Mismatches and Divergences: the Continuum Perspective

Evelyne Viegas

Computing Research Laboratory
New Mexico State University
Las Cruces, NM 88003, USA
viegas@crl.nmsu.edu

Abstract. In this paper, we address the issue of resolving divergences (such as *he swam across the river* translates into French as *il a traversé la rivière à la nage)*[1] and mismatches (such as *fish* translates into Spanish as *pez* and *pescado)* in a uniform way. First, we present empirical evidence that only a continuum perspective on divergences and mismatches can help translate them in different languages. Adopting the continuum perspective implies working within a knowledge-based approach, making use of linguistic and world knowledge to translate divergences, mismatches and difficult cases which are in between these two classes of phenomena. Second, we argue that solving all the cases of "gaps" between languages (whether syntactic, semantic or pragmatic) is not just a dictionary problem, but rather a generation problem requiring the use of planning techniques.

1 Introduction

Resolving divergences and mismatches for Machine Translation (MT) has been and still is a hot issue. We cannot review here all the works which have dealt with this issue, the list is impressive. Very briefly, one can distinguish a divergence (roughly speaking, same meaning but different syntactic structure) from a mismatch (roughly speaking, the grammar and the lexicon of the Source Language (SL) do not make some distinctions which are required by the grammar and the lexicon of the Target Language (TL)) by stating that the former shows a difference in construction (such that *he swam across the river* translates into French as *il a traversé la rivière à la nage),*[2] whereas the latter shows a difference in meanings which are equivalent but not identical from one language to another one (such that *fish* translates into Spanish as *pez* and *pescado,* the former being the "generic" *fish* whereas the latter is the one you eat). More attention has been paid to divergences than to mismatches, for mainly two reasons:

1 divergences have been used to provide arguments in favour of or against transfer-based and interlingua-based approaches[3],
2 divergences, being a syntactic phenomenon, can be detected and resolved more easily than mismatches which involve a semantic treatment, as there is, in this case, hardly any syntactic trigger.

[1] A gloss being: he crossed the river swimming.

[2] We adopt here the term "construction" as used by [Levin &; Nirenburg, 1993], which takes into account not only the "traditional" cases of divergences, but also captures languages conventionalities, so that their Japanese example, *Ikanakute wa ikenai,* should not be translated literally as *Not going won't do* but as the more conventional English expression *You should go.*

[3] See [Dorr, 1995] for a review on the debate.

In terms of divergences, the problem seems to be due to the impossibility of constructing an exhaustive list of all the types of divergences, as discussed by Vandooren, in [Vandooren, 1993], who suggests instead providing a typology of divergences for every language pair. This seems to be a very expensive direction to follow. The divergences examples listed in [Dorr, 1995] are syntactic. The divergences studied in [Levin & Nirenburg, 1993] are also syntactic but include some pragmatic information represented as speakers' attitudes to account for conventionality in languages.

The case of mismatches is even more problematic, as there is need not only for contextual knowledge but also for extra-linguistic knowledge, as discussed in [Kameyama et al., 1991]. We present below the semantic distinctions emphasised by [Heid, 1993]:

2.1 the TL word exhibits more semantic distinctions or finer-grained distinctions than the SL one, such that *fish* is lexicalised in Spanish by *pez* and *pescado,*
2.2 the TL word exhibits fewer semantic distinctions or coarser-grained distinctions than the SL one, such that the Spanish nouns *pez* and *pescado* are both lexicalised in English as *fish,*
2.3 the TL and SL words do not carry the same semantic distinctions; for instance, such that the Spanish verb *madrugar* is lexicalised in English by *get up early,*

We would like to add to the above list:

2.4 the TL or SL share the same semantic features but have different stylistic or pragmatic usage of their lexicalisations;[4]
2.5 the two conceptual worlds between the languages differ; in other words, when we have a conceptual mismatch.[5]
2.6 there is a lexical conceptual gap between the TL and the SL; SL has a lexeme whose meaning is absent in the TL.

We call all the above distinctions "language gaps". Our interest in resolving language gaps (i.e. when there is not a one-to-one mapping between languages, whatever the linguistic level, lexical, semantic, syntactic, etc...) using a knowledge-based approach along with planning techniques comes from noticing that all earlier work ([Lindop & Tsujii, 1993], [Dorr, 1995], [Heid, 1993], [Kameyamaet al., 1991], [Levin & Nirenburg, 1993], [Palmer & Wu, 1995], ...), whatever the approach or MT paradigm adopted, seem to fail to solve completely (i.e., recognise **and** generate) language gaps.

