

A System Architecture for Parallel Corpus-based Grammar Learning

Jonas Kuhn

The University of Texas at Austin,
Department of Linguistics
Austin, TX 78712, USA

jonask@mail.utexas.edu

Type of the paper

Lecture

A System Architecture for Parallel Corpus-based Grammar Learning

Jonas Kuhn

This paper describes an architecture for exploiting implicit information about the grammar of the languages included in a parallel corpus. By initially applying statistical word alignment and defining an appropriate representation format for cross-linguistic structural correspondence, this implicit information can feed a system for bootstrapping grammars. The proposed architecture will be underlying in the new PTOLEMAIOS project.

Dieses Papier beschreibt einer Architektur, mit der die implizit in Parallelkorpora enthaltene Information über die Grammatiken der beteiligten Sprachen ausgenutzt werden soll. Wenn vorab eine statistische Wortalignierung angewandt wird und ein geeignetes Repräsentationformat für die crosslinguistische Strukturkorrespondenz definiert wird, kann diese implizite Information in einem Bootstrapping-Ansatz zum Grammatiklernen verwertet werden. Die vorgeschlagene Architektur wird im neuen PTOLEMAIOS-Projekt zur Anwendung kommen.

1. Introduction

In this programmatic paper, an architecture for grammar learning based on parallel corpora is outlined. This proposed architecture is the target for the PTOLEMAIOS project.¹ More concretely, the project goal is to develop a formal architecture and implement a software system that allows one to train a syntactic grammar for a language L from a parallel corpus including L and multiple other languages, for which a relatively small set of sentences has been hand-annotated for syntactic correspondences across the languages. The resulting grammar should be robust and have a broad coverage, while generally providing reliable analyses at the level of (at least clause-local) head-dependent relations; this will

¹ PTOLEMAIOS is for “Parallel Text-based Optimization for Language Learning—Exploiting Multilingual Alignment for the Induction of Syntactic Grammars.” The project has been accepted in the Emmy Noether program of the DFG (German Research Foundation); this means that a junior research group led by the author will be funded at the Saarland University in Saarbrücken.

make the grammar applicable in NLP applications that involve syntactic/semantic parsing at moderate depth, such as information extraction, question answering, or advanced statistical machine translation.

We can also phrase the project goal as a methodological challenge: to develop a formal framework for grammar learning which is sophisticated enough to allow for the integration of insights and assumptions from linguistic theory, and at the same time surface-oriented, robust, and computationally efficient enough so it can be applied on large amounts of real corpus data, without presupposing time-intensive manual annotation of more than a small subset of the data. We expect that with this methodological goal, our project results will transcend the immediate engineering achievements and contribute to our general understanding of the learnability of linguistic knowledge. In particular, our software architecture will serve as an empirical testbed for linguistic representation systems (e.g., of lexical classes, functional/lexical category distinctions, morphological marking, argument structure, etc.) with respect to learnability properties-an aspect for which so far it has been very hard to test theoretical predictions empirically.

Beyond the technical content of the project, I hope that the interdisciplinary nature of the approach will contribute to bridging the gap that has existed between the various neighboring disciplines concerned with aspects of language learning-a central issue in the cognitive sciences.

2. Project methodology

Let us call the initial system architecture we plan to achieve in the project the “PTOLEMAIOS I” system. Figure 1 illustrates the architecture with a flowchart. The main input for the PTOLEMAIOS I system is a parallel corpus, including translated text in at least two languages. As additional input, a subset of the parallel corpus is annotated with cross-linguistic information about phrasal correspondences and an underlying “pseudo meaning representation” which we will discuss below. We also use standard NLP preprocessing techniques, such as part-of-speech tagging and morphological analysis, to the extent that the required resources are available. As additional preprocessing, (i) the parallel corpus is sentence-aligned, following the standard algorithm of Gale and Church (1991), and (ii), a statistical word alignment is trained using the GIZA++ tool, which implements the standard IBM models (Och and Ney (2003)).

