (token{0,2}
 @seek(soccer_institutions))?  token{0,3}
 @seek(game_teams) & [INTOURNAMENT #tour, TEAM2 #team2] morph & [STEM banned, SURFACE #event])

-> playeraction &
   [SPORTACTIONTYPE #event,
    COMMITTEDBY footballplayer &
            [IMPERSONATEDBY #player],
    INMATCH match &
            [INTOURNAMENT #tour, MATCHTYPE #match, TEAM1 #team1, TEAM2 #team2]].
```

4.2. Knowledge Base Generation

As described in (Buitelaar et al. 2006), the aim of the "SmartWeb Ontology-based Annotation" system (SOBA) is to automatically generate a soccer knowledge base, which is exploited in SmartWeb for knowledge-based question answering. The knowledge base is generated on the basis of information extraction with SProUT from freely available web documents on the soccer world cup – as described above. The web documents include structured as well as textual match reports and images with captions. All available text segments are linguistically annotated to extract semantic structures (class instances) that are compliant with the SWIntO ontology.

In extracting semantic structures, SOBA relies on the LingInfo model to avoid the creation of additional and redundant instances by comparing extracted names of players, countries etc. to LingInfo information of existing instances in the knowledge base.

4.3. Dialog Processing

The Smartweb dialogue integration framework (Reithinger and Sonntag 2005) integrates multiple natural language-intensive processing components such as SPIN (Engel 2005) for speech interpretation.

Usually, the rules for speech interpretation have to be written manually, but with the available LingInfo information we can generate part of the rules automatically. However, as the associated LingInfo information is not task-specific, the annotations are not always useful in a parsing context. To avoid an overgeneration of rules, so called generation rules allow a fine grained control over the rule generation. The generation rules have full access to the ontology and can exploit, e.g., the class hierarchy or the contained instances with LingInfo.

To resolve referential expressions, determiners (definite/indefinite) can be taken into account. This feature is provided by extending the LingInfo class with the property `RefProp`, which represents a definite/indefinite flag. A unit labeled as definite indicates the presence of an anaphoric reference which has to be resolved. This information is passed to FADE, which looks for the referenced item in recent user utterances, and resolves the reference. Additional syntactic information is used for disambiguation when several possible candidates for the referring expression exist.

4.4. Ontology Learning

In the ontology learning components of SmartWeb (Buitelaar et al., 2004; Schutz and Buitelaar, 2005), the representation of linguistic information for ontology classes and properties (relations) allows for the monitoring of any change in the domain model, for instance by tracking the use of soccer terms in subsequent versions of the SmartWeb development corpus.

The use of new terms or of new contexts for existing terms indicates an option for the extension or modification of the SWIntO ontology. For example, the term "Kneipe" ("pub") may be learned from a German text, as well as a potentially hyponymic relation with the term "Gebäude" ("building"). As the LingInfo information for the existing SWIntO class `Building` provides us with a corresponding LingInfo instance for the German term "Gebäude", this information can now be used to introduce a new class `Kneipe` (with a corresponding LingInfo instance for the German term "Kneipe") and integrate it into SWIntO as a subclass of `Building`.

4.5. Other Applications

Additional applications include the integration of the LingInfo model into ECtoloG (Micelli et al. 2006), an ontology that represents a formalization of construction grammar (Chang et al. 2002), and which allows only for one type of linguistic construction - i.e. pairings of form and meaning at different levels of abstraction. Since lexical constructions need linguistic information as provided by the LingInfo model, the LingInfo ontology was converted into OWL and integrated into ECtoloG. Therefore a meta-class `ClassWithLingInfo` (as sub-class of `owl:class`) was defined with the property `linginfo` that links ECtoloG classes and properties with LingInfo instances, enriching the ECtoloG classes with all necessary linguistic information as defined above.

An important challenge arising from this approach is that with the definition of a meta-class the ECtoloG ontology no longer confirms to OWL-DL but rather goes to OWL-Full, which thwarts the employment of Description Logic reasoners.

5. Lexical Acquisition for LingInfo

The LingInfo model enables *flexible interfaces*: by restricting the recompilation of LingInfo to core identifying properties (PoS, lemma, inflectional class), we can exploit a system's independent morphological

components, as in the case of SProUT, or we recompile the full range of information for systems that lack morphological processing components.

For this purpose, we are exploring different methodologies to (semi-) automatically instantiate a LingInfo model for a particular domain ontology with terms and corresponding linguistic information as described above. This is an incremental process, by which some information can be derived from annotated corpora. In this way, lexicons of tools used for annotation (e.g. Petitpierre and Russell 1995, Brants 2000, Lezius 2000) will be in effect tuned to respective domains and become fully integrated with the domain ontology.

Additionally, we can acquire syntactic information for domain-relevant terms from parsed domain corpora and/or existing syntactic lexica. The syntactic information can be defined in LingInfo, and exploited in information extraction tools. We are currently exploring the use of semantically annotated corpora, to acquire specific patterns between morphological and syntactic structures on the one hand and ontology classes on the other, based on the syntax-semantics links provided by LingInfo.