More generally, if we want to account for all types of "language gaps", we suggest distinguishing between four major types of "language gap", corresponding to their level of treatment:

[4] For instance, [Kittredge, 1995] gives examples where in French we use *l'emploi a peu varié* (employment has little changed) whereas English prefers to use a state with modifier *employment remained virtually unchanged.*

[5] For instance, for insurance policies one should not make the same inferences based on driving in left hand-side and right hand-side countries, unless the conceptual worlds have been rendered "equivalent". For instance, the French text extracted from the French UAP corpus *l'adversaire qui prenait son virage complètement* à gauche *m'a heurté et maintenant il profite de ce que j'avais bu pour me donner tous les torts. Honnêtement est-ce qu'il vaut mieux* être saôul à droite ou chauffard à gauche? translates into English as *the adversary who took his turn completely* on the left *[lane] is the one who drove into me, and now he takes advantage of the fact that I had been drinking to make me responsible for all casualties. Honestly, what is the best,* be a drunkard on the right or a road-hog on the left? Having a Natural Language Processing (NLP) system make the same inferences for the two conceptual worlds could lead to wrong inferences in resolving further coreferences.

conceptual: when the conceptual worlds representing different realities can be made "equivalent"
pragmatical: when the languages have different conventional ways of expressing a meaning
semantic: when the language units share some semantics, most of it overlapping; or hardly share any semantics
lexical: when the languages share semantics but differ in lexicalisation.

2 Towards a Theory of "Language Gaps"

In this paper, we focus on semantic and lexical gaps, where we further distinguish 4 types of gap description, as presented in (Figure 1).

Gap Description	Example in SL	Example in TL
synonymy	*peu varié*	*remain virtually unchanged*
hyponymy	*fish*	*pez pescado*
hypernymy	*pez pescado*	*fish*
relevancy	*madrugar*	*get up early*

Figure 1. Four Kinds of "Language Gaps."

We consider the four kinds of gaps as listed in (Figure 1) from a processing viewpoint, specifically, as three sub-problems of the "language gaps" theory for lexical selection in generation: **synonymy, hypernymy, relevancy.**

1. Synonymy. Figure 2 shows a case where the lexical items of SL (sl_{11}) and those of TL (tl_{21}, tl_{22}) share the same semantics **SEM** only differing from the stylistic point of view. Selecting the right lexical item in TL (either tl_{21} or tl_{22}) is a part of lexical selection.

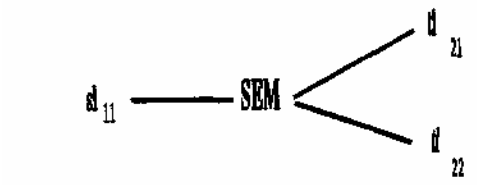
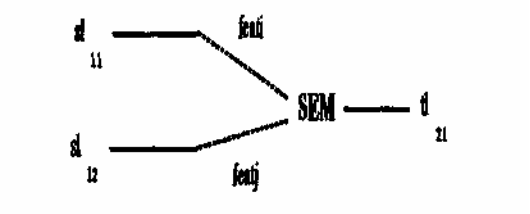


Figure 2. Synonymy.

Both hypernymy and relevancy first involve recognising the language gap. Computationally, this is equivalent to have the language matcher[6] find no lexeme or phrase in the TL that exactly supports the SL lexeme or expression.

[6] The language matcher is the process which tries to "match" or unify lexicon representations for different languages.

2. Hypernymy. Figure 3 shows a case of hypernymy, that is, when the TL does not make distinctions required by the SL. This case is not difficult in the sense that the TL is not ambiguous with respect to itself, but just from the SL perspective. The *fish* example shows that in English it does not matter whether we are talking about food or animal as *fish* "conflates" both interpretations in one single word. One might talk about "vagueness" in this case. From the processing viewpoint, in a knowledge-based approach, selecting the appropriate translation candidate for the Spanish *pez* or *pescado* is equivalent to search for the least common hypernym of the semantics of the Spanish lexical items.



Figures. Hypernymy.