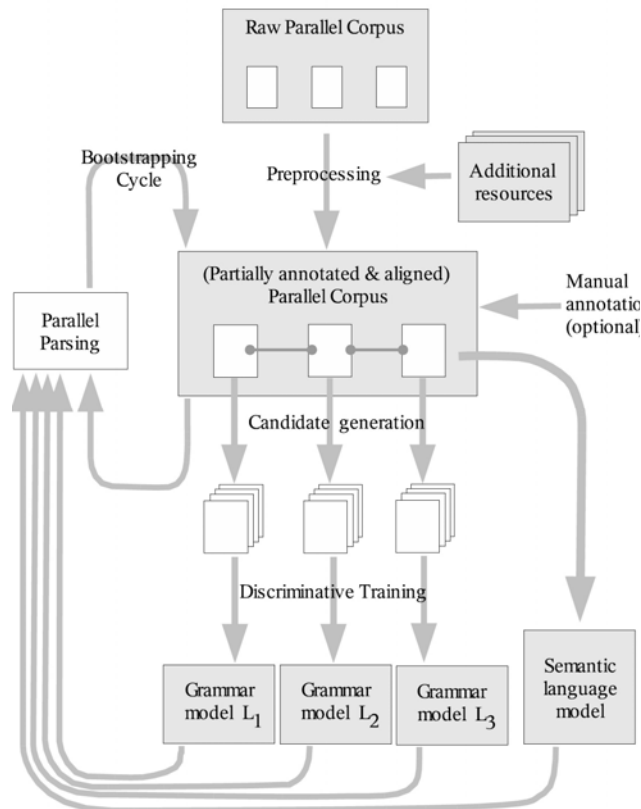


Figure 1: The PTOLEMAIOS I architecture

Besides the preprocessing components, the PTOLEMAIOS I system consists of a bootstrapping cycle for improving grammar models learned from the parallel corpus. The base learning component (section 2.1) at the core of the cycle involves candidate generation (creating generation alternatives to the one observed for each language in the corpus) and discriminative training of a grammar model. The outer loop (section 2.2) applies the grammar models of each stage to the corpus for the creation of a more accurate annotation as the basis for the next learning stage. The outer loop uses a noisy channel model for “parallel parsing”, inspired by the standard model underlying statistical MT.

2.1 Base learning in PTOLEMAIOS I

The central grammar models obtained in the base learning components are log-linear models that predict for some language L , how likely it is to use a particular linguistic realization (let us call it t for “tree”) for a given underlying meaning representation m : we get a language-specific model for $\Pr(t|m)$. The

realization of a meaning with the highest probability can be viewed as our prediction of the correct grammatical realization of the meaning in L .

Note that the model is based on an expressive optimization, i.e., the comparison of alternative realizations of the same underlying meaning (as it is assumed in most linguistic work in Optimality Theory (OT)²), not an interpretive optimization as in classical statistical parsing. However, contrary to classical OT, we do not use a constraint ranking model, but the more robust probability models from statistical NLP. There are strong conceptual arguments for the expressive optimization architecture (compare also section 3): the most crucial aspect in knowing the syntax of English can be paraphrased as knowing that one has to say what did you see?, rather than what saw you? or saw you what? etc. In the initial implementation of the base learning component, we can build on an existing tool: the YASMET system.³ The crucial points to clarify here are (i) what representation formats are used for the underlying meaning m and candidate analyses t , and (ii) how to parameterize the learning (in OT terms, what constraints to assume).

2.1.1. Representations

An important aspect of candidate analyses in OT is the distinction between a part defining the surface form and a part defining the underlying form (in syntactic OT, this corresponds to the meaning). In the PTOLEMAIOS I architecture we use a comparatively simple candidate representation: A candidate analysis is characterized by a tree in which the nonterminal symbols are augmented with atomic-valued features (from a finite set of features, each allowing for values from a finite set); let us call this representation format an augmented phrase structure tree (APT). There may be nonterminal symbols that do not dominate a terminal symbol; all terminal symbols are dominated by non-branching preterminal nodes. Consider for example the tree in figure 2, which we constructed for the English sentence *We will find out all the necessary information*, taken from the Europarl corpus. The EN-Nth and EN-Cat features encode the surface order of the daughters within the local subtree and the part-of-speech category of the syntactic head, respectively. Further feature distinctions for morphological form, sentence type etc. may be added. In our illustrations, only two categories are distinguished in the phrase structure

² For a discussion of directionality in formal OT syntax, see for instance Kuhn 2003.