6. Current and Future Work

In current work, we are preparing the use of *deep parsing* to enhance the coverage and precision of concept recognition rules in the SProUT IE system, in particular for complex, non-local linguistic contexts that involve free word order, coordination, long distance constructions, etc. Via integration of argument structure information gained from deep parsing, SProUT recognition rules can refer to *deep syntactic* input structure, in particular, verbal arguments in non-local configurations. This will allow us to reliably identify concepts in linguistic constructions that are usually beyond the scope of shallow IE recognition systems. Our architecture for the integration of syntactic argument structure is designed as to permit integration of different parsers. The aim of future development in this area is the design of methods for semi-automatic acquisition of argument structure-based recognition rules, and the induction of argument-to-role mappings in the LingInfo model.

Other efforts are focused on the automatic enlargement of initial seed grammars in order to increase both their coverage as well as their inferential capabilities. For this a tight coupling to the Ontology Learning (described in Section 4.4) is vital to ensure consistency between the lexical semantics modeled via the grammar formalism and the descriptive conceptualization of the corresponding entities.

Further work is concerned also with pragmatic knowledge, which in a sense draws on all other knowledge sources cum contextual information. A first proposal on how to integrate such knowledge can be based on (Loos & Porzel 2004).

Acknowledgements

This research has been supported in part by the SmartWeb project, which is funded by the German Ministry of Education and Research under grant 01 IMD01 A. Thierry Declerck has been supported by the eContent project LIRICS[12] under EU grant 22236.

References

M. Alexa, B. Kreissig, M. Liepert, K. Reichenberger, L. Rostek, K. Rautmann, W. Scholze-Stubenrecht, S. Stoye *The Duden Ontology: an Integrated Representation of Lexical and Ontological Information* In: Proc. of the OntoLex Workshop at LREC, Spain, May 2002.

J. A. Bateman, R. Henschel and F. Rinaldi *Generalized Upper Model 2.0* Documentation Report of GMD / Institut für Integrierte Publikations- und Informationssysteme, Darmstadt, Germany, 1995.

T. Brants *TnT - A Statistical Part-of-Speech Tagger*. In: Proc. of 6th ANLP Conference, Seattle, 2000.

P. Buitelaar, M. Sintek and M. Kiesel *Feature Representation for Cross-Lingual, Cross-Media Semantic Web Applications* In: Proc. of the Workshop on Knowledge Markup and Semantic Annotation (SemAnnot2005) at ISWC05, Galway, Ireland, 2005.

P. Buitelaar, P. Cimiano, S. Racioppa and M. Siegel *Ontology-based Information Extraction with SOBA* In: Proc. of the International Conference on Language Resources and Evaluation (LREC), 2006.

N. Chang, J. Feldman, R. Porzel, and K. Sanders *Scaling Cognitive Linguistics: Formalisms for Language Understanding*. In: Proc. of the 1st International Workshop on Scalable Natural Language Understanding (ScaNaLU), Heidelberg, Germany, 2002.

W. Drozdzynski, H.-U. Krieger, J. Piskorski, U. Schäfer, F. Xu *Shallow Processing with Unification and Typed Feature Structures - Foundations and Applications.* In Künstliche Intelligenz, 1/2004.

R. Engel *Robust and Efficient Semantic Parsing of Free Word Order Languages in Spoken Dialogue Systems* In Proceedings of Interspeech 2005, Lisbon, Portugal, 2005

G. Francopoulo, M. George, N. Calzolari, M. Monachini, N. Bel, M. Pet, C. Soria *Lexical Markup Framework (LMF)* In: Proc. of the International Conference on Language Resources and Evaluation (LREC), 2006.

A. Gangemi, N. Guarino, C. Masolo, A. Oltramari and L. Schneider *Sweetening Ontologies with DOLCE.* In: Proc. of the 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW), Siguenza, Spain, pp. 166-181, 2002.

A. Gangemi, R. Navigli, P. Velardi *The OntoWordNet Project: extension and axiomatization of conceptual relations in WordNet.* In: Proceedings of ODBASE03, Springer, 2003.

H.-U. Krieger and U. Schafer *TDL---a type description language for constraint-based grammars* In Proceedings of the 15th International Conference on Computational Linguistics (COLING), pp. 893-899, 1994.

W. Lezius *Morphy - German Morphology, Part-of-Speech Tagging and Applications* In: Proc. of the 9th EURALEX International Congress, pp. 619-623, Stuttgart, Germany, 2000.

[12] http://lirics.loria.fr/

B. Loos and Porzel, R. *Towards Ontology-based Pragmatic Analysis*. In Proceedings of DIALOR 2005, June 9 - 11, Nancy, France, pp. 163-166.

V. Micelli, and R. Porzel *Tying the Knot: Ground Entities, Descriptions and Information Objects for Construction-based Information Extraction.* In: Proc. of OntoLex06. Genoa, Italy, 2006.

G. A. Miller *WORDNET: A Lexical Database for English*. Communications of ACM (11): 39-41, 1995.

I. Niles and A. Pease *Towards a standard upper ontology*. In: Proc. of the international conference on Formal Ontology in Information Systems (FOIS01), ACM Press, 2001.

D. Petitpierre and G. Russell *MMORPH - The Multext Morphology Program.* Multext deliverable report for task 2.3.1, ISSCO, University of Geneva. 1995.

N. Reithinger and D. Sonntag *An Integration Framework for a Mobile Multimodal Dialogue System Accessing the Semantic Web*. Proceedings of the 9th European Conference on Speech Communication and Technology (Interspeech) 2005.

Appendix: LingInfo Model