3. Relevancy

Figure 4 shows the most challenging case of semantic gap. This type of gap does not directly support a translation between SL and TL, but only some approximate translation that we call **relevancy.** By relevancy, we mean to focus on the most relevant information from the SL text to be carried across to TL to best match the most equivalently relevant information in TL. From a processing viewpoint, this case involves taking into account static and dynamic resources: conceptual world model, "script-like" information, and an engine to draw inferences on the static resources in context. Although we cannot detail the process in this paper, we will illustrate it through an example.

The relevancy process determines for a particular word or phrase in SL (sl_{11}) the set of possible candidates, whether lexicalised or not: words and phrasals (tl_{21}, ..., tl_{2n}), as well as semantic representations (sem_k). This set will be added to the set of candidates, input to the lexical selection process. The hyper and hypo in Figure 4 stand for hypernymy and hyponymy respectively. The most difficult case of relevancy concerns when SL has a lexical item or expression which meaning is not found in TL. There, the SL lexeme(s) must be given a *definiens* trying to find the best words in TL to express it, this process might involve using hypernymy and hyponymy treatments and will require an inference engine.[7]

Hyponymy can be understood as a sub-type of the relevancy type: further specifying the meaning of a SL word (sl11) to best "match" the meanings of the words from TL (tl_{21}, tl_{22}), requires contextual processing, but not necessarily extralinguistic knowledge.[8] For instance,

[7] One can think of this gap in terms of acquisition of a language.

[8] In this sense, the hyponymy treatment includes Nirenburg's notion of saliency which holds at the lexical level only. By saliency the author meant to lexicalise in as few lexemes as possible in the TL, the most semantic information of the input. For instance for *madrugar* → *get up early,* we would rightly match the pairs instead of generating for *madrugar* say *get up in the morning before 6am.*

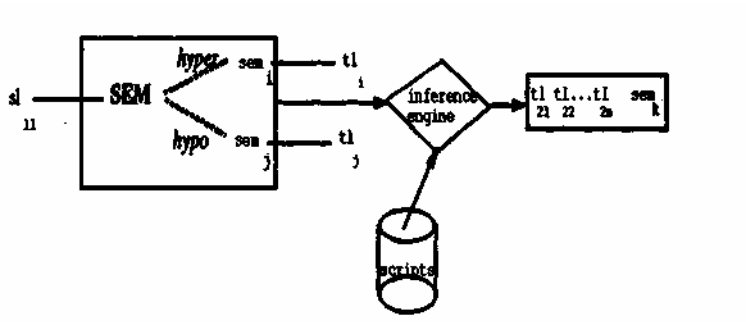


Figure 4. Relevancies.

assuming the semantics for *fish, pez, pescado,* given below, going from English to Spanish might require more or less contextual reasoning to match the SL text:

fish(X,)
    sem: FISH(X)
pez(X)
    sem: FISH(X), LOCATION(WATER)
pescado(X)
    sem: FISH(X), EDIBLE(X)

In the presence of LOCATION(WATER) in the context of FISH, the language matcher will try to best match as much semantic as possible in TL, selecting the Spanish *pez* as in the example *I saw many fish in Lake Powell.* However, more contextual processing might be involved for the language matcher to find the best solution, in particular in the case of non literal language such as in *I liked the fish I had at noon, what was it?,* where the event ellipsis EAT has first to be reconstructed ([Viegas & Nirenburg, 1995]), to find that in this context FISH, as a potential theme of EAT, is of type EDIBLE and therefore *pescado* will be selected. EAT illustrates a case of sem_k in Figure 4.

3 Handling "Exceptions" as "Typical" Cases: the Continuum Perspective

In the following, we first address the dictionary issue showing how to treat "language gaps" uniformly, then we show that this issue can be reduced to a generation problem. Finally, we demonstrate the need for using planning techniques to actually generate the best translations or relevancies in TLs.