³ <http://www.fjoch.com/YASMET.html>.

backbone of the feature grammar: clausal nodes (CL) and non-clausal nodes (X).

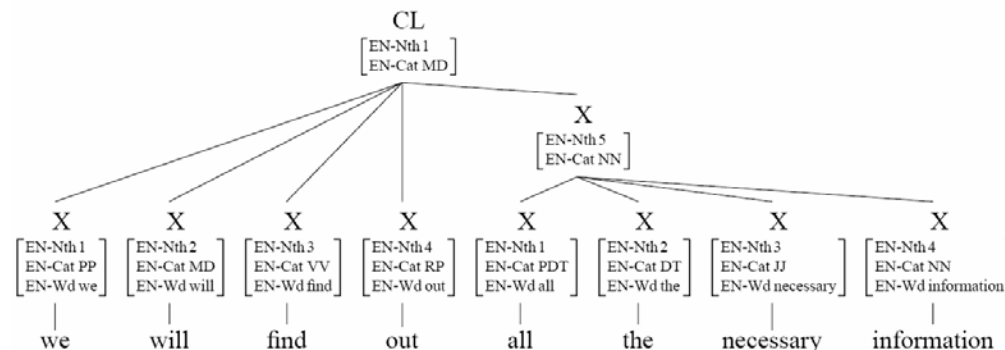


Figure 2: Augmented Phrase Structure Tree

The flat analysis of the clause-internal and NP-internal structure is motivated by the cross-linguistic context in which the analysis is used. We can assume very similar trees for translational correspondents as in (1) or the translations in (2).

- (1) EN We will find out all the necessary information
 FR Nous chercherons toutes les informations nécessaires
 We will seek all the information necessary
- (2) EN I am not satisfied with what happened
 DE Nicht einverstanden bin ich jedoch mit dem was geschehen ist
 not agreeing am I however with that what happened has

We go so far as to assume that translations have the same hierarchical graph structure—we may think of this as the consensus representation. Formally, the multilingual consensus tree is generated by a variant of an inversion transduction grammar (ITG; compare Wu 1997). In the type of transduction grammar we assume, there are three ways in which the realization in a particular language may differ from the others: it is possible (i) to use different orderings of the daughters of a nonterminal, (ii) to use different nonterminal symbols for realizing a preterminal, and (iii) to leave certain nonterminal symbols unrealized.⁴

⁴ A fourth option might be added at some point, using a special re-ordering operation for non-local realization (similar to the subtree cloning operation proposed by Gildea (2003)).

Given these assumptions we can draw a single graph for all translational realizations of a sentence. The graph for (2) in figure 3 includes the English and the German tree. (The actual terminal symbols and the sequencing for English is shown, but the reader may verify that the German version of (2) can be obtained by ordering the sisters in each local subtree according to the DE-Nth feature.) We can now specify the candidate analyses for PTOLEMAIOS I. A candidate analysis is an augmented phrase structure tree (APT) like in figure 3, including information about n other languages from a parallel corpus. The surface part of a candidate analysis for language L is simply the word string obtained by forming the sequence of the L -Wd feature values when traversing the tree according to the order of daughter nodes according to the L -Nth feature. The meaning part of a candidate analysis is essentially the consensus tree representation.

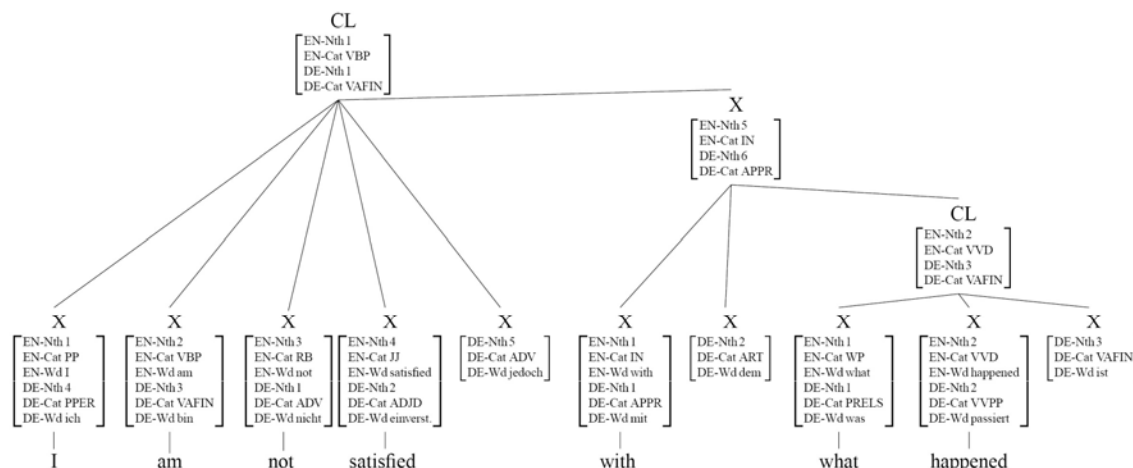


Figure 3: Candidate analysis for English, including pseudo meaning representation based on a German/English corpus

We define a pseudo meaning representation (PMR) relative to a language L as the (in principle unordered) graph obtained from the APT by leaving out all L -specific features and all terminal nodes. A PMR relative to L may contain nodes that have no overt realization in L (e.g., the *jedoch* node for the PMR for English in figure 3). The meaning part of a candidate defines the expressive competitors, i.e., the candidate set in discriminative raining: the alternative realizations of the underlying PMR, i.e., variants of the hierarchical graph structure that may contain different words and a different ordering.

2.2.1. Parameterization of the expressive competition model.

Our grammar models determine $\Pr(t|m)$ for a particular language L . To estimate this conditional probability we apply a log-linear or Maximum Entropy model (Ratnaparkhi, 1999). This allows us to use a large number of constraints or learning features, which do not have to be statistically independent. Since many aspects of t are fixed by the PMR format that we assume for m , only the following information needs to be determined by the model (relative to a specific language L): for a nonterminal node, whether or not it gets realized in L , and if it does get realized which is the syntactic head daughter, and what is relative order of the daughters. For preterminal nodes, the terminal realizing the node in language L has to be determined.

In the flat APT representation, an entire clause is represented as a single subtree of depth 1, containing all argument and modifier phrases. Nominal phrases or prepositional phrases are also internally flat. Thus, most of the systematic cross-linguistic variation in the grammar of clauses and nominal phrases is reflected in ordering alternatives and the addition/omission of certain function words, i.e., these decisions are local to a subtree of depth 1. Therefore we will work with the following **separability assumption**: up to certain limited feature distinctions in our tree representations, the probability $\Pr(t|m)$ for the entire tree can be separated out into a product of probabilities for local subtrees; these can be computed separately and (at least approximately) combined by dynamic programming techniques for unification grammars.⁵ This means that the candidate sets that have to be effectively computed and compared are comparatively small.

Since we plan to use a log-linear model, we can use fairly complex constraints in the training, reflecting linguistic insights into the levels of representation involved. Of course, we do not have access to a true, reliable semantic representation; it is a central hypothesis however that information from the other languages in the parallel text may be used as a substitute. Thanks to the assumption of an expressive optimization we will be in a position to experiment with the actual constraint sets assumed in the theoretical OT literature (to the extent that they can be adjusted to the specific representations we will assume).

⁵ This is related to locality observations for OT competitions based on extended projections, which was made in (Kuhn, 2001). Dynamic programming techniques for log-linear models of syntax are discussed by Geman and Johnson (2002).

2.2 The outer loop in PTOLEMAIOS I

Where does the PMR information (which is required to learn information about the grammar of language L) come from if no full manual annotation of the corpus is planned? The answer is that expressive optimization decisions are learned using a *bootstrapping approach* involving several languages. Initially, either a small set of manually annotated sentence tuples from the parallel corpus is used (indicating clauses and major nominal/prepositional phrases and their syntactic heads), or an unsupervised grammar induction process exploiting just the statistical word alignment is applied (compare Kuhn (2004)).