Upon a close exam of empirical data, it is often difficult to classify a translation pair as a lexical gap (a clear example of predicative gap: *he limped up the stairs* → *il monta les marches en boitant* (French) (he went up the stairs limping)) or a semantic gap (a clear example of semantic underspecification: *pez, pescado* (Spanish) → *fish).* For instance, the English *cook* and *bake,* could both be translated into French as *cuire* or they could be translated as *cuire (sur le feu)* and *cuire au four* respectively, thus presenting a case of divergence or conflation as discussed in [Talmy, 1985].[9] This last observation favours the continuum perspective, allowing

[9] Note that whereas *sur le feu* can be elided, *au four* cannot be elided, it can just be conflated as part of the argument of the verb, for instance if the argument is *pain* (bread).

for a compositional treatment on a scale ranging from fully compositionality *(cuire sur le feu → cook on the stove)* to absence of compositionality *(kick the bucket)* with in between a "semi-compositionality", via constructions *bake → cuire au four.* Providing a uniform treatment within a continuum perspective entails the use of world knowledge along with linguistic knowledge as one cannot "freeze" language pairs (as thought for instance by [Melamed, 1996]) as we explain in section 4. Moreover, the continuum perspective allows for a homogeneous treatment without any need for classifications or typologies.

Some confusion with respect to semantic gaps seems to come from a widely held belief that an SL which has fewer lexical units corresponding to a greater number of lexical units in the TL is ambiguous from a monolingual perspective, such as in the examples:

fish → pez/pescado (Spanish)
cuire (French) → *bake/cook*
se trouver (French) → *stand/lie*

The word *fish* becomes ambiguous only with respect to Spanish, *cuire* or *se trouver* (French) with respect to English.[10]

We will treat semantic gaps as cases of underspecification (elsewhere called vagueness).[11] We exemplify below the treatment of language gap using French examples. We consider the lexeme cuire as unambiguous in French[12] but consider it as underspecified with respect to English. We assume a knowledge-based approach, that is semantics based, and a conceptual world model (such as the one described in [Nirenburg et al., 1994] or [Mahesh, 1996]) where we have a concept labeled COOK, which minimally contains the following relevant information, presented in a frame-based form:

COOK
AGENT: HUMAN
THEME: PHYSICAL-OBJECT
INSTRUMENT: COOKING-EQUIPMENT
LOCATION: PLACE

The lexical entries for the English words *cook* and *bake* minimally contain the following semantic information, presented here in an informal way, where OVEN is a hyponym of COOKING-EQUIPMENT:[13]

cook(X.Y)
sem: COOK(X,Y), AGENT(X), THEME(Y), INSTRUMENT(COOKING-EQUIPMENT)
bake(X,Y)
sem: COOK(X,Y), AGENT(X), THEME(Y), INSTRUMENT(OVEN)

Now let us assume the following entries for *cake* and *pasta* along with the concepts PASTA and CAKE:

[10] We will not address here the cases of acquisition where a word can definitely be ambiguous for a native speaker in his/her own language until he/she knows how to use the word(s) correctly in any situation.

[11] There is no consensus on what is underspecification (see [Van Deemter & Peters (eds.), 1996] for different approaches). In this paper, we will consider a lexeme as semantically underspecified when its meaning can be further specified for a particular truth value in context. For instance, *fish* is underspecified with respect to its ANIMAL or FOOD meanings in *I bought two fish.* It becomes specified in *I bought two fish to put them in the aquarium.*

[12] We only consider here the meaning of *cuire* which is translated into the English verbs examined here.

[13] We avoid here the discussion on how to write the semantics in a lexicon entry or how concepts are inherited for explanatory purposes.

PASTA	pasta(X)
ISA: PREPARED-FOOD	sem: PASTA(X)
CAKE	cake(X)
ISA: BAKED-POOD	sem: CAKE(X), INSTRUMENT(OVEN)

Relevant elements of a semantic representation for the simplified French sentence *Jean a fait cuire le gâteau,* should look like:

COOK(X,Y), AGENT(JEAN), THEME(CAKE), INSTRUMENT(OVEN)

Translating the above semantic representation into English does not require extra processing, as would be the case in [Kameyama et al., 1991], where this would be treated as a specialisation,[14] as there are two entries, *cook* and *bake,* which can lexicalise the concept COOK in the English lexicon.[15] However, in the lexicon, *bake* requires its instrument to be of type OVEN, which we find in the semantics of the English word *cake* mapped to CAKE. Therefore, in this case, the gap is treated as a generation problem where the semantic constraints taken into account come not only from the predicate "bake(X,Y)" but also from the arguments of the predicate "cake(X)". Note that in the case of translating from English to French we would again treat it as a generation problem with no more processing than previously, whereas [Kameyama et al., 1991] would require a generalisation mechanism for such cases. Let us now examine more complex examples for *cook* and *bake* which translate into *cuire (sur le feu); cuire au four]*. One could try to list them in multilingual dictionaries, such that the English verbs *bake* and *cook* can be realised as the French expressions *cuire au four* and *cuire sur le feu*.[16] However, treating this language gap requires more than representational issues as we illustrate in next section.