In each bootstrapping cycle, expressive-optimization-based log-linear grammar models are trained for each of the languages, using only training data for which the PMR has been assigned with high confidence. The result is a model for the conditional probability $\Pr(t_L|m)$. Furthermore, the PMRs are used to train an unconditional model for $\Pr(m)$, using standard PCFG training on a canonical form of the underlying PMR trees with some compiled-out features. The models are then (re-)applied to the full training data in “parallel parsing mode” (see below), assigning PMRs with a certain confidence. The resulting data points for which a PMR tree could be determined with high confidence are used as training data for the next bootstrapping cycle of training log-linear models for the individual languages. Crosstalk between the grammars and the inherent redundancy of parallel corpora leads to an increase of the amount of usable training data.

2.2.1. (Parallel) Parsing with a noisy channel model

This is the central procedure in the PTOLEMAIOS learning approach. The input to the parallel parsing step are corresponding sentences in two or more languages from a parallel corpus and statistical word alignment(s) for the sentence tuples. The *resources* used are the unconditional PMR model and a log-linear grammar model for each of the languages under consideration. Let us call the models used a family of PMR transduction grammars.

To keep the discussion simple, we will discuss parallel parsing for the case of parsing sentences in just two languages; the situation with more languages is analogous. The *output* of parallel parsing is the most probable PMR underlying each pair (or tuple) of input sentences. The application of the various probability models follows the noisy channel model, which is underlying in all work on Statistical Machine Translation (MT). There, the idea is to solve the

problem of translating, say, from French to English, not by training a direct model for translating a given French sentence into an English sentence (finding the English sentence e that maximizes $\Pr(e|f)$). Instead, Bayes' law is used as in (3)⁶ to transform the problem into an equivalent combination of different subproblems, for which separate models are trained: (i) a *translation model* for determining how likely a French sentence is to be a translation of a given English source sentence (note that the assumed direction of translation underlying the training is reversed: a model for $\Pr(f|e)$ is trained), and (ii) a translation-independent *language model* for determining how likely a sequence of words is as an English sentence ($\Pr(e)$). To apply the two models in an actual translation task from French to English (!), a third component, a *decoder* is needed, which searches the space of candidate word sequences e to maximize the product of the probabilities $\Pr(f|e)\Pr(e)$.⁷

$$\begin{aligned}
 (3) \quad & \text{a.} \quad \operatorname{argmax}_e \Pr(e|f) \\
 & \text{b.} \quad = \operatorname{argmax}_e \frac{\Pr(f|e) \Pr(e)}{\Pr(f)} \\
 & \text{c.} \quad = \operatorname{argmax}_e \Pr(f|e) \Pr(e)
 \end{aligned}$$

Parsing can be viewed as a translation problem too: a sentence in a natural language is given, and the output is a translation of the sentence into some structural representation format. In statistical parsing, it is standard to approach this translation problem using direct estimation of the probability $\Pr(\text{tree}|\text{sentence})$.⁸ In the PTOLEMAIOS project, we apply the noisy channel model to the parsing problem. In order to find the right tree for a given sentence, we search for the tree that maximizes the product $\Pr(\text{sentence}|\text{tree})\Pr(\text{tree})$. The idea can be easily generalized to a scenario dealing with several languages (and using a pseudo meaning representation (PMR) m instead of the *tree*, as the common goal for parsing all corresponding sentences $s_{L1} \dots s_{Ln}$ (here we show the case for two languages):⁹

⁶ The denominator can be ignored in the step from (3b) to (3c) since it remains constant when we are looking for the best e , given a fixed f .

⁷ The split of the translation problem in two subproblems gives much more reliable results. Many of the strings receiving a high probability in a $\Pr(e|f)$ model barely resemble English sentences, but since in the noisy channel model the probability is multiplied with the probability assigned by a language model (which in turn doesn't "know" anything about translation), the overall result is quite satisfactory.

⁸ Typically, models for the joint probability $\Pr(\text{tree}, \text{sentence})$ are trained; such a model can also be straightforwardly applied to compare different possible trees for a given sentence.

⁹ The step from (4c) to (4d) makes the assumption that generating a sentence from a PMR in one language is independent from generating from the same PMR in other languages. This

$$\begin{aligned}
(4) \quad & \text{a.} \quad \text{argmax}_m \Pr(m | s_{L1}, s_{L2}) \\
& \text{b.} \quad = \text{argmax}_m \frac{\Pr(s_{L1}, s_{L2} | m) \Pr(m)}{\Pr(s_{L1}, s_{L2})} \\
& \text{c.} \quad = \text{argmax}_m \Pr(s_{L1}, s_{L2} | m) \Pr(m) \\
& \text{d.} \quad = \text{argmax}_m \Pr(s_{L1} | m) \Pr(s_{L2} | m) \Pr(m)
\end{aligned}$$

The scheme involves our conditional grammar models that assign a probability to surface realizations in a particular language, given an underlying meaning representation (or PMR), and an additional model for the prior probability of PMRs (we can call this a “semantic language model”, following Miller et al. (1994)). In section 3.2, we will discuss the advantages of applying a noisy channel model in parsing. The actual search in parsing will be performed by a chart-based algorithm for parallel parsing (compare e.g., Wu 1997), but for space reasons we cannot include the discussion in this paper.

2.3. Evaluation

Evaluating grammar induction systems (as opposed to treebank-trained grammars) is somewhat problematic; no acknowledged standard methodology exists (compare van Zaanen et al. (2004), who come to the same conclusion). Since in PTOLEMAIOS I, the structural restrictions enforced by the PMR format convention are rather limited, this evaluation problem applies in part to our project as well. It is a research question for PTOLEMAIOS to identify adequate evaluation techniques. We will perform comparative evaluations for two variants of parsing: parallel parsing (with a multilingual input) and simple parsing with a monolingual input. A set of evaluation measures commonly employed in work on grammar induction is the comparison of the induced grammar’s behavior against an existing treebank. Besides such an evaluation against structural gold standard representations, we will work on a more task-oriented evaluation methodology. In collaboration with other projects, we will apply the induced grammars in multilingual information extraction or question answering tasks. Finally, a highly adequate application scenario, which at an advanced stage we will use for evaluation, is the use of parallel parsing of a corpus in order to improve statistical word alignment (and ultimately statistical MT). By exploiting the phrase correspondences of the most likely PMR, it can be expected that a simple word-based alignment can be improved.

reflects the assumption that all relevant information is encoded in the PMR (which is of course a simplifying assumption, but an important one to make the approach practical).

3. Discussion of the PTOLEMAIOS methodology

The PTOLEMAIOS project will develop a novel technique for building the central component for NLP systems—grammars. The technique requires only small amounts of hand-annotated text, hence it will be broadly applicable, even to languages for which there is no high commercial interest in language technology. For languages like English, the methodology will be interesting too, since it can be applied to special sublanguages, such as scientific English of a particular community.

3.1. Applicability of the PTOLEMAIOS grammars

Initially, the character of the underlying pseudo meaning representation is primarily determined by its function as the common part in a comparison of (cross-linguistic and language-internal) generation alternatives. However, its language-independent character will be an important factor for the broad usability of the PTOLEMAIOS grammars in various NLP tasks. The great advantage of the grammars resulting from induction with the PTOLEMAIOS system is that they produce parallel tree representations in parsing, i.e., they can be used directly in multilingual applications such as multilingual information extraction, question-answering, and machine translation (for example to improve statistical alignment, or in work on hybrid MT systems).

3.2. Exploitation of a noisy channel model in parsing

The use of a noisy channel model in parsing (i.e., using a combination of a “semantic language model” $\Pr(m)$ and a conditional model $\Pr(s|m)$ to determine the most likely parse/meaning representation m for a given string s) is largely unexplored.¹⁰ Most statistical parsing approaches take a more direct approach, applying a model of $\Pr(m|s)$ (or $\Pr(m, s)$). When fully supervised learning is applied, the direct approach is of course very natural. However in a weakly supervised scenario building on an unannotated corpus, training has to live with meaning representations (m) that are not 100% reliable. In contrast, the actual sentence s is always known. Incorrectly labelled training data will thus always cause wrong predictions if we estimate and apply $\Pr(m|s)$ directly, whereas we may be luckier if we estimate $\Pr(s|m)$ and use it in a noisy channel model: $\Pr(m)$ may be low for incorrect meaning representations, so the better alternative may

¹⁰ But compare (Miller et al., 1994).

indeed win. This presupposes of course that we have a broader empirical basis for estimating the semantic language model $\text{Pr}(m)$ (otherwise it would always suffer from the same labeling mistakes). The PTOLEMAIOS hypothesis is that using a multilingual parallel corpus will provide this additional empirical breadth. For training a model $\text{Pr}(m)$, there are various ways of extending the training data beyond what is usable for training the conditional model for a particular language: the consensus requirement may be relaxed, so sentences for which no high-confidence PMR can be obtained using, say 11 languages, but for which there is a consensus among 4 or 5 languages, could be used in training the language-independent model for PMRs. This may expand the empirical basis considerably, so we can use more fine-grained learning features, e.g., lexicalization features, which may be too sparsely instantiated for the language-specific grammar models.