4 Planning "Language Gaps" in Generation

Let us now, for the sake of simplicity, look at some isolated sentences involving *bake* and *cook* and *cuire; cuire au four; cuire sur le feu:*

3 (a) **Cuis** le pain → **Bake** the bread
3 (b) **Cuis** les pâtes (al'dente) → **Cook** the pasta (al'dente)
3 (cl) **Cuis** les pâtes **au four** → **Bake** the pasta
(c2) **Cuis** les pâtes **au four** → **Cook** the pasta in the oven
3 (d1) **Cuire** les pâtes **au gratin** pas plus de 20 minutes → **Bake** the pasta **au gratin** no longer than 20 minutes
(d2) **Cuire** les pâtes **au gratin** pas plus de 20 minutes → **Cook** the pasta **au gratin** no longer than 20 minutes
3 (e) I prefer **baked** meals to meals **cooked** on the stove top → Je preferre les plats **au four** aux plats (cuisinés) **sur le feu**
3 (f) **Cuire** le pain et les pâtes → **bake** the bread, then **cook** the pasta

[14] See [Kameyama et al., 1991] for the use of specialisation and generalisation for mismatches; and [Palmer & Wu, 1995] for their treatment of divergences.
[15] In reality the concept COOK can have many other lexicalisations such as *boil, fry, grill, braise, ...* that we do not consider here for the sake of simplicity.
[16] Note the metonymy on *feu* (fire) for *stove,* which would have to be resolved in further processing.

In what follows we will consider the lexical entries for *cook* and *bake* as defined previously, and will give *cuire* the following semantic representation:
cuire(X,Y)

sem: COOK(X,Y), AGENT(X), THEME(Y), INSTRUMENT(COOKING-EQUIPMENT)

The problem with the approach of [Kameyama et al., 1991] is illustrated in 3(dl,d2) where if we want to specialise, we have to rely on the semantics of the noun which sometimes is ambiguous, such as in 3 (dl,d2) where, although there is a preference for generating 3(dl) rather than 3(d2), it is still acceptable to have 3(d2). Moreover, contextual constraints present in the semantic representation will help to eventually generate *bake the pasta* if in the linguistic context we are told that *pasta* is a reference for *lasagna;* in this case it would be misleading to generate *cook the pasta* as a coreference for lasagna. The point we want to make is that it is impossible to "freeze" the meanings of *bake* and *cook* as being equivalent to *cuire au four* and *cuire (sur le feu)* respectively.

Finally, example 3(f) shows that generating a mismatch requires that lexical selection be done contextually: in other words, requires text planning. By planning we mean that we try to find a best match for the input semantics (as described in [Beale & Viegas, 1996]), whether originally input to the generator or calculated dynamically via the theory of language gaps, with maximal adherence to the lexicon constraints, while taking into account prior and further context. For instance, the French sentence 3(f) can be planned as a coordination of events, the second *cuire* being elided as it is a coreferential lexical anaphora with the first *cuire* [Tutin & Viegas, 1996]; in English, however, we might prefer to lexicalise both events *bake* and *cook* and to plan them as temporally successive rather than as a coordination.

5 Conclusion

In this paper, we focused on two points. First, we discussed representational issues for solving language gap situations between SLs and TLs. We advocated a continuum perspective entailing a semantics based approach, in order to be able to treat all the cases of gaps. Second, we demonstrated that solving "language gaps" goes beyond the issue of dictionary representation and is part of the larger process of lexical selection. We advocated solving the language gaps by using planning techniques at the generation level. Finally, one must remember that lexical selection is a complex process. In most generation systems, lexical choice is done "vertically" from concepts to the lexemes. This type of treatment is not adequate to treat language gaps, as the actual lexicalisations must take into account prior and current lexicalisations and use many lexico-semantic relations.

Further research involves theoretical and practical issues. From a theoretical viewpoint, there is still a lot of work to be done to understand how best to deal with the trade-offs between the lexicon and the conceptual world. We believe that work on underspecification might help reduce the needs for specialisation or generalisation procedures for cases of semantic distinctions which overlap between languages.