3.3. A (more) realistic model of the interplay of cognitive systems

The envisaged noisy channel model thus approximates a very basic split of cognitive systems and information sources involved in human sentence processing: as an input sentence \mathbf{s} in language L is parsed, seeking to obtain the correct meaning m for it that was intended by the speaker, (i) grammatical knowledge (**linguistic competence**) is applied (i.e., the language-specific grammar model $\text{Pr}(s|m)$), and (ii) plausibility of the arising meaning representation m is checked, taking into account **contextual and encyclopaedic knowledge** and making inferences as required. Part (ii) is modeled rather crudely by a context-independent statistical distribution $\text{Pr}(m)$, but note that the isolation of this group of knowledge sources is already a considerable conceptual advance over classical statistical parsers which indistinguishably accumulate grammar-specific knowledge, general world knowledge (reflected by certain patterns observed in the corpus) and domain-specific knowledge. Besides the advantage of being able to train the semantic language model on larger amounts of data, the conceptual split of models for different cognitive resources will facilitate error analysis and the improvement of the various subparts of the model.

The combination of knowledge sources (i) and (ii) leads to an infinite search space;¹¹ this means that a heuristic search procedure is required: the decoder. This is the technological correspondence to (iii) the **performance system** in

¹¹ Any meaning whatsoever is a candidate for the correct m .

human sentence processing (which also doesn't cover the full search space of the competence grammar).

4. Conclusion

In this paper, the system architecture and structural representation format that we plan to use in the PTOLEMAIOS project was outlined. The project will be of both theoretical and practical interest, as it addresses computational and representational issues in grammar learning and learnability, and it will lead to an implemented system for bootstrapping robust grammars for language-technological applications like Information Extraction, Question Answering and Machine Translation.

References

- Gale, William A., and Kenneth Ward Church. (1991). A program for aligning sentences in bilingual corpora. In *Proceedings of the 29th Meeting of Annual the Association for Computational Linguistics*, pp. 177–184.
- Geman, Stuart, and Mark Johnson. (2002). Dynamic programming for parsing and estimation of stochastic unification-based grammars. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, pp. 279–286, Philadelphia.
- Gildea, Daniel. (2003). Loosely tree-based alignment for machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL'03), Sapporo, Japan*, pp. 80–87.
- Kuhn, Jonas. (2001). Generation and parsing in Optimality Theoretic syntax – issues in the formalization of OT-LFG. In Peter Sells (Ed.), *Formal and Empirical Issues in Optimality-theoretic Syntax*, pp. 313–366. Stanford: CSLI Publications.
- Kuhn, Jonas. (2003). *Optimality-Theoretic Syntax—A Declarative Approach*. Stanford, CA: CSLI Publications.
- Kuhn, Jonas. (2004). Experiments in parallel-text based grammar induction. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004)*, pp. 470–477.
- Miller, Scott, Robert Bobrow, Robert Ingria, and Richard Schwartz. (1994). Hidden understanding models of natural language. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL 1994)*, pp. 25–32.

- Och, Franz Josef, and Hermann Ney. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29, 19–51.
- Ratnaparkhi, Adwait. (1999). Learning to parse natural language with Maximum Entropy models. *Machine Learning*, 34, 151–175.
- van Zaanen, Menno, Andrew Roberts, and Eric Atwell. (2004). A multilingual parallel parsed corpus as gold standard for grammatical inference evaluation. In *Proceedings of the Workshop on the Amazing Utility of Parallel and Comparable Corpora, LREC 2004*, Lisbon, Portugal.
- Wu, Dekai. (1997). Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23, 377–403.