From a practical point of view, one could envisage that statistical techniques might help reduce the need for putting so much burden on the knowledge sources and still be able to get very good results in the cases of hypernymy and synonymy (e.g. [Melamed, 1996]). However, the case of relevancy, the most challenging from a computational linguistic perspective, requires a knowledge-based approach, and should receive more attention in the future.

Acknowledgments: I am grateful to the Mikrokosmos team for many discussions we had on the research presented here, however all errors remain mine. Research reported in this paper was supported in part by Contract MDA904-92-C-5189 from the U.S. Department of Defense.

References

[Beale & Viegas, 1996] Beale, Stephen and Evelyne Viegas. 1996. Intelligent Planning meets Intelligent Planners, In *Proceedings of the Workshop on Gaps and Bridges: New Directions in Planning and Natural Language Generation, ECAJ'96,* Budapest, pp. 59-64.

[Dorr, 1995] Bonnie J. Dorr. 1995. A lexical-semantic solution to the divergence problem in machine translation. In P. St-Dizier P. and E. Viegas (eds), *Computational Lexical Semantics,* Cambridge University Press, pp. 36T-395.

[Heid, 1993] Ulrich Heid 1993. Le lexique: quelques problèmes de description et de représentation lexicale pour la traduction automatique. In P. Bouillon and A. Clas (eds), pp. 169-196.

[Kameyama et al., 1991] Kameyama, M., R. Ochitani and S. Peters. 1991. Resolving Translation Mismatches With Information Flow. In *Proceedings of the Association for Computational Linguistics, 1991,* pp. 193-200.

[Kittredge, 1995] Kittredge, R. 1995. Efficiency vs. Generality in Interlingual Design: Some Linguistic Considerations. In *Working Notes of the Multilingual Text Generation Workshop at the 14th International Joint Conference on Artificial Intelligence, Montréal, 64-74*

[Levin & Nirenburg, 1993] Levin, Lori L. and Sergei Nirenburg. 1993. Principles and Idiosyncrasies in MT Lexicons, In *Proceedings of the 1993 Spring Symposium on Building Lexicons for Machine Translation,* Stanford, California.

[Lindop & Tsujii, 1993] Lindop, J. and J. Tsujii. 1993. Complex Transfer in MT: A Survey of Examples. Technical report, num 91, 5, Center for Computational Linguistics, Manchester, UMIST.

[Mahesh, 1996] Mahesh Kavi. 1996. Ontology Development for Machine Translation: Ideology and Methodology Technical report, MCCS-96-292, CRL, New Mexico State University.

[Melamed, 1996] I. Dan Melamed 1996. Automatic Construction of Clean Broad-Coverage Translation Lexicons. In Proceedings of *AMTA-96,* Montréal, Québec, Canada, pp. 125-134.

[Nirenburg et al., 1994] Nirenburg, Sergei, Victor Raskin and Boyan Onyshkevych. 1994. Apologiae ontologiae. Memoranda in Computer and Cognitive Science MCCS-95-281. New Mexico State University: CRL.

[Palmer & Wu, 1995] Palmer, M. and Z. Wu 1995. Verb Semantics for English-Chinese Translation. *Machine Translation,* Volume 10, Nos 1-2.

[Talmy, 1985] Talmy Leonard. 1985. Lexicalization Patterns: semantic structure in lexical forms. In Shopen, T. (ed.) *Language Typology and Syntactic Description III: Grammatical Categories and the Lexicon.* Cambridge University Press.

[Tutin & Viegas, 1996] Agnès Tutin and Evelyne Viegas. 1996. Generating Coreferential Anaphoric Definite NPs. In Proceedings of *Discourse Anaphora and Anaphor Resolution Colloquium,* Lancaster University.

[Van Deemter & Peters (eds.), 1996] Van Deemter Kees and Stanley Peters. 1996. *Semantic Ambiguity and Underspecification.* CSLI Publications.

[Vandooren, 1993] Vandooren Françoise. 1993. Divergences de traduction et architectures de transfert. In P. Bouillon and A. Clas (eds) *La traductique* AUPELF-UREF, Montréal, Canada.

[Viegas & Nirenburg, 1995] Viegas Evelyne and Sergei Nirenburg. 1995. The Semantic Recovery of Event Ellipsis: its Computational Treatment. In Proceedings of the Workshop *Context in Natural Language Processing, of the 14th International Joint Conference on Artificial Intelligence (IJCAI9S),* Montréal, Québec.